UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

School of Electronics and Computer Science
School of Geography

Spatiotemporal modelling of Health Management Information System data to quantify malaria treatment burdens in the Kenyan Government's formal health sector

By Peter W. Gething

Doctor of Philosophy

June 2006

UNIVERSITY OF SOUTHAMPTON

Abstract

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE and SCHOOL OF GEOGRAPHY

Doctor of Philosophy

SPATIOTEMPORAL MODELLING OF HEALTH MANAGEMENT INFORMATION SYSTEM DATA TO QUANTIFY MALARIA TREATMENT BURDENS IN THE KENYAN GOVERNMENT'S FORMAL HEALTH SECTOR

By Peter W. Gething

Reliable and timely information on disease-specific treatment burdens within a health system is critical for the planning and monitoring of service provision. Health Management Information Systems (HMIS) exist to address this need at national scales across Africa but are failing to deliver adequate data due to widespread under-reporting by health facilities. Faced with this inadequacy, vital public health decisions often rely on crudely adjusted regional and national estimates of treatment burdens. This study has taken the example of presumed malaria in outpatients within the largely incomplete Kenyan HMIS database and has developed geostatistical modelling frameworks for the prediction of the monthly tally of treatments for malaria (MC) at all facilities and months where this value is missing. Three different kriging methodologies were compared to test the effect on prediction accuracy of (a) the extension of a spatial-only to a space-time prediction approach, and (b) the replacement of a globally-stationary with a locally-varying random function model. Space-time kriging was found to produce predictions with 98.4% less mean bias and 14.8% smaller mean imprecision than conventional spatial-only kriging. A modification of space-time kriging that allowed space-time variograms to be recalculated for every prediction location within a spatially-local neighbourhood resulted in a larger decrease in mean imprecision over ordinary kriging (18.3%) although mean bias was reduced less (87.5%). Because the MC variable included non-spatial variation caused by differences between individual facilities and their catchment populations, a series of studies were conducted to model catchment population size. These predictions require refined models that incorporated rich local data that were not available at the national level so directly estimated catchment population values were not available. An alternative approach was developed that incorporated data on the total number of outpatients seen at facilities each month as a proxy measure of catchment size. Two modelling frameworks were developed to implement this approach and the most accurate model was identified in a cross-validation exercise. A model-based and an empirical method were developed to measure the uncertainty of predictions of MC and how this changed as sets of predictions were aggregated in space and time. The final set of predictions enabled the national treatment burden for presumed malaria in the government health sector to be defined during the 1996-2002 period. During this time, the national annual treatment burden was predicted as 6.8 million cases, with an expected margin of error of 1.3%. The modelling framework presented here provides for the first time reliable information from imperfect HMIS data to support evidence-based decision making at national and sub-national levels.

Contents

Acknowledgements

I would like to take the opportunity to express my thanks to everyone that has helped me during the course of this PhD. Fist and foremost, my profound thanks to my supervisors, Professors Pete Atkinson and Mark Nixon, who have provided me guidance and support throughout and from whose energy, motivation, and experience I have gained immeasurably. My profound thanks also to Professor Bob Snow and Dr. Simon Hay for framing the HMIS problem and providing me with invaluable guidance and expertise along the way. I am hugely grateful for having had the opportunity to address this problem and hope I have made a valuable contribution.

I am extremely grateful to Dr. Abdisalan Noor who has been instrumental in providing the data on which this thesis is based and who has been a constant source of advice and knowledge about Kenya and its health system. Without Noor's painstaking work in constructing the national health service database and in obtaining, formatting, checking, and cleaning the HMIS outpatient database, the opportunity to carry out the work in this thesis would have never arisen. Thanks also in this regard to everyone else who helped in the development of these data sources including Priscilla Gikandi in Nairobi, and Briony Tatem in Oxford. Thanks also to Professor David Rogers who developed code for the formatting of the HMIS database. My thanks to Drs. Andy Tatem and Mike English who provided comments on research output. The data used in the two catchment modelling studies resulted from surveys carried out in 2001-2002 by the Kenya Medical Research Institute – Wellcome Trust Collaborative Programme in Nairobi, particularly Drs. Abdinasir Amin and Dejan Zurovac. I am grateful to both of them. I would also like to acknowledge Dr. James Nyikal, the Director of Medical Services, Ministry of Health, Government of Kenya for his support and policy framework for this work and Dr. Esther Ogara, Department of Health Management Information System, Ministry of Health, Government of Kenya, for facilitating the acquisition of the outpatient data. My thanks to the Engineering and Physical Sciences Research Council who part-funded this PhD and to the School of Electronics and Computer Science and School of Geography, University of Southampton.

I would like to express my thanks to the examiners, Professors Margaret Oliver and David Martin, for their thorough efforts in reviewing this project. Their suggestions have helped to improve this document considerably.

Lastly, I want to say thanks to Mum and Dad, Lucy, James, and the rest of my family for their love, support, and help throughout.

List of Abbreviations

ACT	artesunate-based combination therapy
AQ	amodiaquine
ARI	acute respiratory illness
ccdf	conditional cumulative distribution function
cdf	cumulative distribution function
D	dispensary
DEM	digital elevation model
DHS	Demographic and Health Survey
DOMC	Division of Malaria Control
DPTWG	Drug Policy Technical Working Group
DSS	Demographic Surveillance System
EA	enumeration area
EIR	entomological inoculation rate
GIS	geographic information system
GoK	Government of Kenya
GPS	geographic positioning system
GSLIB	Geostatistical Software LIBrary
H	hospital
HC	health centre
HIS	health information system(s)
HMIS	health management information system(s)
ICD	International Classification of Diseases
IID	independent and identically distributed
IMCI	Integrated Management of Childhood Diseases
ITNs	insecticide treated (bed) nets
KEMRI	Kenya Medical Research Institute
LMC	linear model of coregionalisation
LSTOK	local space-time ordinary kriging
MAE	mean absolute error
MC	malaria cases
MDGs	Millennium Development Goals

ME	mean error
MICS	multiple indicator cluster survey
MMTC	mean monthly total cases
MP	malaria proportion
NHSD	National Health Service Database
OK	ordinary kriging
PARIS21	Partnership in Statistics for Development in the 21st Century
PPV	positive predictive value
RBM	Roll Back Malaria
RF	random function
RUR	relative utilisation rate
RV	random variable
SEIR	susceptible, exposed, infectious, recovered
sGs	sequential Gaussian simulation
SK	simple kriging
SMC	standardised malaria cases
SMR	standardised morbidity ratio
SP	sulphadoxine pyrimethamine
STKED	space-time kriging with an external drift
STKT	space-time kriging with a trend
STOK	space-time ordinary kriging
ST-sGs	space-time sequential Gaussian simulation
STSK	space-time simple kriging
TC	total cases
TS	time series
USAID	United States Agency for International Development
UTL	useful therapeutic life
VR	vital registration
WLS	weighted least-squares
WHO	World Health Organisation
WHO/AFRO	World Health Organisation Regional Office for Africa
WHS	World Health Survey

List of Figures

List of Tables

Chapter 1

Introduction

Chapter 1

1. Introduction

1.1 Project motivation

The United Nations Millennium Development Goals (MDGs) reflect the principal development challenges facing the international community (UN, 2000). The eight goals, agreed by 147 member states of the United Nations in 2000, consist of 18 specific targets to be met by 2015, and are accompanied by 48 quantifiable indicators. Three of the eight goals relate explicitly to the most prominent public health issues faced by low-income countries: the reduction of child mortality; the improvement of maternal health; and the combating of HIV/AIDS, malaria, tuberculosis, and other diseases. Underpinning all efforts to meet these public health challenges is the need to strengthen health care systems in low-income nations (WHO, 2000a). With this renewed focus on health system functionality has come growing recognition of the fundamental importance of health information to form an evidence base for decisions about health system organization, financing, management, and delivery (Murray et al., 2004; AbouZahr and Boerma, 2005; Macfarlane, 2005; World Economic Forum, 2006). Reliable and timely information on physical, human and financial health system resources, the type, quality and coverage of health services offered to the population in need, and the impact of those services on population health is essential for the effective planning of service provision, the implementation of targeted public health programmes, the allocation of resources, the monitoring of intervention strategies, and the evaluation of policies and programmes (Murray et al., 2003, 2004).

Although large quantities of health data are collected worldwide through population-based surveys, surveillance programmes, vital registration systems, and routine health system records, data in low and middle-income countries remain incomplete, inconsistent and inadequate to meet the challenges set by the MDGs. A basic information requirement for the planning and delivery of drugs, staff and other commodities within a national health system is accurate and up-to-date data on the number of patients utilising different health facilities and the types of illness for which they are treated. Such information requirements are addressed in most countries by some form of national Health Management Information System (HMIS) that, among many other functions, coordinates the routine acquisition of treatment records from health facilities and the transfer, compilation and analysis of these data through district, regional and national levels.

A perfect HMIS requires all health facilities to report promptly at regular intervals, allowing comprehensive quantification of treatment events through time and space across the health system. The reality of HMIS in Africa and elsewhere stands in marked contrast to this ideal (WHO/SEARO, 2003; WHO/AFRO, 2003; Setel et al., 2005; WHO, 2005a; Health Metrics Network, 2005a). Typically, many facilities never report or report only intermittently resulting in spatially and temporally incomplete national data (Al Laham et al., 2001; MoH Kenya, 2001a; Rudan et al., 2005; Health Metrics Network, 2005b). Following several decades of donor investment in HMIS across Africa the incomplete nature of routine national reporting has shown little improvement (Evans and Stansfield, 2003; AbouZahr and Boerma, 2005). The widespread inadequacy of national HMIS data sets presents a substantial obstacle to evidence-based public health decision-making (Snow et al., 2003a). Faced with poor data coverage, important public health metrics are often estimated using rudimentary methods to account for missing values (WHO, 1995; Kindermans, 2002; Derriennic, 2003).

The Kenyan HMIS is typical of many in Africa, with widespread under-reporting by health facilities and a largely incomplete national database (MoH Kenya, 2001a). A particularly important health system metric in Kenya, as elsewhere in sub-Saharan Africa, is the treatment burden for malaria, defined here as the total number of malaria diagnoses that are made at Government of Kenya (GoK) health facilities in a given month or year. Malaria is the most common diagnosis in outpatients across Kenya (MoH

Kenya, 2001a) and the treatment protocol for this disease is undergoing a period of transition in Kenya due to the decreasing efficacy of existing drugs and the introduction of new, more expensive ones (Kindermans, 2002; WHO, 2005b; MoH Kenya, 2005c). The procurement of these new drugs requires accurate quantification of the number of treatments that are required (WHO, 2006). Such quantification is also required if donor assistance is to be obtained from new investment mechanisms such as the Global Fund to Fight AIDS, Tuberculosis, and Malaria (Murray et al., 2003; Ashraf, 2005; GFATM, 2005).

This project addresses the problem of producing reliable estimates of the treatment burden for malaria in the Kenyan government's formal health sector using incomplete data from the national HMIS database.

1.2 Project aims and approach

1.2.1 Project aims and objectives

The overall aim of this project is to provide reliable national and sub-national estimates of the annual outpatient treatment burden for malaria at health facilities in the formal government health sector in Kenya. This overall aim will be achieved by meeting the following specific objectives:

(1) to develop models by which missing values in the Kenyan HMIS outpatient malaria record can be predicted (along with estimates of prediction accuracy) to produce a spatiotemporally complete database;

(2) to evaluate these models and identify the best-performing approach in terms of prediction accuracy; and

(3) to implement this best-performing approach to estimate monthly and annual outpatient treatment burdens for malaria at the national, provincial, and district levels with accompanying uncertainty estimates.

1.2.2 Project approach

Whilst the Kenyan HMIS is characteristic of others in Africa in terms of the incomplete status of the national database, it is distinct in that the vast majority of health facilities have been recently georeferenced (Noor et al., 2004, Noor 2005). This process involves the identification and recording of latitude and longitude co-ordinates for each facility, allowing the construction of a spatial database. This referencing means that data (monthly facility records) and unknown values (where facilities have failed to report in a given month) can be placed in a spatial framework, and the task of predicting the unknown values can be considered a spatial modelling problem (Cressie, 1993; Bailey and Gatrell, 1995).

Geostatistics is a field of spatial modelling that incorporates established tools such as kriging for the prediction of unknown values in space from spatially distributed data (Matheron, 1971; Journel and Huijbregts, 1978; Isaaks and Srivastava, 1989; Goovaerts, 1997). Although geostatistics incorporates a wide array of tried-and-tested spatial modelling techniques it is more than simply a 'tool-box' of algorithms to accomplish specific tasks with spatial data. Geostatistics is founded on the *theory of regionalized variables* (Matheron, 1971), a conceptual probabilistic modelling paradigm that centres on the characterisation and exploitation of spatial autocorrelation in the variable of interest. The robust yet flexible nature of the geostatistical paradigm has led to its extension, beyond its original set of core tools, to new approaches that are adapted to suit a wide range of data scenarios and modelling objectives.

This project has adopted a primarily geostatistical approach to the objectives defined above. The strategy throughout has been to extend the application of established geostatistical techniques to other approaches that may be better suited to the characteristics of the HMIS data set and to compare quantitatively the predictive performance of different techniques. The HMIS database consists of monthly records collected at a set of spatially distributed facilities over several years. As such, an alternative to representing these data as a series of independent spatial data sets is to place all data in a single space-time framework. A particular focus of this project has been the use of space-time geostatistical approaches that can utilise simultaneously spatial and temporal autocorrelation in the property of interest in order to predict

unknown values distributed across space and through time. A further focus has been the local characterisation of spatial autocorrelation structures, taking advantage of the large number of data distributed across space with the aim of more accurately characterising the spatial heterogeneity displayed by the HMIS data.

A central theme of this project has been the issue of spatial standardisation. The HMIS data represent monthly counts of malaria diagnoses at each facility. When count data such as these are used in public health modelling settings, the raw data are usually standardised by some measure of the population from which that count was generated (Lawson, 2001). In the current case, each count is influenced by a range of factors specific to each facility such as its size, function, and catchment population, and accounting for these factors may reveal greater spatial structure in the property of interest which may, in turn, allow missing values to be predicted with greater accuracy. Because very little facility-specific information is available across Kenya, a key concern of this project has been to develop novel strategies that allow a degree of spatial standardisation between count data at different facilities. Again, the effect of these different strategies on the ultimate accuracy of predictions of missing values was evaluated.

1.3 Contributions made by the project

This project has resulted in, for the first time, reliable estimates of the annual outpatient treatment burden for malaria at health facilities in the formal government health sector in Kenya at the national, provincial and district levels. These estimates represent a tangible resource to assist evidence-based decision-making for the provision of anti-malarial resources in this under-funded health service. In meeting this primary objective, a series of further contributions have been made by this project and these are summarised below.

HMIS outpatient data on disease-specific counts are generally collected each month from a set of spatially-distributed health facilities, resulting in a space-time data set. This study has shown that such data may display spatial autocorrelation even before any standardisation to account for facility-specific factors and that, where present, this spatial structure can be exploited using established geostatistical techniques to predict

missing values. Furthermore, this study has shown that these data may display substantial temporal, as well as spatial, autocorrelation. By comparing spatial-only and space-time models for geostatistical prediction, this study has shown that, for the case of Kenya, the space-time approach, by exploiting temporal as well as spatial autocorrelation in the data, is able to predict missing values with greater accuracy than the spatial-only approach.

The spatial structure of disease-specific HMIS data is determined, in part, by the spatial pattern of the disease in the population across a country which can often display substantial spatial heterogeneity in both first and second order characteristics, driven by climatic, topographic and demographic factors. Where such spatial characteristics are present, the adoption of a globally-stationary random function (RF) model for geostatistical prediction may be less appropriate than one in which only smaller sub-regions are considered stationary. This study has developed a local approach to space-time geostatistical prediction which enables space-time autocorrelation structures to be estimated and modelled within a spatially-local neighbourhood around each prediction. By comparing the global and local space-time models this study has shown that, for the case of Kenya, the local approach is able to predict missing values with greater accuracy than the global approach. However, this increase in accuracy was modest and came at a substantial price in terms of labour and computational resources.

Public health data in the form of counts are usually standardised by some measure of the population from which the counts were generated. For HMIS data the appropriate denominator is the catchment population of each facility, but such information is not available for facilities across Kenya. One response is to estimate catchment populations from census-derived population maps within a GIS. A simple and widely-used catchment model can be generated by defining Thiessen polygons around each facility. As part of a wider project (led by Abdisalan Noor at the Kenya Medical Research Institute (KEMRI)-University of Oxford-Wellcome Trust Collaborative Programme) to develop catchment models for Kenyan health facilities, this study has developed methods for assessing the assumptions that underpin the use of Thiessen polygons in this context and has shown that, for a set of test facilities in four Kenyan districts, these assumptions are invalid.

In the absence of facility catchment population estimates, an alternative approach to the standardisation of HMIS count data was developed in this study. Data on the total count of all-cause outpatient cases (i.e. not limited to any one disease) at each facility were used to define denominators with which to standardise the disease-specific count data, with the rationale that the total outpatient count acts as a proxy measure of facility catchment populations. In the Kenyan HMIS, however, the use of data on the total count as a denominator was limited by the fact that these values were only available for the same points as the disease-specific count data, meaning that the denominator itself had to be predicted for unsampled locations. By developing and testing an alternative prediction framework that incorporated total count data in this way, this study showed that this approach can substantially increase the spatial autocorrelation in the resulting standardised count data, which can then be predicted using geostatistical techniques with a greater accuracy than the raw count data. However, this study also showed that the uncertainty introduced by the need to predict the denominator at unsampled locations can negate much of the benefit of using a standardised numerator, and a second predictive framework was developed in which this uncertainty was reduced.

By developing and testing a series of geostatistical modelling frameworks that incorporate spatial and space-time techniques, globally-stationary and locally-stationary models, and alternative ways of standardising HMIS count data, this study has been able to compare quantitatively the effect of each modelling strategy and identify that framework that is most suitable for predicting missing malaria outpatient data in the Kenyan HMIS. A stochastic simulation approach was developed and tested that adapted a sequential Gaussian simulation algorithm in order to generate a model of the uncertainty associated with predictions made within the final modelling framework.

In summary, this study has developed and tested geostatistical models that can predict missing values within the Kenyan HMIS data base to an acceptable level of accuracy, with realistic accompanying measures of prediction uncertainty. This study, therefore, serves as an example to other public health practitioners faced with the task of delivering reliable metrics from imperfect HMIS data. The use of such techniques was made possible in Kenya by the existence of a comprehensive georeferenced database of government health facilities. As such, this project also serves as an example of the importance and potential benefit of developing nationwide health service GIS

frameworks to health planning agencies across the developing world.

1.3.1 Publications

The research carried out for this project has resulted (in part or in full) in the following publications (or submissions) and conference presentations.

Peer-reviewed journal papers

Gething, P.W., Noor, A.M., Zurovac, D., Atkinson, P.M., Hay, S.I., Nixon, M.S., and Snow, R.W., 2004. Empirical modelling of government health service use by children with fevers in Kenya. *Acta Tropica*, 91, 227-237.

Gething, P.W., Noor, A.M., Gikandi, P.W., Ogara, E., Hay, S.I., Nixon, M.S., Snow, R.W., and Atkinson, P.M., 2006. Improving data from imperfect health management information systems in Africa using space-time geostatistics. *PLoS Medicine*, 3.

Gething, P.W., Noor, A.M., Gikandi, P.W., Hay, S.I., Nixon, M.S., Snow, R.W., and Atkinson, P.M., 2006. Developing geostatistical space-time models to predict malaria outpatient treatment burdens in Kenya. Submitted to *Geographical Analysis*, under review.

Gething, P.W., Atkinson, P.M., Noor, A.M., Gikandi, P.W., Hay, S.I., and Nixon, M.S., 2006. A local space-time kriging approach applied to a national outpatient malaria data set. *Computers & Geosciences*, under review.

Noor, A.M., Amin, A.A., **Gething, P.W.**, Atkinson, P.M., Hay, S.I., and Snow, R.W., 2006. Modelling distances travelled to government health services in Kenya. *Tropical Medicine and International Health*, 11, 188-196.

Conference presentations

Empirical modelling of Kenyan health facility catchments as an aid to predicting malaria risk. *Epidemiology: a spatial perspective 2003*, Salford, UK, 2003.

Optimising the utility of national malaria data for health system planning in Kenya using spatiotemporal analysis. *Geostats UK 2005,* Belfast, UK, 2005.

Optimising the utility of national malaria data for health system planning in Kenya using spatiotemporal analysis. *Workshop on Recent Advances in Modelling Spatio-temporal Data*, Southampton Statistical Sciences Research Institute, Southampton, UK, 2005.

1.4 Thesis Outline

Chapter 2 provides detailed information on specific issues that provide a contextual backdrop to the study including the current status of health information in low-income countries and the role of HMIS in providing important health information. Relevant information is given on Kenya and its health system, how the Kenyan HMIS operates within this system, and the particular relevance of malaria as a public health problem. Chapter 3 presents the two main data sets on which this project was based: routine malaria outpatient data from the Kenyan HMIS, and a corresponding spatial database of health facilities. The construction and compilation of these data sets are described and exploratory analysis is presented that describes the broad spatial and temporal characteristics of the outpatient data set. Analysis is also included that examines the extent and patterns of missing data. Chapter 4 describes the principal established methods used in this project starting with an overview of the conceptual underpinnings of geostatistics, and the key concepts and tools by which the approach can be used to characterise and predict spatial phenomena. The extension of spatial-only geostatistical techniques to space-time settings is also discussed and a brief review is included of the use of geostatistical methods in public health and malaria settings. Chapter 5 presents the conceptual framework that was developed in this project to meet the aims stated in this chapter. The various factors that are likely to influence malaria treatment burdens at different facilities are identified and the way in which these are likely to vary in space

and time is discussed. In light of this discussion two distinct modelling strategies are proposed and the rationale for these is explained. The first involves the prediction of catchment populations for facilities across Kenya, and this work is presented in Chapter 6. The second involves the inclusion of a second outpatient variable from the HMIS in a geostatistical prediction framework and two such frameworks are developed and tested. Chapter 7 addresses different techniques for geostatistical prediction and implements a comparison of spatial-only and space-time approaches. Globally stationary and locally-varying models are also compared. In Chapter 8, the different prediction frameworks are implemented to predict missing malaria cases (MC) for a test data set and their predictive accuracies are compared. An uncertainty model is also developed that provides model-based measures of prediction uncertainty using an adapted space-time sequential Gaussian simulation technique adapted to represent the modelling framework used. In Chapter 9, the final modelling framework is implemented using the full data set to predict malaria treatment burdens across Kenya. An empirical model validation approach is also developed to provide estimates of the expected prediction errors at different levels of spatial and temporal aggregation. Chapter 10 is a discussion chapter that considers some of the most important issues that have arisen during the project, and looks ahead to future avenues for research that have resulted. The thesis ends with a brief conclusion in Chapter 11.

Chapter 2

Background

Chapter 2

2. Background

2.1 Introduction

Having laid out the motivation for this study and its specific objectives in Chapter 1, this chapter is designed to offer detailed information on specific issues that provides a comprehensive backdrop to the problem addressed. Information is given across a spectrum of detail, from a general description of the health information field in which this study is best categorised to a specific account of the niche that it aims to fill: the use of incomplete routine outpatient data from HMIS to estimate malaria treatment burdens in Kenya. The over-arching issue of inadequate health information in low-income countries is introduced first, and the importance, current status, and potential of HMIS to provide vital information is discussed. An account is given of the provision of health care in Kenya, and how the Kenyan HMIS currently contributes to the functionality of the health system. A summary of the burden of malaria in Kenya is provided and the importance of producing reliable estimates of treatment burdens is explained, along with the pivotal role of HMIS in providing such estimates.

2.2 Health information in low-income countries

2.2.1 Requirements for health information

The MDGs have brought into sharp focus the need for wholesale improvements in the availability of reliable and timely health information as part of international efforts to strengthen health systems in low-income countries (WHO, 2000a; Murray et al., 2004;

AbouZahr and Boerma, 2005; Macfarlane, 2005; UN, 2006; World Economic Forum, 2006). The need to improve the evidence base for public health decision-making (PARIS21, 2004; Scott, 2005) is reflected in numerous established and emerging programmes from international health and donor agencies such as the World Health Organisation (WHO) (WHO, 2000a, 2000b) and the Roll Back Malaria partnership (RBM) (RBM, 2000), and the launch of new initiatives such as the Ellison Institute (Murray et al., 2004; Horton, 2005), the Health Metrics Network, and the Partnership in Statistics for Development in the 21st Century (PARIS21).

Health information is required in a multitude of forms to meet the diverse requirements of different regional, governmental, and international public health actors. Information is required on the health status of populations, the characteristics of the health system and services available, and the efficacy of the health system in benefiting public health. The list of indicators included in the MDGs (UN, 2001) serves to highlight those public health metrics considered most important to the international development agenda. These include both population health indicators (such as infant and child mortality, maternal mortality and HIV prevalence, and the prevalence and death rates associated with malaria and tuberculosis) and health service indicators (such as child immunization coverage, the provision of obstetric care, the coverage of malaria preventative and curative measures, the detection and appropriate treatment of tuberculosis, and the proportion of the population with sustainable access to essential drugs).

Whilst the establishment of a small number of clearly defined public health indicators within the MDGs is a powerful tool to drive global public health policies, the implementation of inclusive and effective national-level health systems requires more comprehensive information on the full spectrum of public health challenges and the type, quality, and coverage of services that are provided (AbouZahr and Boerma, 2005). Furthermore, these data are required at finer spatial and temporal resolutions in order to identify intra-national discrepancies in public health status or service provision to allow targeted public health programmes and the allocation of resources, and to drive decentralized policy making. In addition to the monitoring and evaluation of population health and service provision, effective health system management requires comprehensive, timely, and reliable information on the demand for services (e.g. the number of treatments being administered for a given condition) and the status and flux

of human and financial resources, physical assets, and logistics supply (Health Metrics Network, 2005c).

2.2.2 Sources of health information

The mechanisms and resources for generating and archiving health data are generally substantially under-developed in low-income countries compared to the systems in place in higher-income nations. Nevertheless, large numbers of health data are collected each year in low-income countries from a range of sources and by a variety of agencies addressing different information needs. A useful distinction is between active and passive data collection. Active data collection involves proactive methods of obtaining health data such as surveys and surveillance systems and is usually motivated by specific information requirements. Passive collection refers to systems that record data on a routine basis such as the collection of patient records at health facilities, and is often motivated only by a general, and often nebulous, requirement for health information. Both these approaches can provide information on service provision and the health status of the population, although each has inherent strengths and weaknesses in meeting specific information needs.

Active population-based mechanisms for generating health data are implemented in many low-income countries worldwide and include household surveys and surveillance programmes. Population surveillance and survey efforts are becoming increasingly internationally coordinated (Health Metrics Network, 2005d). Three of the most prominent international household survey initiatives are the Demographic and Health Survey (DHS), the Multiple Indicator Cluster Survey (MICS), and the World Health Survey (WHS). DHS is funded largely by the United States Agency for International Development (USAID) and is the largest international survey programme, having been implemented in over 75 countries since its inception in 1984. DHS surveys are implemented approximately every five years in the target countries and consist of nationally representative household surveys with large samples of up to 30,000 households providing population and health data on a wide range of indicators. MICS is a UNICEF initiative developed in 1994 to provide household survey data on a range of child health and development indicators (UNICEF, 1999). MICS surveys have been

carried out in over 60 countries during 1995 and again during 2000, with a further round currently in progress. The WHS is a recent WHO initiative that aims to enlist Ministries of Health around the world to implement a standardised module-based population survey to collect data on population health status, risk factors, and the responsiveness, coverage, access, utilisation, and cost of health services. These three internationally coordinated population-based survey programmes are designed to provide statistically sound, internationally comparable estimates of key population health indicators and have been developed or adapted largely with the aim of monitoring the health-related targets set out in the MDGs.

A further source of actively-collected population health data is provided by longitudinal sentinel surveillance studies. In contrast to the national snap-shot provided by individual household surveys such as DHS and MICS, longitudinal sentinel surveillance studies are ongoing efforts that aim to monitor prospectively at regular intervals the same set of individuals within a sample community in order to better assess health status and interventions through time. As with household surveys, such efforts are becoming internationally coordinated to enhance cross-study comparability, principally through the INDEPTH Network Demographic Surveillance Systems (DSS) which currently incorporates 31 surveillance sites in 17 low-income countries (IDRC, 2002; Sankoh et al., 2004).

The substantial international investment that has allowed the establishment of coordinated active data-collection programmes such as DHS, MICS, WHS and DSS has led to tangible increases in the availability and quality of population-based health data with which to monitor international targets such as the MDGs. Whilst such data are widely used by national policy-makers, their utility is limited by their spatial and temporal coverage and resolution. National household surveys such as DHS are designed to be nationally representative and data are generally available at no finer level of spatial disaggregation than the first-level administrative unit (usually the province). This severely limits their use for decentralized (e.g. district level) decision-making. Furthermore, the large duration between repeat surveys (~ 5 years) limits the scale of temporal patterns that can be monitored. Surveillance studies such as DSS represent, in some respects, the inverse situation, providing fine resolution spatial and temporal coverage but only over a small local area that is not necessarily nationally

representative. Whilst surveillance studies often collect high-quality data on a rich array of health variables, national household surveys are generally limited to a relatively small set of variables.

National health systems cannot rely on active population-based survey and surveillance data alone to meet their health information requirements. If limited health system funds are to be used efficiently to maximise the efficacy of service delivery in low-income countries, health systems require reliable data on a much more comprehensive suite of population health and service variables at substantially finer spatial and temporal resolutions. Inclusive information on health service demand, provision, and functionality can only be feasibly supplied from within the health system itself, using passively collected data obtained on a routine basis by health system facilities and practitioners. Two data collection mechanisms that often constitute the backbone of information systems within low-income health services are vital registration (VR) systems and health management information systems (HMIS). VR systems are designed to provide the most fundamental population metrics such as births, deaths, and, crucially, the causes of death, ideally based on an internationally standardised classification procedure such as the International Classification of Diseases (ICD) (WHO, 2004b). In low-income countries, VR systems generally provide only an incomplete and fragmented record of vital events and are often based on unreliable methods for determining cause of death such as verbal autopsy (Snow et al., 1992b; Mahapatra and Chalapati Rao, 2001; Morris et al., 2003; Silvi, 2003; Sibai, 2004; Mathers et al., 2005). HMIS often perform a number of functions but their core role is generally to coordinate the routine collection and collation of facility-based records of morbidity and mortality along with management and financial data. As discussed in Chapter 1, data from HMIS are the principal focus of this project and these systems are now discussed in more detail.

2.3 HMIS in Africa

2.3.1 Definitions and functionality

The term *health information system* (HIS) is often used in the broadest sense to encompass all data collection instruments, actors, resources, and institutions involved in

the collection of health data at the national level (Üstün et al., 2003). Different Ministries of Health define different information subsystems as being either components of, or external to, their national HIS. Furthermore, other countries use the term *health management information system* which tends to be used more specifically to refer to a well defined information system operated by a Ministry of Health to coordinate routine data collection. In some cases, a formal restructuring has been implemented in which an HIS has been modified and relaunched as an HMIS. This adjustment has often been motivated by a desire to refocus data collection and use, often in line with a policy of health service decentralisation and with an emphasis on managers utilizing health information at the facility and district level (MoH Kenya, 2000; Gladwin et al., 2002; Mutemwa, 2006). For the purposes of this discussion, the term HMIS will be used exclusively, accepting that in some cases HIS and HMIS are one and the same entity, whilst in others they are quite distinct.

Definitions of HMIS abound and rarely correspond exactly (Health Metrics Network, 2005c). If active population-based data collection mechanisms can be reasonably excluded from the definition then the core function of HMIS can be stated as the provision of routine facility-based data. In a fully-functional HMIS, these routine facility data provide both service and management information. Service information relates to both supply (the services available at each facility, and the quality, capacity, and coverage of service provision) and demand (records of service use, the numbers of inpatients and outpatients being treated and what they are being treated for). Management information relates to the wide range of ancillary data needed for the efficient planning, monitoring, and implementation of health service delivery. Information is required on human resources (e.g. health personnel, staffing levels and turnover, ratio of population to health workers), finance (e.g. disbursement and expenditures, efficiency monitoring, annual budgets and accounts, ratio of population to expenditure, expenditure by the public), physical assets (e.g. records of capital investment, buildings and equipment supply, status, maintenance requirements and life span), and logistics (e.g. data to estimate requirements of drugs, vaccines, contraceptives and other essential medical supplies).

Some form of HMIS is operated by most Ministries of Health across sub-Saharan Africa and elsewhere in the developing world. In the overwhelming majority of low-income

nations, however, these systems are considered to be severely underperforming or failing completely, and unable to meet basic health information requirements (Keller, 1991; Avgerou, 1993; Cibulskis and Hiawalyer, 2002; Gladwin et al., 2002, 2003; Littlejohns et al., 2003; Chaulagai et al., 2005; Mutemwa, 2006). Data from HMIS are seen as fragmentary and biased by both policy-makers and the academic community. This scepticism about the quality of HMIS data has led to a gross disparity between the level of resources that are invested in their generation and the extent to which they are used as a basis for either research or decision making. This disparity can only be reduced by improving the reliability of information that can be obtained from HMIS. Whilst wholesale improvements in HMIS infrastructure must remain the long-term goal, the reliability and, hence, utility of current HMIS data can be improved using statistical modelling, and it is this rationale that has motivated the current project.

2.3.2 Limitations of HMIS outpatient data in Africa

The focus of this project is on HMIS data as a means of quantifying treatment burdens for a given condition, using the example of malaria in Kenya. The metric of interest is the number of outpatients that are treated for the disease in a given month or year in facilities around the country. Such information is best supplied from routine outpatient records. A principal function of all HMIS is the collection and transfer of these records from all facilities through district and provincial levels and their ultimate collation into a national database. Whilst the problems and deficiencies in HMIS-generated outpatient data vary between countries, a common set of limitations can be identified that restrict the utility of these data for the delivery of information to health decision-makers that is reliable, accurate and representative. These generic limitations can be divided into those internal to HMIS – failings or weaknesses at specific points within the system – and those external to it – operating outside the system but reducing the value of the information that can be obtained.

2.3.2.1 Internal constraints

In an effective HMIS, routine outpatient data are based on accurate diagnosis and comprehensive registration of cases at the facility level and consistent reporting of these

records (usually monthly) to the next level in the administrative framework. Numerous studies have been undertaken across Africa to assess HMIS functionality, along with others that have focused on related clinical practices. Many health facilities, especially those in peripheral rural areas, have limited or no access to laboratory facilities for the analysis of samples with which to confirm diagnoses of malaria and other prevalent communicable diseases. In a study of 81 government health facilities in four districts of Kenya, 42% had access to a functional microscope (Zurovac et al., 2002). In a study in Ethiopia, 53% of sampled facilities had the capacity for laboratory confirmation of malaria, although this was as low as 33% for health centres alone (WHO, 2001b). A Ugandan study reported that 51% of facilities had the capacity for laboratory confirmation of malaria, 44% for tuberculosis, and 21% for meningococcal meningitis (CDC, 2000). Furthermore, even where laboratory tests are available, they are often inaccurate (Zurovac et al., 2006), and are often ignored in diagnoses (Barat et al., 1999). In the absence of laboratory facilities, diagnoses for malaria and other diseases are generally based on clinical examination. Although efforts have been made by the WHO and others to provide guidelines and diagnostic algorithms to assist clinical diagnosis, increases in diagnostic accuracy have been limited (Redd et al., 1992, 1996; Smith et al., 1994). This is partly due to the overlap of symptoms displayed by common communicable diseases. The symptoms of malaria can overlap with those of pneumonia, hepatitis, influenza, viral encephalitis, haemorrhagic fever and meningitis, among others (Warrell, 2002; Kallander et al., 2004). In practice, presumptive diagnosis and treatment of all fevers as malaria is the norm in many malarious areas (Bloland et al., 2003). The WHO Integrated Management of Childhood Illnesses (IMCI) programme, developed to improve clinical diagnosis and treatment practices in areas with limited laboratory access, advocates that all febrile children in high risk areas be considered to have, and be treated for, malaria (WHO, 1997; Perkins et al., 1998; Gove, 1998). Various studies have retrospectively tested clinically diagnosed malaria cases using laboratory techniques to asses the accuracy of clinical diagnosis for outpatients. Zurovac et al. (2002) found that, although sensitivity (the proportion of patients with malaria who were correctly diagnosed as such) exceeded 90% in four Kenyan sentinel districts, specificity (the proportion of patients without malaria who were correctly diagnosed as such) was much lower (39%). A study of peripheral health facilities in one district of Tanzania reported a positive predictive value (PPV – the proportion of positive diagnoses that were correct) of clinical diagnosis of 44% across all ages (Font et al., 2001). A study in

rural Mozambique found a PPV of 89% (Loveridge et al., 2003). The poor availability of accurate diagnostic methods and the institutionalized presumptive treatment of fevers as malaria means that there is likely to be a substantial discrepancy between HMIS outpatient records of diagnosis and treatment for malaria and the true extent of outpatient malaria morbidity.

A further widespread and fundamental limitation of HMIS in Africa is the extensive under- and non-reporting of routine data from health facilities (Al Laham et al., 2001; MoH Kenya, 2001a; Rudan et al., 2005; Health Metrics Network, 2005b). Reasons for under-reporting include resource constraints, such as a shortages of outpatient registers and reporting forms, limited means of sending data, and lack of staff training (WHO, 2001b; Chilundo et al., 2004; Health Metrics Network, 2005c). Furthermore, whilst time-demands on front-line health workers for data recording and reporting are often considerable (Braa et al., 1997), feedback of HMIS information from higher levels to peripheral facilities is usually limited. This inequity can lead to diminished motivation for busy health workers to commit time to data recording and reporting. Widespread under-reporting inevitably leads to substantial gaps in the national HMIS database which prevents the straightforward quantification of basic outpatient health and service metrics.

2.3.2.2 External constraints

The principal external constraint on the utility of routine outpatient data for assessing public health is the under-utilisation of health facilities by the community. Factors such as the high cost of treatment, poor access, and inadequate service delivery have led to low utilisation rates of formal health services across Africa (Fosu, 1994; Oranga and Nordberg, 1995; Mwenesi et al., 1995; Foster, 1995; Ryan, 1998; Molyneux et al., 1999, 2002). Low utilisation means that a significant proportion of morbidity due to communicable disease is never presented to the formal health sector and is therefore not included in HMIS data. Many African studies have investigated the behaviour of those seeking care for communicable diseases and a wide divergence in attendance rates has been reported (McCombie, 1996, 2002). Studies in Kenya that have investigated treatment-seeking for malaria, or malaria-like fevers, have found attendance rates at formal facilities of between 18% and 43% (Ruebush et al., 1995; Hamel et al., 2001;

Amin et al., 2003; Guyatt and Snow, 2004). A study of fatal malaria cases in Tanzania found that 65% had presented at government or private health facilities (de Savigny et al., 2004) whilst a large household survey in rural Ethiopia reported attendance of 80% for suspected malaria cases (Deressa et al., 2003). The most comprehensive picture of attendance rates to formal health services in low-income countries is provided by DHS surveys. A standard module of these surveys includes the percentage of children with symptoms of acute respiratory illness (ARI) or fever that were taken to a formal health

Table 2.1 DHS estimates of national attendance rates of children with ARI or fever to formal health facilities in nine east African countries.

Country	DHS Survey year	Child attendance rate (%)*
Eritrea	2002[1]	43
Kenya	2003[2]	45
Malawi	2004[3]	20
Mozambique	2003[4]	55
Rwanda	2000[5]	15
Tanzania	2004-2005[6]	57
Uganda	2000-2001[7]	65
Zambia	2001-2002[8]	69
Zimbabwe	1999[9]	50

*This variable is obtained in the DHS surveys as the percentage of children under five years who had a cough accompanied by short, rapid breathing (symptoms of ARI) and / or fever in the two weeks preceding the survey for whom treatment was sought from a formal health facility or provider. 1. (NSEO Eritrea and ORC Macro, 2003); 2. (CBS Kenya and ORC Macro, 2003); 3. (NSO Malawi and ORC Macro, 2005); 4. (INE Moçamique and ORC Macro, 2003); 5. (ONAPO Rwanda and ORC Macro, 2001); 6. (NBS Tanzania and ORC Macro, 2005); 7. (UBOS Uganda and ORC Macro, 2001); 8. (CSO Zambia and ORC Macro, 2003); 9. (CSO Zimbabwe and ORC Macro, 2006).

service provider. Table 2.1 lists these attendance rates from the most recent DHS surveys for nine East African countries. Rates ranged from 15.1% in Rwanda (ONAPO Rwanda and ORC Macro, 2001) to 69.1% in Zambia (CSO Zambia and ORC Macro, 2003). The causes and extent of under-utilisation vary considerably both within and between African countries. Treatment seeking decisions at the household level are influenced by a diverse range of factors including physical and economic access, social and cultural beliefs and values, the actual or perceived quality of care offered by formal sector facilities, and the range of alternative treatment options available. In the above DHS surveys, the most frequently cited reasons given by mothers for non-attendance included lack of money to pay for treatment, the distance to the health facility, lack of availability or means to pay for transport to the facility, and previous experience of lengthy queues for treatment once at the facility. The unavailability of suitable drugs at health facilities due to stock-outs is also widely cited as a reason for non-attendance.

2.3.2.3 Implications

Despite decades of donor assistance in HMIS across Africa, and a growing awareness of their importance to health system delivery, tangible progress in increasing data availability and quality has generally been slow (WHO, 1993; Lippeveld et al., 2000; Evans and Stansfield, 2003; WHO, 2004a; AbouZahr and Boerma, 2005; Health Metrics Network, 2005c). The internal and external constraints considered here present a challenge to users of routine outpatient data. If due consideration is not given to these limitations, there is a potential for data to be misinterpreted and misleading conclusions to be made. A critical realisation is that the spatiotemporal pattern of outpatient cases of malaria and other diseases recorded in the database does not represent directly the underlying pattern of the disease in the community. Instead, the relationship is characterised by uncertainty introduced by the limitations discussed above. This uncertainty leads to the potential for misapplication of HMIS outpatient data. The crude use of such data to estimate the total burden of a given illness in the population, for example, is likely to result in gross under-estimation, even after missing data are taken into account, due to the extensive under-utilisation of government facilities by care-seekers. Similarly, attempts to infer the relative prevalence of different diseases in a population should be treated with caution due not only to differential utilisation patterns but also to the unreliability of diagnosis. Even greater uncertainty is associated with the relationship between outpatient morbidity and the transmission dynamics of a given communicable disease (Snow et al., 1997) and, as such, attempts to link the two should focus on assessing the characteristics and extent of this uncertainty.

Whilst factors such as under-utilisation and misdiagnosis reduce the suitability of outpatient data to provide information on morbidity in a population, they are less obstructive when the data are to be used to analyse within-system treatment patterns and resource requirements. The number of outpatient treatments administered for a given condition within a health system is partly determined by the proportion of those afflicted in the population who attend health facilities for treatment, and by the proportion of those attending who are correctly diagnosed. As such, it is not necessary to correct for the effects of misdiagnosis and under-utilisation when using HMIS outpatient data to estimate the treatment burden for a given condition, since these factors contribute to defining that burden. Widespread under-reporting of outpatient records by health

facilities, however, is a more serious limitation to the use of HMIS data for estimating treatment burdens. A complete HMIS database allows the straightforward quantification of the number of patients treated for each disease at each facility each month, allowing the comprehensive assessment of specific resource requirements. Where large proportions of data are missing each month, however, and the number and location of missing data varies between months, an HMIS database cannot provide such quantifications directly. The incompleteness of HMIS databases contributes substantially to their under-use by health system decision-makers.

The remainder of this chapter considers the provision of health care in Kenya and presents details of the Kenyan HMIS. The burden of malaria is then discussed along with the importance of defining antimalarial drug demand. The potential role of HMIS data in estimating drug demand is then examined.

2.4 Kenya and its health system

2.4.1 Kenya country profile

Kenya straddles the equator and is situated on the eastern coast of the African continent with borders to Tanzania to the south, Uganda to the west, Ethiopia and Sudan to the north, Somalia to the northeast, and the Indian Ocean to the southeast (Figure 2.1). The country is divided into 8 provinces and 72 districts and has a land area of 571,466 square km of which the majority (~ 80%) is arid or semi-arid. Table 2.2 contains summary information for each province. The bulk of Kenya's 30 million people live in the extensive highland region that makes up the country's south west quadrant (Figure 2.2). This fertile elevated plateau stretches from Lake Victoria in the west to the lowlands in the east and is bisected by the Great Rift Valley running approximately north-south. The most densely populated regions are centred around the shores of Lake Victoria and around the capital, Nairobi. A further region of dense population is found along the Indian Ocean coastline, incorporating the country's principal port city, Mombassa. Average temperatures and rainfall vary considerably across Kenya due to the varying altitude and proximity to the lakes or ocean. A marked seasonal pattern is evident with a short dry season from January to March, a long rainy season from March to May, a long

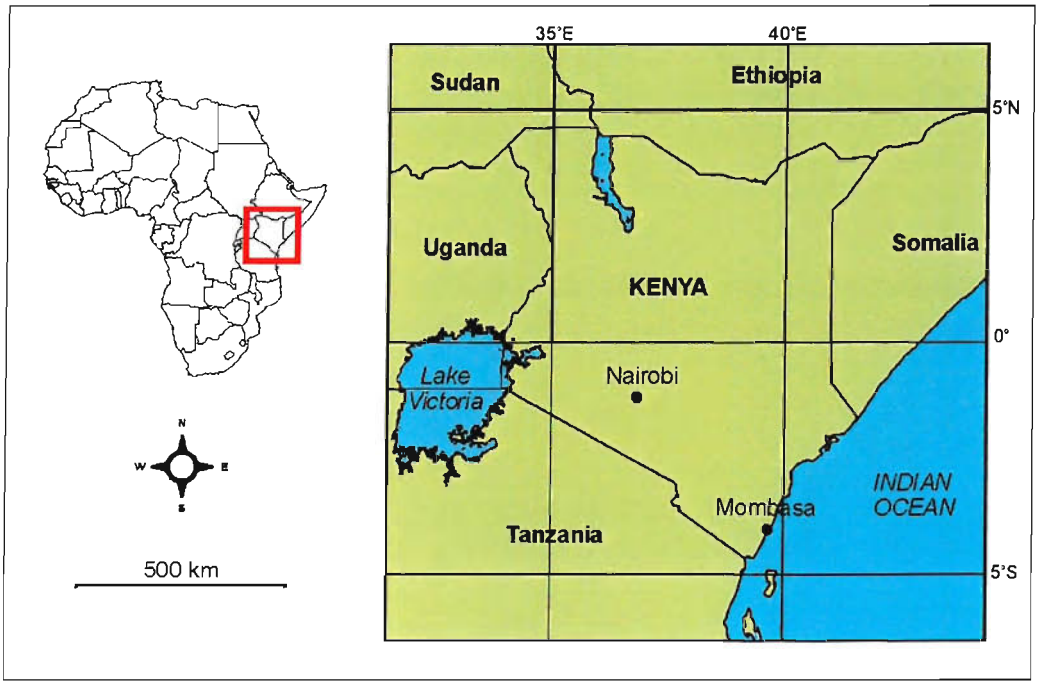


Figure 2.1 Overview map showing the position of Kenya and its neighbours on the east of the African continent.

dry season from May to October, and a further short rainy season from October to December. Kenya has endured decades of poor economic performance and slow economic growth that has contributed to an overall deterioration in the welfare of the population. Government estimates state that around 56% of the population were living in poverty in 2003, and that the proportion living below the poverty level has steadily increased (CBS Kenya, 2003). Increasing poverty has gone hand in hand with rising

Table 2.2 Summary information for the Kenyan provinces: population and land area.

Province	Population*	Area (Sq. km)
Central	3,724,159	13,191
Coast	2,487,264	83,603
Eastern	4,631,779	159,891
Nairobi	2,143,254	684
North-Eastern	962,143	126,902
Nyanza	4,392,196	16,162
Rift Valley	6,987,036	173,854
Western	3,358,776	8,361
KENYA	28,686,607	582,648

* Population as recorded in the 1999 census (CBS Kenya, 2001a)

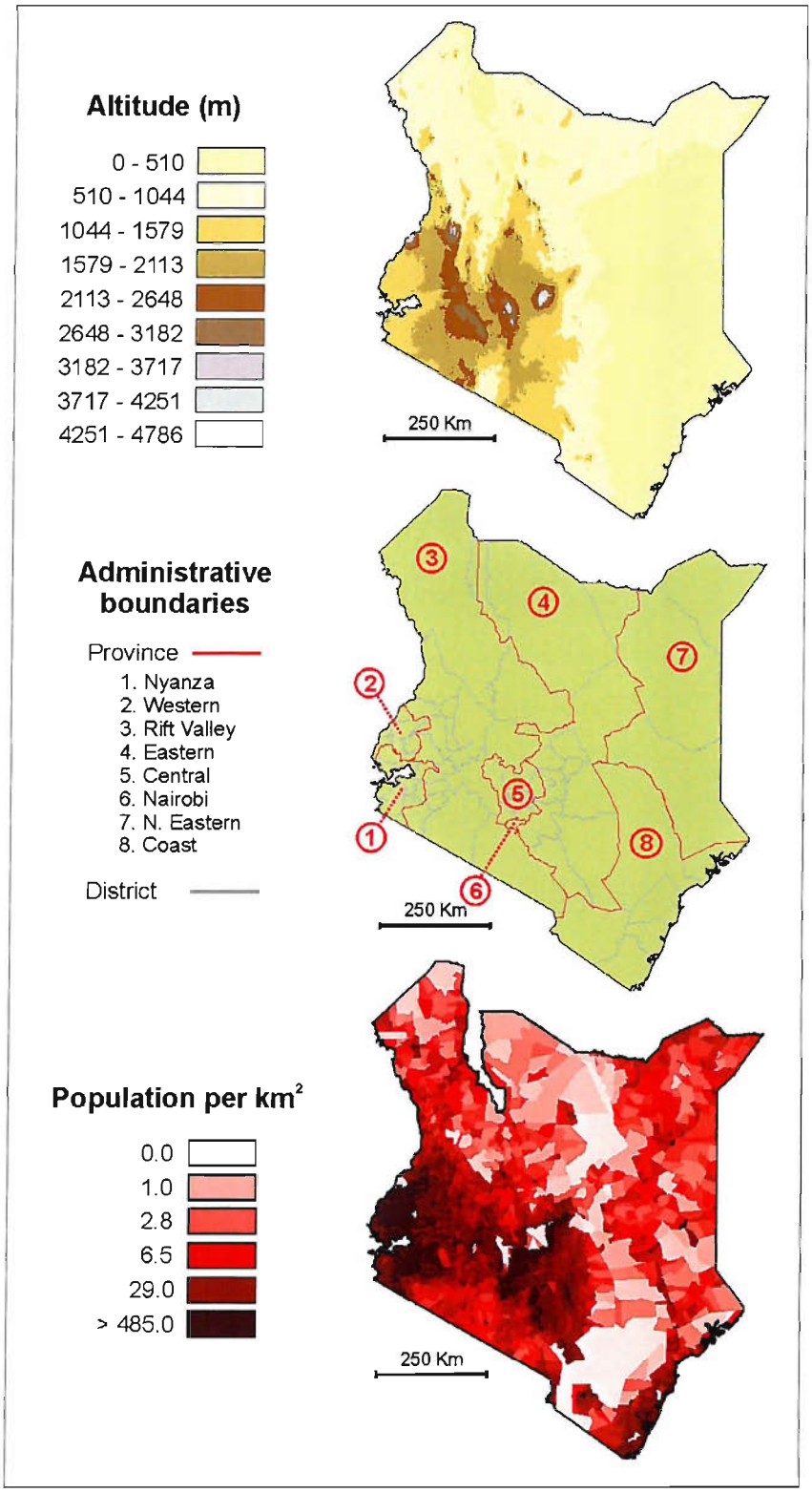


Figure 2.2 Maps of Kenya showing: (top) altitude in metres; (middle) district and provincial administrative boundaries, and province names; and (bottom) population density per km^2 in each sub-location (5^{th} level administrative level) based on the 1999 census.

unemployment and illiteracy rates and the concomitant decline in living standards has been reflected in worsening public health. Having shown signs of improvement during the 1970s and 1980s, the crude death rate and infant mortality rate increased during the 1990s whilst life expectancy declined (CBS Kenya and ORC Macro, 2003).

2.4.2 Health service provision in Kenya

The provision of health services to the Kenyan population is implemented by various governmental and non-governmental organisations and is delivered through a hierarchical administrative system that coordinates multiple levels of service delivery (MoH Kenya, 2003). The organisation of formal health care delivery through the Ministry of Health operates at three levels: national, provincial, and district. The national level centres on the Ministry of Health headquarters, housing the Central Board of Health. The provincial level incorporates the Provincial Health Management Boards and Provincial Health Management Teams and acts as an intermediary between the national and district levels, overseeing district level health policy and quality standards and coordinating district level provision. The district level comprises District Health Management Teams that concentrate on the delivery of health services and generate their own expenditure and budgetary plans within the provincial and national framework. Whilst the Ministry of Health is the primary source of health care, operating around 52% of health facilities nationwide, a substantial proportion of service provision comes from other service providers. Non-governmental and charitable organisations, including the religious missions, mostly operate services in underserved, often rural, areas. They provide both curative and preventative services and receive some governmental support as well as income from external donors and user fees. Around 40% of the country's health services are offered by the private-for-profit sector, including clinics and hospitals that generally specialise in curative services with limited preventative services being offered (MoH Kenya, 2003; NCAPD/MOH/CBS Kenya and ORC Macro, 2005).

The network of health facilities operated by the providers listed above are themselves organised in a hierarchical framework. The most basic and numerous facilities are the dispensaries, followed by sub-health centres, health centres, sub-district hospitals,

district hospitals and provincial hospitals. At the apex are the nation's two teaching and referral hospitals. Dispensaries are generally the first point of contact with patients and are staffed by enrolled nurses and medical assistants providing services such as antenatal care as well as basic outpatient curative care. Health centres offer a wider range of curative and preventative outpatient services including minor surgical procedures, and are staffed by midwives or nurses, clinical officers and occasionally by doctors. District hospitals act as the first referral points for health centres and dispensaries within each district and offer 24 hour inpatient and outpatient care in a range of clinical services backed up by laboratory and other technical support. Provincial hospitals act as the next level of referral, offering specialised care not available at district hospitals including intensive care services. They also provide supervision, monitoring, and technical assistance to the district hospitals. Kenyatta National Hospital in Nairobi and the Moi Teaching and Referral Hospital, 320 km to the north-west of the capital in the Rift Valley province, are Kenya's centres of excellence providing complex health care and highly skilled personnel and representing a high concentration of the nation's health care resources.

Despite a series of major policy initiatives (MoH Kenya, 1999, 2005b), the Kenyan health service remains inefficient and public health status remains poor. A series of indicators point to a worsening level of health care provision over the past two decades. The doctor-to-population ratio declined during the 1980s and 1990s and the overall use of public health services also declined, falling from 0.6 new consultations per person in 1990 to 0.4 in 1996 (CBS Kenya, 2003). Although public-sector spending on health has increased in absolute terms, it has not kept up with population growth such that public per-capita health expenditure fell from US$12 in 1990 to US$6 in 2002 (NCAPD/MOH/CBS Kenya and ORC Macro, 2005).

2.4.3 The Kenyan health management information system

Each level of the Kenyan health system hierarchy described above requires health information at different spatial and temporal scales to assist planning and decision-making. The need for a national mechanism to collect health data was recognised at an early stage by the Ministry of Health and they established the Division of Health

Information Systems in 1974, charged with the responsibility of collecting, processing, analysing and disseminating health and health-related data. This system collected epidemiological data (inpatient and outpatient records) and produced annual bulletins containing data from the facilities. The system was revised and partly computerized through USAID funding during the 1980s. An assessment of this system was made as part of the 1994 Kenya Health Policy Framework (MoH Kenya, 1994) which led to the establishment in 1999 of the Division of Health Management Information Systems in an attempt to address shortcomings and provide more reliable information with the broader scope needed for planning, budgeting, monitoring, and evaluation at all levels (MoH Kenya, 2000).

A primary function of the current Kenyan HMIS is the coordination of routine outpatient data collection, collation, and analysis. In principle, each health facility makes a record of each outpatient visit and the resulting diagnosis or diagnoses that were made. These data are compiled onto a standard form at each health facility, and completed forms are passed each month on to District Medical Records Officers based at district hospitals. At the district level, data from each facility within the district are collated into a District Outpatient Morbidity Summary and sent through the hierarchy to the national HMIS headquarters at the Ministry of Health, where all data are received by Medical Records Technicians who order, collate, and check data before entering them into the national database operating on a rudimentary computer system. This database lists outpatient counts for each health facility under a suite of diagnostic categories. National outpatient data are made available publicly in HMIS reports that are published at intervals of approximately four years.

The most recent major health policy initiative in Kenya is the National Health Sector Strategic Plan for the period 2005-2010 (MoH Kenya, 2005a). This policy document identifies the continuing inadequacy of the HMIS to provide the health information that is required, and makes the improvement of the system a priority. The principal limitation of the HMIS as a source of routine outpatient data is the low reporting rate of monthly outpatient records by both facilities and districts. The Kenyan HMIS report for the 1996-1999 period states an overall monthly reporting rate of between 33% and 40% over all facilities (MoH Kenya, 2001a). Because the number and identity of health facilities that report each month changes, statistics on the total count of cases of each disease seen

each month are not a reliable way of estimating resource requirements, or of analysing differences over time or between different regions. Because of this inadequacy, the use of HMIS data for monitoring and evaluation, problem-solving, decision-making, and trend analysis is extremely limited at national, provincial, and district levels.

2.5 The burden of Malaria

Malaria is a life-threatening infectious disease caused by protozoan parasites of the genus Plasmodium. The vector for the human form of malaria is the female Anopheles mosquito which transmits the parasite to the human host during a blood meal and forms part of the complex parasite life-cycle. The symptoms of malaria typically appear 9 to 14 days after infection and can initially include fever, headache and vomiting. If left untreated, the infection can progress rapidly to become life-threatening, by destroying red blood cells (anaemia) and by clogging the capillaries that carry blood to the brain (cerebral malaria) or other vital organs. Young children and pregnant women are particularly at risk of the disease.

Malaria is found throughout the tropical and sub-tropical regions of the world and remains a leading global cause of morbidity and mortality (WHO, 2005a), with an estimated 300-660 million clinical cases occurring annually worldwide (Snow et al., 2005). Africa, particularly the sub-Saharan region, bears a grossly disproportionate proportion of the worldwide burden of malaria with two thirds of its population estimated to be at risk of the disease (Hay et al., 2004). Africa is estimated to account for two thirds of the overall clinical burden, and three quarters of that caused by *Plasmodium falciparum*, the most severe and life-threatening malaria parasite (Korenromp, 2004). Africa is also thought to account for 89% of the global mortality burden (WHO, 2003b) with between three-quarters to one million African children under the age of five dying each year (Snow et al., 1999a, 2003a). That the continent bears the brunt of the worldwide malaria burden is due partly to the prevalence of the most potent malaria parasite and vector species and partly to the limited economic and public health resources to combat the disease effectively.

2.5.1 The burden of malaria in Kenya

Malaria is a major cause of morbidity and mortality in Kenya and presents a substantial barrier to development with around 20 million Kenyans living at risk of the disease. Although current estimates are of unknown accuracy, recent figures state that, each year, an estimated 145,000 children under the age of five are admitted to hospital due to malaria and 34,000 children of this age die of the disease (MoH Kenya, 2001b). Malaria is the leading cause of outpatient cases and places a heavy burden on the health system and public health expenditure. Personal expenditure due to malaria is also high with a recent survey estimating that every affected household spends around US$20 each year on malaria treatment (CBS Kenya, 2001b), which represents a substantial financial burden to a large proportion of the population. The morbidity associated with malaria has a significant impact on productivity and is estimated to result in 170 million lost working days each year (MoH Kenya, 2001b). This economic burden has the largest impact on the rural poor who rely largely on small-scale agriculture as a source of income. The underlying level of morbidity due to malaria is not uniform across Kenya, but varies considerably due to a complex set of climatic, human, and vector interactions. These factors are discussed in detail in section 5.2.1 of Chapter 5.

2.5.2 Using effective drugs to combat malaria in Kenya

Government led efforts to combat malaria in Kenya are coordinated by the Division of Malaria Control (DOMC) at the Ministry of Health. The current policy framework on combating malaria (MoH Kenya, 2001b) is centred around four key strategic objectives: (1) clinical management through provision of effective and prompt treatment; (2) management of malaria and associated anaemia in pregnancy; (3) control of the malaria vector (anopheline mosquitoes) using insecticide treated bed nets (ITNs) and spraying of insecticides; and (4) improving epidemic preparedness and response. The first objective in this list is seen as the cornerstone of the national malaria strategy, aiming to guarantee that people have rapid access to effective, affordable, acceptable, and available antimalarial drugs to enhance prompt and effective treatment of malaria episodes. A critical factor in the provision of effective antimalarial drugs is the need to continually monitor drug efficacy. That is, to monitor the extent to which the malaria parasites within the population are developing resistance to the drugs being used. The

development of resistance is a largely inevitable evolutionary process, occurring as a result of innate genetic diversity in the parasite population. Those genotypes that are better suited to survive the effects of a given drug are more likely to survive and hence their prevalence grows until a large proportion of the parasite population shares this resistant physiology (Wernsdorfer, 1994; Warrell et al., 2002). As such, all antimalarial drugs have a finite useful therapeutic life (UTL), although this can vary widely depending on the mechanism and utilisation of the drug.

Since its introduction in the 1930s, the primary antimalarial drug for *P.Falciparum* malaria in Kenya has been chloroquine (Shretta et al., 2000). Resistance to chloroquine grew rapidly across sub-Saharan Africa during the 1980s and 1990s, leading to substantial increases in mortality (Marsh, 1998; Trape et al., 1998, 2001). Chloroquine was belatedly replaced in Kenya with a new alternative, sulphadoxine pyrimethamine (SP), in 1998. However, resistance to SP, and the main alternative, Amodiaquine (AQ), developed quickly following their introduction which led to a substantial reassessment of drug policy in 2003. A review of studies carried out by the DOMC Drug Policy Technical Working Group (DPTWG) concluded that both SP and AQ had fallen below an acceptable level of efficacy, with an average SP treatment failure rate of 33% amongst children under five (MoH Kenya, 2005c). In response to this declining efficacy, and in line with international recommendations (WHO, 2001a, 2003a, 2005b), the DPTWG recommended that a change in treatment policy be introduced, with the introduction of artesunate-based combination therapy (ACT) as the new first-line antimalarial treatment. The formal plan for transition to the new ACT, artemether-lumefantrine (brand name Coartem®), was launched by the Ministry of Health in 2005 (MoH Kenya, 2005c). ACTs contain a newly developed and potent antimalarial component derived from the Chinese herb *Artemisia annua*. By using this compound in combination with a second drug that works in a different way, ACT is both highly effective and likely to resist the growth of rapid parasite resistance. It is hoped that ACT, if implemented carefully across Africa, will allow substantial and long-term reductions in malaria mortality and morbidity (White, 1998; White et al., 1999; Garner and Graves, 2005).

2.6 Estimating drug demand

A critical aspect of the transition from SP to Coartem® in Kenya is to ensure that the correct quantity of the drug is obtained. Accurate estimates of demand are required for efficient procurement and delivery. It is vital that adequate supplies of the drug are available since under-stocked health facilities will lead to a rapid loss of public confidence, further discouraging utilisation and, hence, drug dissemination (MoH Kenya, 2005c). However, Coartem® is currently substantially more costly than previous drugs (Kindermans, 2002) and has a limited shelf-life, meaning that excessive drug stocks are an expensive waste of resources (UN, 2005). Well defined demand estimates are also required by international donors who can provide funding for the new drug (RPM plus, 2005; GFATM, 2005; WHO, 2006).

The challenge of defining drug requirements in Kenya and across sub-Saharan Africa has been the subject of numerous studies by academic and public health institutions (Kindermans, 2002; Snow et al., 2003b, 2003c; MoH Kenya, 2004, 2005c). A critical distinction is between estimating the drug requirements of a health system for treating a given condition and estimating the theoretical drug requirements for treating all cases of that condition in the population. In low-income countries, and for a disease such as malaria, the difference between the two is likely to be substantial due to factors discussed previously, primarily the low utilisation of formal health services to obtain treatment. In recognition of the widespread inadequacy of health system efforts to estimate drug demand, the WHO defined a set of best-practice procedures to be implemented when quantifying the amount of drugs needed to treat a given condition within a health service (WHO, 1988). Two simple but distinct approaches are advocated: the morbidity method and the consumption method. The consumption method is based on assessing the quantity of drugs that have been consumed, either at the full set of health facilities within a health system, or at a representative sample. Where demand for a new drug is to be estimated, the consumption method relies on assessing the consumption of the previous, outgoing, drug. Even when reliable data are available for the previous drug, a key shortcoming is that the patterns of drug stocking and utilisation may be very different than those for the new drug, given differences in cost and shelf-life, for example. The morbidity method takes a different approach and requires estimates of the number of treatment episodes for a given disease in all health facilities.

This is then multiplied by the quantity of drugs that are administered for each treatment episode to obtain an estimate of the total drug requirement.

2.6.1 The role of HMIS in estimating drug demand

Accurate drug demand estimates made using the morbidity method are dependent on reliable estimates of the number of treatment episodes (the treatment burden). In Kenya, the appropriate source of data on which to base these estimates is the database of routine health facility records collected within the HMIS. Because of the known incompleteness of the database, however, these data have not been used to generate such estimates. A number of important definitions must be considered when using routine facility data in drug demand estimates. The variable of interest is the number of treatment episodes. A treatment episode differs from other forms of contact in that it requires a standard drug treatment. Follow-up visits that do not result in further treatments, for example, do not strictly constitute a treatment episode. Furthermore, a single patient visit may result in more than one treatment episode if he/she is treated for multiple conditions. Another complication is that the standard drug treatment may vary with the age of the patient and the severity of the condition. Where this is the case, the number of treatment episodes within each age group or severity category must be quantified separately. In principle, routine patient records should contain all the necessary information about each patient visit including patient age, sex, the condition for which they were treated, and whether the visit was a first contact, a follow-up visit, or a referral. Unfortunately, much of these data are often never collected, or are lost when data are aggregated within the HMIS.

2.7 Chapter summary

This chapter has presented information that provides the contextual backdrop to the problem addressed in this project: the estimation of outpatient treatment burdens for malaria in Kenya. The importance of health information in low income countries is becoming increasingly recognised, just as the gross deficiency of current information for meeting a wide range of requirements is becoming apparent. Although increasing numbers of health data are being collected in low-income countries, much of this is aimed at monitoring broad national-level indicators and is insufficient for detailed

system management. HMIS are one mechanism for routine data collection that should provide a wealth of invaluable information to decision-makers about the supply and demand of resources at health facilities within a health system, but are failing to do so in most low-income countries for a range of reasons including low utilisation rates, unreliable diagnosis of common diseases, and under-reporting of data from health facilities. A primary drain on health system resources in Kenya is the burden of malaria. Because of the introduction of expensive new drugs, a critical responsibility of the HMIS in Kenya is the provision of routine outpatient data which will allow the malaria treatment burden to be quantified. Widespread under-reporting, however, means that the HMIS data cannot be used directly to make these estimates. It is the problem of making reliable estimates with this incomplete data that is addressed in this project.

Chapter 3

Data

Chapter 3

3. Data

3.1 Introduction

This project is based on data integrated from two independent data sets. The principal data of interest originated from a routine outpatient database generated from health facilities across Kenya that was collected and collated within the Kenyan HMIS. These data were matched to a second database that contained the latitude and longitude coordinates of each health facility. In this chapter, the HMIS outpatient data set and the georeferenced health facility data set are both described in detail. Exploratory analysis is presented that describes the broad spatial and temporal characteristics of the outpatient data set. Analysis is also included that examines the extent and patterns of missing data in the outpatient data set.

3.2 The Kenyan national health service database (NHSD)

The stated aim of this project is to predict the total count of outpatients treated for malaria in all health facilities across Kenya using incomplete HMIS outpatient data. A prerequisite in this task is that the total number of health facilities in the country is known. Furthermore, the spatiotemporal techniques that are brought to bear on this task in this project require that the spatial and temporal locations of both data and unsampled points are known. Outpatient data within the HMIS are temporally referenced, with each count corresponding to a known month. In common with most HMIS in low-income countries, however, the Kenyan system does not include spatial data, meaning that the

spatial locations of the health facilities from which data were generated are unknown. The current project was conceived, in part, because of the construction of a new and unique spatial database that resulted from an extensive exercise to list and georeference all health facilities in Kenya. This database is now known as the national health service database (NHSD) and represents the first such resource of its type in Kenya. A summary of the construction and key characteristics of the NHSD is presented below. For detailed accounts, the reader is pointed to Noor et al. (2004) and Noor (2005).

3.2.1 Construction of the NHSD

The NHSD was developed and made available for this study by a team led by Dr. Abdisalan Noor at the Malaria Public Health & Epidemiology Group, Centre for Geographic Medicine, part of the KEMRI-University of Oxford-Wellcome Trust Collaborative Programme in Nairobi. The first stage in the creation of this resource was the establishment of a single comprehensive list of health facilities from all service providers around the country. Various independent records of public and private health facilities held by the Ministry of Health and various other Governmental and NGO bodies were cross-checked and compiled into a single list. This list was then augmented with information obtained directly from each district such as hand-drawn maps, local listings, telephone directories, and reports. Provisional lists were then sent out to District Health Management Teams and relevant NGOs and other parties for cross-checking and corrections. Having established a single list of Kenyan health facilities, a set of spatial coordinates were obtained for each using georeferencing data from a variety of sources. Previous projects run by various research and NGO bodies around Kenya had led to around half of the 72 Kenyan districts having global positioning system (GPS) coordinates for some or all of their constituent health facilities. Where health facilities on the list did not have GPS data, coordinates were obtained instead from various mapping sources or by matching facilities to known village or sub-location coordinates.

3.2.2 Use of the NHSD in this project

The NHSD does not represent a single exercise to provide a snapshot of Kenyan health facilities at a particular instant in time. Rather, the database is continually updated as

new facilities open, others close, and new information is obtained. In light of this dynamic nature, analysis in this study operated on two different versions of the NHSD. The bulk of model development and testing is described in Chapters 7 and 8 and was carried out using a version of the NHSD made available to this study in August 2004. Implementation of the developed model to obtain final predictions was carried out using an updated version of the NHSD obtained in October 2005. In this section, the former version is described in more detail. The updated version is summarised alongside the presentation of the final model implementation in Chapter 9.

3.2.3 Health facilities in the NHSD

Table 3.1 lists all health facilities included in the August 2004 version of the NHSD. These include facilities operated by the Ministry of Health, the charitable missions, private-for-profit organisations, NGOs, and other minor service providers including the armed forces, local authorities, and other governmental ministries. In addition to the main hierarchy of facility types discussed in Chapter 2, specialist facilities such as nursing homes and maternity hospitals, and institutional health facilities were included. Not all health facilities were georeferenced and, in this version of the database, not all health facilities had a unique identification code (an HMIS number) that was necessary to match outpatient data in the HMIS with the corresponding health facility.

In this study, the focus was on Ministry of Health facilities only and this decision was driven by several factors. Firstly, the need for reliable estimates of treatment burdens for malaria is particularly pressing for government facilities because the phased introduction of Coartem® as the first-line antimalarial will begin in this sector (MoH Kenya, 2005c). Secondly, the government sector represents a relatively stable and formally documented set of health facilities around the country. Although the construction of the NHSD revealed a substantial number of Ministry of Health facilities that were not on formal lists at the Ministry, the exercise resulted ultimately in a comprehensive inventory of health facilities within this sector, of which only a handful could not be georeferenced. This is in contrast to facilities provided by the other major sectors such as the charitable missions and private organisations, about which far less complete and reliable information was available, with a lower proportion being georeferenced.

Service provider		1.Hospitals (district, sub-district)		2.Hospitals (referral and provincial)		3.Health centres		4.Dispensaries		5.Private hospitals		6.Private clinics and Medical centres		7.Nursing homes & maternity hospitals		8.Special treatment hospitals		9. Institution health facilities		All facility types	
		n	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)
MOH	Georef.	119	(100.0)	10	(100.0)	475	(99.2)	1387	(96.7)	-	-	-	-	1	(50.0)	6	(100.0)	37	(94.9)	2035	(97.4)
	HMIS no	116	(97.5)	10	(100.0)	448	(93.5)	1230	(85.8)	-	-	-	-	2	(100.0)	4	(66.7)	32	(82.1)	1842	(88.2)
	Both	116	(97.5)	10	(100.0)	445	(92.9)	1194	(83.3)	-	-	-	-	1	(50.0)	4	(66.7)	31	(79.5)	1801	(86.2)
	Total	**119**		**10**		**479**		**1434**		-		-		**2**		**6**		**39**		**2089**	
MISS	Georef.	83	(96.5)	-	-	132	(95.7)	665	(86.1)	-	-	-	-	12	(66.7)	5	(71.4)	3	(75.0)	900	(87.8)
	HMIS no	75	(96.5)	-	-	117	(95.7)	593	(86.1)	-	-	-	-	12	(66.7)	6	(71.4)	4	(75.0)	807	(87.8)
	Both	73	(84.9)	-	-	112	(81.2)	518	(67.1)	-	-	-	-	9	(50.0)	4	(57.1)	3	(75.0)	719	(70.1)
	Total	**86**		-		**138**		**772**		-		-		**18**		**7**		**4**		**1025**	
PRIV	Georef.	1	(100.0)	-	-	15	(71.4)	126	(52.5)	81	(81.0)	1135	(50.9)	207	(68.5)	9	(26.5)	78	(65.5)	1652	(54.2)
	HMIS no	1	(100.0)	-	-	13	(61.9)	197	(82.1)	47	(47.0)	565	(25.3)	199	(65.9)	3	(8.8)	97	(81.5)	1122	(36.8)
	Both	1	(100.0)	-	-	9	(42.9)	104	(43.3)	39	(39.0)	266	(11.9)	142	(47.0)	2	(5.9)	69	(58.0)	632	(20.7)
	Total	**1**		-		**21**		**240**		**100**		**2231**		**302**		**34**		**119**		**3048**	
NGO	Georef.	2	(100.0)	-	-	3	(75.0)	32	(76.2)	-	-	-	-	3	(100.0)	16	(64.0)	1	(25.0)	57	(71.3)
	HMIS no	0	(0.0)	-	-	2	(50.0)	16	(38.1)	-	-	-	-	1	(33.3)	24	(96.0)	2	(50.0)	45	(56.3)
	Both	0	(0.0)	-	-	2	(50.0)	12	(28.6)	-	-	-	-	1	(33.3)	15	(60.0)	0	(0.0)	30	(37.5)
	Total	**2**		-		**4**		**42**		-		-		**3**		**25**		**4**		**80**	
LA	Georef.	-	-	-	-	47	(92.2)	36	(85.7)	-	-	-	-	3	(75.0)	1	(100.0)	1	(50.0)	88	(88.0)
	HMIS no	-	-	-	-	50	(98.0)	42	(100.0)	-	-	-	-	4	(100.0)	1	(100.0)	2	(100.0)	99	(99.0)
	Both	-	-	-	-	46	(90.2)	36	(85.7)	-	-	-	-	3	(75.0)	1	(100.0)	1	(50.0)	87	(87.0)
	Total	-		-		**51**		**42**		-		-		**4**		**1**		**2**		**100**	
AF	Georef.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	21	(70.0)	21	(70.0)
	HMIS no	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	21	(70.0)	21	(70.0)
	Both	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	16	(53.3)	16	(53.3)
	Total	-		-		-		-		-		-		-		-		**30**		**30**	
OTHER MIN	Georef.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	105	(84.7)	105	(84.7)
	HMIS no	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	109	(87.9)	109	(87.9)
	Both	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	94	(75.8)	94	(75.8)
	Total	-		-		-		-		-		-		-		-		**124**		**124**	
All Providers	Georef.	205	(98.6)	10	(100.0)	672	(97.0)	2246	(88.8)	81	(81.0)	1135	(50.9)	226	(68.7)	**37**	(50.7)	246	(76.4)	4858	(74.8)
	HMIS no	192	(92.3)	10	(100.0)	630	(90.9)	2078	(82.1)	47	(47.0)	565	(25.3)	218	(66.3)	38	(52.1)	267	(82.9)	4045	(62.3)
	Both	190	(91.3)	10	(100.0)	614	(88.6)	1864	(73.7)	39	(39.0)	266	(11.9)	156	(47.4)	26	(35.6)	214	(66.5)	3379	(52.0)
	Total	**208**		**10**		**693**		**2530**		**100**		**2231**		**329**		**73**		**322**		**6496**	

Table 3.1 Breakdown of health facilities by type and service provider in the August 2004 version of the national health service database. Total facility counts are shown along with the count (and percentage) of facilities that have georeferencing data (Georef.), that have an HMIS number (HMIS no.), and that have both (Both). Service providers are Ministry of Health (MOH), mission (MISS), private (PRIV), non-governmental organisations (NGO), local authority (LA), armed forces (AF), and other Ministries (OTHER MIN). The section highlighted in grey shows the set of 1765 facilities that were used in this project for model development and testing.

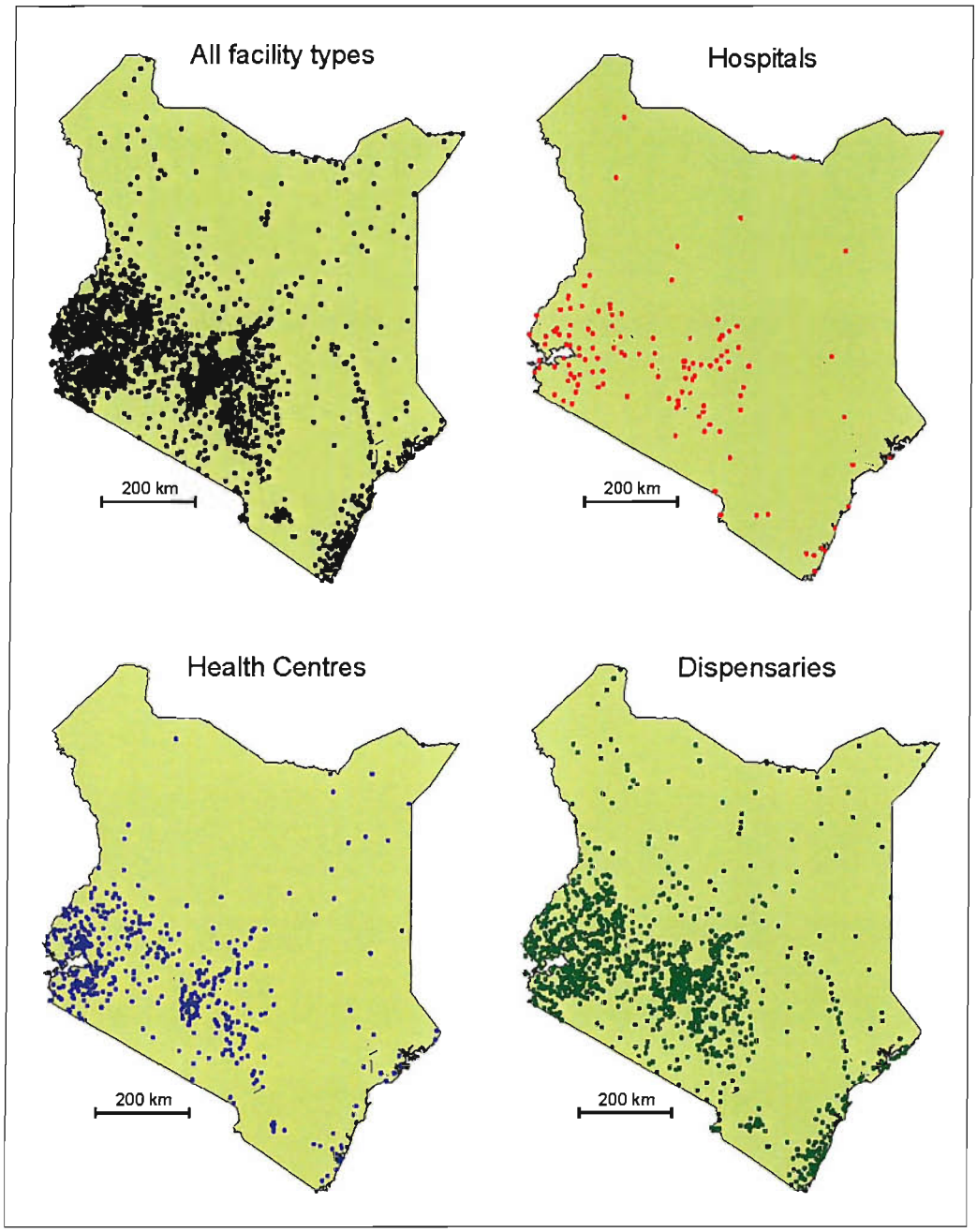


Figure 3.1 Maps of Kenya showing the locations of the 1765 Ministry of Health facilities used in this study for model development and testing. Each of the three facility categories used are shown along with the combined set of all three types.

For the purposes of model development and testing, a set of 1765 facilities was selected in this project from the August 2004 version of the NHSD (Table 3.1). This set was composed of all mainstream Ministry of Health outpatient facilities that were both georeferenced and had a unique HMIS number. This set excluded specialist non-outpatient facilities such as nursing and maternity homes. The various categories of

outpatient facility were condensed into three principal types: hospitals (referral and teaching hospitals, provincial, and district hospitals); health centres (subsuming sub-district hospitals); and dispensaries (subsuming sub-health centres). This simpler classification represents a broad grouping of facilities according to the generic levels of service they provide. This set consisted of 126 hospitals, 445 health centres, and 1194 dispensaries. The spatial distribution of these health facilities across Kenya (Figure 3.1) reflects approximately the underlying population density, although it is well established that access to health facilities is not equitable across the country with rural areas, for example, being generally under-served in relation to urban areas (Noor et al., 2003, 2006).

3.3 The Kenyan HMIS outpatient data set

The routine outpatient data on which this project is based were obtained directly from the Division of HMIS at the Ministry of Health by the KEMRI-University of Oxford-Wellcome Trust Collaborative Programme team. This team obtained the data in a simple TXT format and imported them into MS Excel (Microsoft Corp., USA) and subjected them to extensive checks for duplication and inconsistencies before porting them into MS Access (Microsoft Corp., USA) using an MS QuickBasic (Microsoft Corp., USA) script. Each record included a unique facility code, which allowed the entire data set to be linked to the NHSD such that data for each facility were integrated with the corresponding latitude and longitude coordinates. This formatted and spatially referenced database was then made available for this project.

The HMIS data consisted of monthly records of diagnoses made at outpatient departments of health facilities across Kenya over an 84-month period (January 1996-December 2002). Each record included the total number of outpatients attending a given facility during a given month. The number of diagnoses made under a wide range of diagnostic codes was also available for each monthly record per facility. Records were not structured by age, sex or distinguished as initial or follow-up visits. Due to the limitations on diagnosis accuracy discussed in Chapter 2, diagnoses could only be interpreted as representing a *presumed* case of a given condition.

In this section, those data that corresponded to the set of 1765 Ministry of Health

facilities defined earlier are presented. Analysis was conducted to assess the number of missing data within this set and to describe any patterns in when and where data were missing. Further exploratory analysis was then carried out to describe the broad spatial and temporal characteristics displayed by the various diagnostic codes.

3.3.1 Exploratory analysis of missing data

An initial aim in the analysis was to characterise and quantify the extent of incompleteness due to missing data. Missing data in the national database may result from failings at different points in the HMIS framework such as failure by individual facilities to submit their monthly records, failure of the District Medical Records

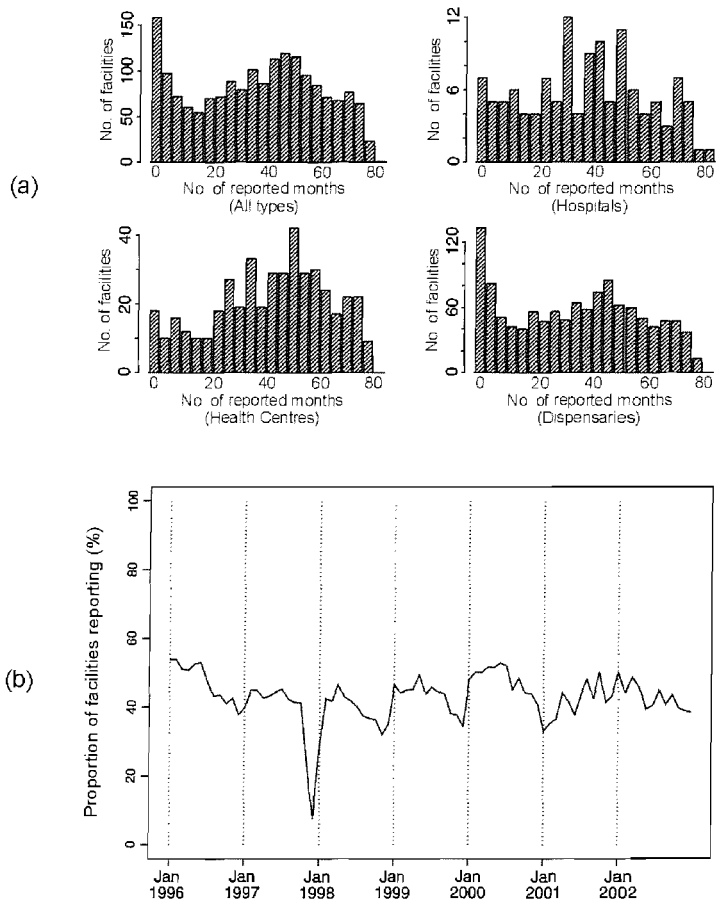


Figure 3.2 (a) Histograms showing the distribution of the number of months reported for each facility type. Complete reporting would result in 84 monthly records from each facility. (b) Time series plot showing the proportion of health facilities that reported in each of the 84 months.

Officers to collate data correctly or to submit their monthly District Outpatient Morbidity Summary, failure by HMIS headquarters to enter these data correctly into the main database, or simple physical loss of the relevant form between or at the various levels in the HMIS. In this study, the relative amount of missing data from each facility was quantified by a *reporting rate*, defined as the percentage of months for which an outpatient record was available for each facility. Facilities with less than 100% reporting rates were deemed to have *under-reported*, accepting that in some cases this term will not accurately describe the cause of the missing data. Under-reporting was assessed by province for each facility type.

3.3.1.1 Results

Under-reporting was found to be widespread, although there was considerable variation between facility types and provinces (Table 3.2). Figure 3.2 (a) shows the distribution of reporting rate values for each facility type. No facilities reported in all 84 months whilst 158 facilities (9%) did not report in any month. A complete 84-month data set for the 1765 facilities would contain 148,260 records. Only 63,543 records were present representing an overall reporting rate of 42.9% over all facility types and provinces. This ranged from 12.2% in Nairobi to 52.4% in Central Province. Health centres had the highest reporting rate (50.0%), followed by hospitals (44.2%) and dispensaries (40.1%). Overall reporting rate varied both within and between years, with a minimum of 7.6% in December 1997 and a maximum of 52.8% in June 2000 (Figure 3.2 (b)).

	Rift Valley	Nairobi	Nyanza	Central	Eastern	Coast	Western	N. Eastern	KENYA
Hospitals	47.8	12.3	47.2	36.1	47.0	37.5	50.9	37.7	44.2
Health centres	39.7	15.8	61.1	58.3	54.0	43.7	59.4	19.0	50.0
Dispensaries	31.9	9.0	46.7	51.7	43.3	41.1	40.6	20.2	40.1
All facilities	34.6	12.2	51.2	52.4	45.9	41.3	49.9	20.9	42.9

Table 3.2 Overall reporting rate in each Kenyan province by facility type. These values represent the total percentage of the expected monthly outpatient records that were available within the HMIS database.

It is likely that the majority of missing data are caused by facility-level failures to submit monthly reports. However, detailed examination of the spatial patterns of under-reporting revealed evidence that district-level processes also affected reporting rate in some cases. An example is provided by the neighbouring districts of Kericho and Nyando in the west of Kenya. Plots of district-wide reporting rate for both districts were derived and these are shown in Figure 3.3 along with maps showing the health facilities in each district. Both plots display evidence of national-level factors, specifically the very low reporting rate recorded in December 1997 associated with the national nurses strike. Whilst both plots display a clear temporal trend, these are very different for each district. Reporting in Kericho (Figure 3.3 (b)) decreases steadily throughout 1996 and 1997, and then steadily rises for the remainder of the data period. In contrast, reporting in Nyando (Figure 3.3 (c)) is consistently high throughout 1996 to 1998 (excepting the December 1997 event) and then consistently low throughout 1999 to 2002. This marked difference between neighbouring districts suggests the influence of factors operating at the district level. A district-wide near-cessation of reporting such as occurred in Nyando from 1999 onwards clearly results in a contiguous spatiotemporal 'hole' in the HMIS data set, which has clear implications for attempts to predict missing values based on data proximate in space and time.

The observed overall reporting rate of 42.9% confirmed that the incompleteness of this database is substantial and presents a significant challenge to users. Temporal variations in reporting rate may have been caused by factors such as impairments to transport during the rainy season which impedes the effectiveness of the data delivery network, and seasonal variations in the availability of staff. The pronounced dip in reporting rate in December 1997 was likely to be related to a national nurses strike at that time. The fact that the set of facilities that report in any given month changes through time is important when attempting to identify and explain temporal trends. An observed national increase in cases of a certain illness in a given month, for example, could be brought about by a relative increase in reporting from areas where that illness is more prevalent.

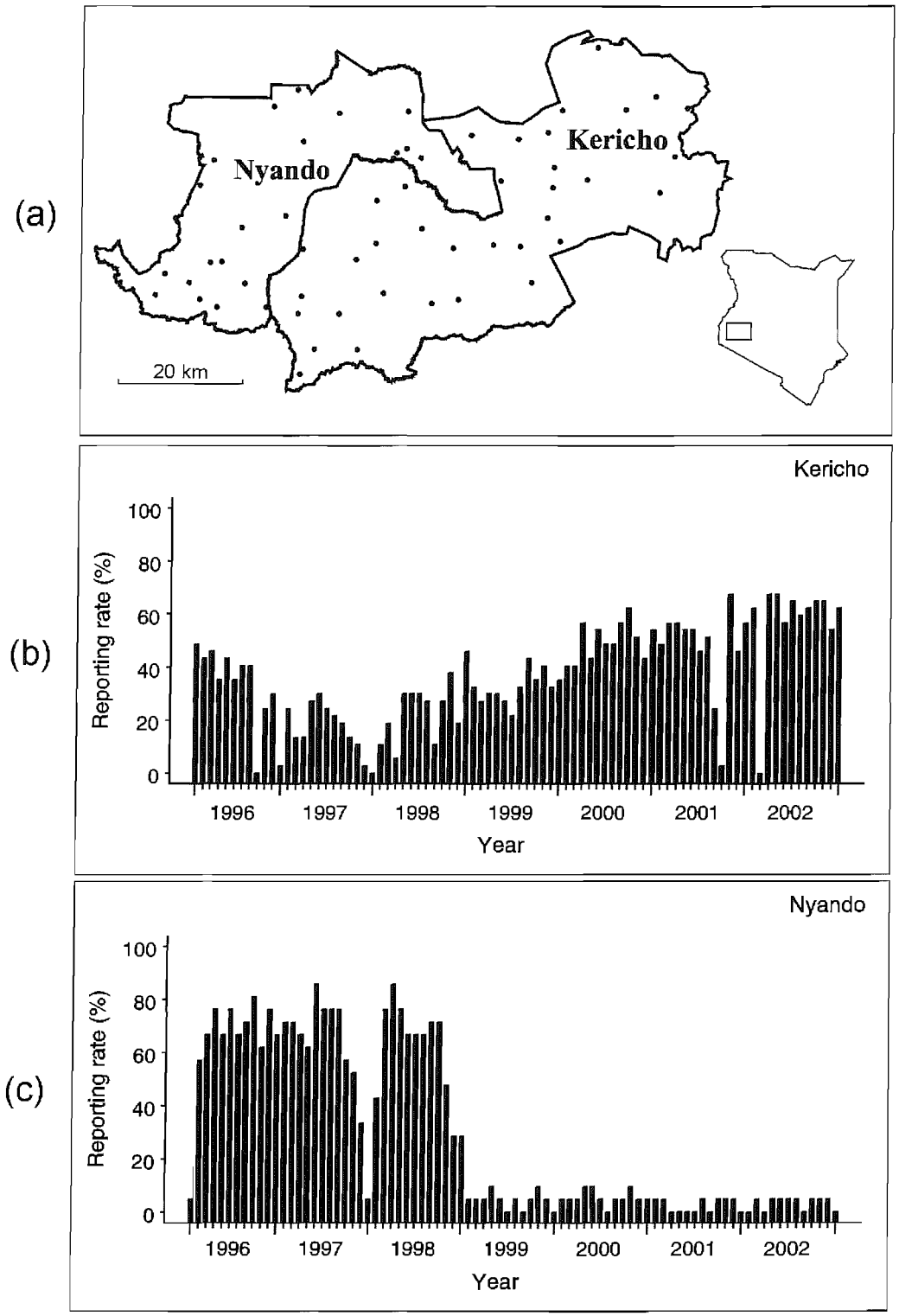


Figure 3.3 District-level reporting patterns in two western districts of Kenya. (a) Map showing the boundaries and health facilities (dots) of Nyando and Kericho districts. The two plots show the corresponding district monthly reporting rate (i.e. the percentage of facilities in each district that reported in each of the 84 months) for (b) Kericho and (c) Nyando.

3.3.2 Exploratory analysis of outpatient diagnosis patterns

The purpose of this section is to present a series of exploratory analyses that aimed to describe the broad spatial and temporal patterns displayed by the outpatient data for 11 diagnostic codes relating to the principal illnesses and communicable diseases of public health importance in Kenya. Along with malaria, these were anaemia, diarrhoea, ear infections, malnutrition, measles, meningitis, pneumonia, pyrexia, respiratory diseases, and tuberculosis. It should be re-emphasised that, given the ambiguity associated with outpatient diagnosis, these categories should be interpreted as representing presumed rather than confirmed causes of illness.

3.3.2.1 Methods

For each of the 11 diagnostic codes, the total number of cases reported during the 84-month period was determined along with summary statistics that describe the variation between facilities in the mean number of cases per month. The relative contribution of individual diagnoses to total outpatient morbidity was also determined by comparing each diagnosis-specific case count to the corresponding total case count. Relative contributions of each illness were then determined by month for the eight Kenyan provinces. This allowed a broad assessment of both the spatial variation in disease composition across Kenya and the way in which this composition varied during the study period.

The pattern of seasonal variation in cases was investigated for each diagnostic code by determining the percentage of cases that occurred in each calendar month, averaged over the study period. Care was taken to exclude possible bias introduced by monthly differences in reporting rate by standardising monthly case counts by the number of facilities that reported. The standardised percentage of cases, $p(i,j)$, occurring in a given calendar month, i, and year, j, was calculated by first dividing the total count of cases, c, for that month and year by the number of corresponding facility records, r, and then standardising by the sum of this value for all m = 1,2,...12 months in year j (3.1)[1]:

[1] Please note that, from here on, numbers presented in parentheses within the text in this way refer to the numbered equations that appear throughout the document.

$$p(i,j) = \left(\frac{c_{ij} r_{ij}^{-1}}{\sum_{m=1}^{12} c_{mi} r_m^{-1}} \right) \times 100 \quad (3.1)$$

These values were calculated nationwide and by province for each diagnostic code. In each case, an average seasonal profile was also determined as the mean percentage, q, for each calendar month, i, over the n years j= 1,2,...n, where n = 7 (3.2).

$$q(i) = \frac{1}{n} \sum_{j=1}^{n} p(i,j) \quad (3.2)$$

3.3.2.2 Results

A total of 55.9 million cases were included in the database, of which the selected 11 illnesses contributed 40 million. The total number of cases recorded under each diagnostic code during the seven year period ranged from under 7000 for meningitis to 18.5 million for malaria (Table 3.3). For malaria, the mean monthly case count at facilities ranged from zero to 2044, with a mean of 205.0 for dispensaries, 299.1 for health centres and 634.2 for hospitals. The distribution of mean monthly case count per facility was positively skewed for all illnesses, indicating that a small proportion of facilities had monthly counts that were far greater than those of the majority of facilities. The skewness statistic was smaller (i.e. less skewed distributions) in the more common illnesses (e.g. malaria = 3.1) than the comparatively rarer ones (e.g. meningitis = 39.7), and was also smaller when facility types were considered individually. Monthly case counts were largest at hospitals although this pattern was again more pronounced for the less common illnesses. Malaria was the most common of the illnesses studied, contributing 33.2% of all cases over the seven years. Respiratory conditions were second most common with 23.4%, then diarrhoea (4.5%), pneumonia (2.3%), ear infections (1.3%), anaemia (0.6%), pyrexia (0.3%), malnutrition (0.2%), measles (0.2%), tuberculosis (0.1%), and meningitis (0.01%). The relative contribution made by each illness varied between provinces (Figure 3.4). Respiratory conditions, for example, ranged from 31.6% of total cases in Central province to 17.2% in Nairobi. Malaria ranged from a contribution of 41.2% in Western province to 6.9% in Nairobi.

Table 3.3 Total number of diagnoses and summary statistics for 11 selected diagnostic codes for the 84-month study period. Facility types are abbreviated as H (hospitals), HC (health centres) and D (dispensaries).

Diagnosis	Facility type	Total cases	Mean cases per month per facility				
			Mean	Standard deviation	Minimum	Maximum	Skewness[a]
Anaemia	**ALL**	**348,401**	5.8	13.5	0.0	241.3	7.4
	H	101,194	20.8	32.7	0.0	241.3	3.8
	HC	88,086	4.9	6.9	0.0	48.8	2.9
	D	159,121	4.5	10.5	0.0	130.8	6.5
Diarrhoea	**ALL**	**2,537,495**	38.7	47.6	0.0	823.0	6.9
	H	596,442	124.2	124.9	10.0	823.0	2.8
	HC	755,291	39.8	24.8	0.3	172.5	1.9
	D	1,185,762	28.6	23.0	0.0	238.5	2.6
Ear infections	**ALL**	**716,443**	11.5	56.8	0.0	2044.0	30.7
	H	223,184	56.0	202.1	0.0	2044.0	8.7
	HC	177,437	9.1	8.8	0.0	135.4	7.8
	D	315,822	7.5	7.7	0.0	101.0	4.9
Malaria	**ALL**	**18,559,406**	262.4	218.6	0.0	2889.0	3.1
	H	3,351,694	634.2	430.9	28.0	2889.0	1.8
	HC	6,036,608	299.1	177.3	11.0	1137.0	1.4
	D	9,171,104	205.9	140.8	0.0	934.8	1.3
Malnutrition	**ALL**	**113,887**	1.9	6.4	0.0	164.8	15.8
	H	37,179	7.7	20.1	0.0	164.8	5.8
	HC	30,159	1.7	3.6	0.0	54.2	8.6
	D	46,549	1.3	2.7	0.0	33.9	5.2
Measles	**ALL**	**111,321**	1.7	2.8	0.0	66.8	9.6
	H	15,492	3.4	6.5	0.0	66.8	8.1
	HC	37,864	2.0	2.2	0.0	15.4	2.9
	D	57,965	1.4	2.2	0.0	24.6	4.5
Meningitis	**ALL**	**6746**	0.2	4.1	0.0	163.3	39.7
	H	5330	1.6	15.0	0.0	163.3	10.9
	HC	645	0.0	0.1	0.0	1.1	5.6
	D	771	0.0	0.3	0.0	6.4	14.7
Pneumonia	**ALL**	**1,301,272**	20.2	42.1	0.0	1130.1	13.8
	H	310,052	71.9	121.3	0.0	1130.1	6.2
	HC	406,856	21.5	26.5	0.0	323.0	4.9
	D	584,364	13.9	20.5	0.0	290.0	5.7
Pyrexia	**ALL**	**159,987**	2.5	6.3	0.0	70.4	5.0
	H	24,866	6.2	10.5	0.0	56.9	2.5
	HC	53,570	2.6	6.5	0.0	59.4	4.9
	D	81,551	2.0	5.4	0.0	70.4	5.9
Respiratory	**ALL**	**13,089,152**	186.7	189.8	0.0	3731.1	7.2
	H	2,451,705	499.3	485.8	38.2	3731.1	3.6
	HC	3,839,460	194.7	125.5	10.0	895.3	2.0
	D	6,797,987	148.3	97.6	0.0	770.4	1.6
Tuberculosis	**ALL**	**55,203**	1.4	12.7	0.0	353.4	20.7
	H	39,349	13.5	42.7	0.0	353.4	6.2
	HC	9092	0.5	1.5	0.0	21.0	8.1
	D	6762	0.3	4.7	0.0	148.0	29.9

[a] Skewness values between -1 and 1 indicate an approximately normal distribution.

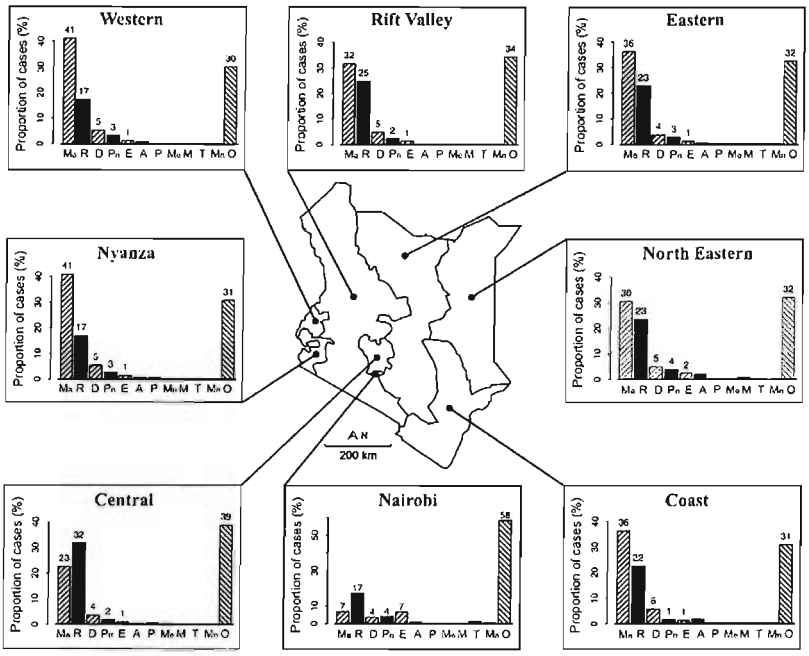


Figure 3.4 Bar charts showing the relative contribution of eleven selected diagnoses to total outpatient morbidity at facilities in each Kenyan province. Initials refer to malaria (Ma), respiratory conditions (R), diarrhoea (D), pneumonia (Pn), ear infections (E), anaemia (A), pyrexia (P), measles (Me), malnutrition (M), tuberculosis (T), meningitis (Mn), and all other causes (O). Percentages are given (above bars) for all other causes and for the first five illnesses.

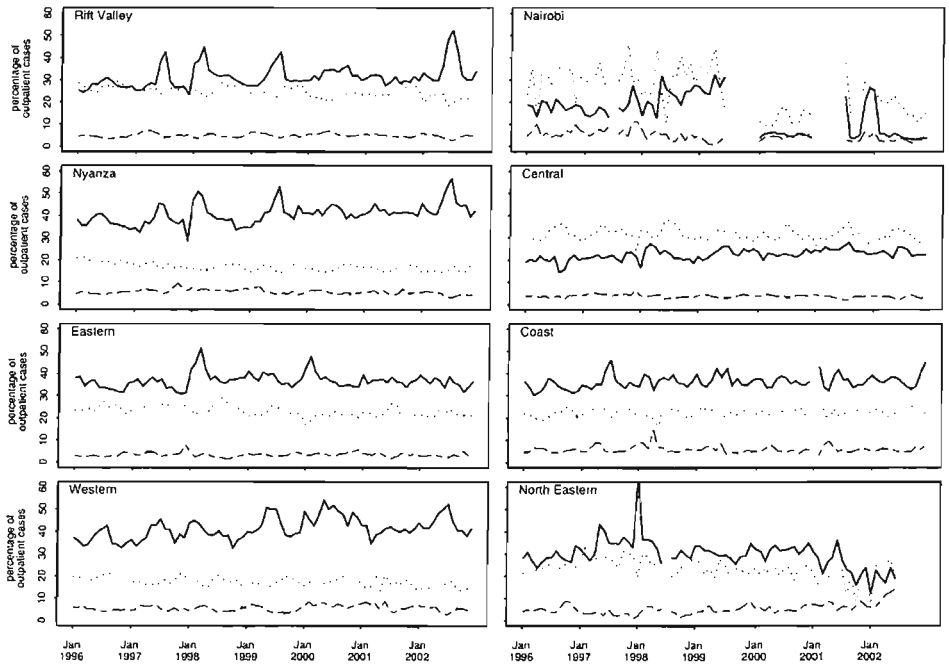


Figure 3.5 Percentage contribution of diagnoses of malaria (solid line), respiratory diseases (dotted line) and diarrhoea (dashed line) to total monthly outpatient cases at facilities in each Kenyan province for the 84-month study period January 1996 - December 2002. Discontinuities indicate that no records were present for that period.

Differences in percentage were also observed between facility types, with both malaria and respiratory conditions making a relatively greater contribution at dispensaries than at hospitals. Relative contribution varied temporally as well as spatially. The nationwide malaria contribution ranged from a maximum of 42.8% of total cases in March 1998 to a minimum of 26.9% in December 1997. Respiratory conditions ranged from 27.2% (July 1998) to 20.3% (October 2002), and diarrhoea ranged from 6.1% (April 1997) to 2.9% (July 2002). Figure 3.5 presents time-series plots of the relative contributions of these three illnesses to total outpatient morbidity for each province over the 84-month period. All three diagnoses displayed inter- and intra-annual patterns of variation. Distinct peaks in malaria contribution were present in various years, of which some occurred simultaneously in several provinces, whilst others were unique to a single province. Some regular (i.e. seasonal) intra-annual variation could be detected in the contribution of malaria, although this is exposed more clearly for this, and other diagnoses, by the seasonality profiles shown in Figures 3.6 and 3.7.

The pattern of seasonality differed between diagnoses (Figure 3.6). The mean nationwide seasonality profile for malaria revealed a characteristic peak in July, with a smaller peak in February and March. The individual values for each of the seven years plotted around this mean expose the extent of inter-annual variation in seasonal pattern which was greatest during the peak months. This seasonality of malaria was similar to that of anaemia, respiratory conditions and pneumonia, with each showing a relatively consistent pattern over the seven years. Ear infections, diarrhoea and malnutrition also had a consistent pattern over the study period with each exhibiting modest seasonality peaking in May, March, and July respectively. The three least common diagnoses (meningitis, measles, and tuberculosis) displayed no clear pattern of seasonality, with large inter-annual variations. For most diagnoses, the pattern of seasonality varied between provinces. Provincial seasonality profiles for malaria are shown in Figure 3.7. Strong peaks in cases in July along with smaller peaks in February and March were generally evident in provinces to the west of the country (Rift Valley, Central, Nyanza and Western provinces), along with Coast province. Eastern and North Eastern provinces had less pronounced seasonality whilst Nairobi province exhibited no clear pattern with erratic inter-annual variation.

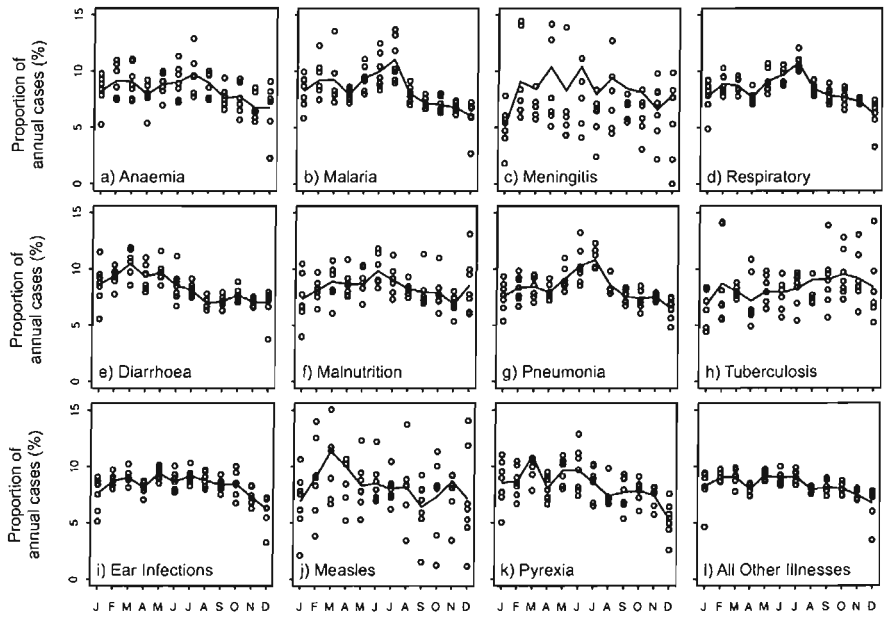


Figure 3.6 Seasonality plots showing the percentage of annual cases occurring each month for 11 selected illnesses at outpatient facilities for the 84-month study period January 1996 – December 2002. Shown are the monthly case proportions for each individual year (circles) as well as the seven-year mean (continuous line).

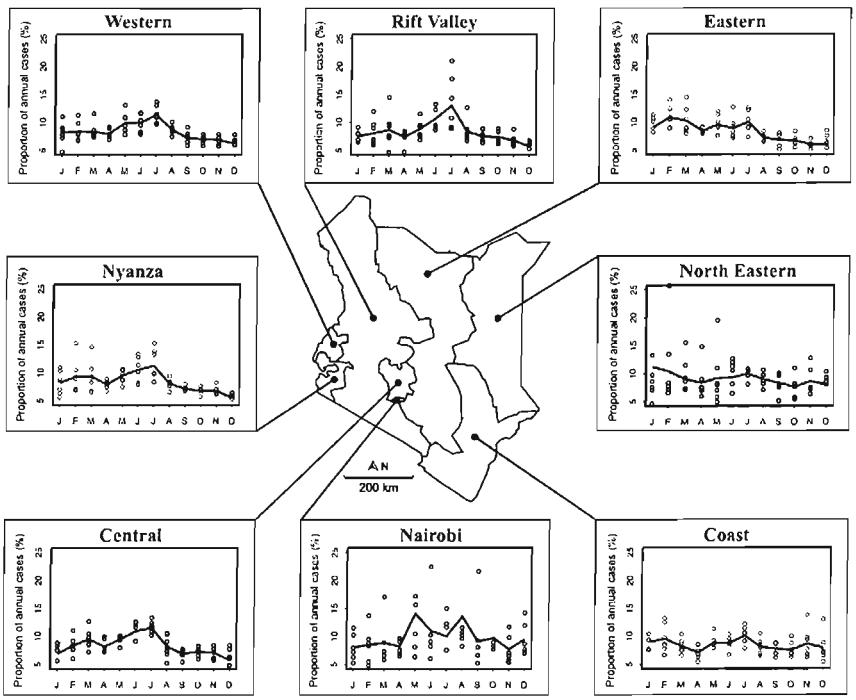


Figure 3.7 Seasonality plots showing the percentage of annual malaria cases occurring each month at outpatient facilities in each Kenyan province during the 84-month study period January 1996 – December 2002. Shown are the monthly case proportions for each individual year (circles) as well as the seven-year mean (continuous line).

3.3.2.3 Discussion

The relative contribution of each diagnostic code to the total number of outpatient diagnoses, and the varying nationwide pattern of these contributions has been presented, along with temporal trends for selected illnesses over the seven-year period. One result that requires explanation is the large disparity between Nairobi and Central provinces in the relative contribution of malaria and respiratory diagnoses to total outpatient morbidity, given that these provinces are contiguous neighbours. A tentative explanation is that the proportion of patients diagnosed under these two most common diagnoses is less in the Nairobi data due to a higher standard of diagnostic accuracy. A larger proportion of patients may be diagnosed more specifically with less common conditions and are therefore included in the 'other' category rather than as malaria or respiratory diagnoses. This argument is supported by the unusually large proportion of data from Nairobi province that originates from hospitals and health centres (which have relatively good diagnostic capabilities) rather than from dispensaries (which have relatively poor diagnostic capabilities) which was brought about by the unusually low reporting rate for dispensaries in Nairobi province of just 9%.

Distinct peaks in the time series showing the proportion of malaria cases can be linked to known malaria emergencies. A pronounced peak in early 1998 in Rift Valley, Nyanza, Eastern and North Eastern provinces corresponds to dramatic rises in malaria morbidity in highland and semi-arid areas of Kenya following exceptional rainfall in October to December 1997 associated with the 1997-1998 El Niño event (Brown et al., 1998; Karanja and Mutua, 2000). Distinct peaks are also evident in mid-1999 and mid-2002 in Western, Nyanza and Rift Valley provinces which again correspond to reported epidemic events following heavy rains (WHO, 1999; Hay et al., 2003).

Nationwide seasonality profiles (Figure 3.6) represent the overall distribution of cases across an average year for each illness. Whilst such profiles have important implications for health-system planning, it is vital that both temporal (i.e. inter-annual) and spatial variation in seasonality is considered in their interpretation. The patterns of malaria seasonality in each province shown in Figure 3.7 can be explained in broad terms by the corresponding seasonal pattern of rainfall. The strong peak in June and July in the four western provinces (Central, Nyanza, Western, and Rift Valley) follows the main rainy

season during March, April, and May. The period between peak rainfall and peak malaria incidence represents the time taken for the increased rainfall to result in more favourable mosquito habitats, the establishment of large mosquito populations, increased infective biting of humans, increased infection and, finally, increased morbidity. In more easterly provinces, the short rainy season that occurs during October, November, and December has a more pronounced effect, with the corresponding peak During January and February.

Inclusion of monthly case proportions for each individual year as well as the seven-year mean exposes the extent of inter-annual variation in seasonality, which can be attributed to a range of causes. It has been proposed, for example, that super-annual cycles in directly transmitted diseases can be driven by population dynamics under the susceptible, exposed, infectious, recovered (SEIR) model (Aron and Schwartz, 1984). An analysis of a 30-year time series of malaria admissions in Kericho in the western highlands of Kenya revealed that super-annual cycles accounted for over 30% of total variance (Hay et al., 2001). Long-term trends may also be present for many illnesses driven by factors such as population movement and demographic change. For malaria, such trends may also be driven by the decreasing drug efficacy (Marsh, 1998; Shanks et al., 2000; Bloland, 2001; Trape, 2001), and by climate change, although the relative significance of the latter is contested (Hay et al., 2002; Reiter et al., 2004). The provincial variation in seasonality described for malaria (Figure 3.7) reflects the spatial heterogeneity in the influence and timing of a range of controlling factors including meteorological determinants of suitability for transmission and the ecology and population dynamics of the mosquito vectors, parasites and at-risk populations (Snow et al., 1997, 1998; Craig et al., 1999).

By summarising the relative contributions and seasonal patterns of the 11 selected diagnostic codes at outpatient departments in each province over the seven-year study period, a profile can be constructed that begins to describe the spatial and temporal pattern of outpatient morbidity in Kenya during that time. However, due to the chronic level of missing data within the HMIS database, the extent to which these summaries are representative of the true national picture is unknown. Using these incomplete data to infer unknown properties of the complete set requires that the temporal and spatial heterogeneity described in this section is considered fully.

3.3.3 Choice of malaria as the diagnostic code of interest

Of the many illnesses and conditions contributing to public ill-health in Kenya that are included in the HMIS outpatient data set, malaria was chosen exclusively as the diagnostic code of interest in this project, and the disease for which treatment burdens would be estimated. This choice was driven by several factors. Firstly, there is a pressing case for the need of these estimates. As described in the previous chapter, malaria presents an overwhelming public health challenge in Kenya, blighting the lives of millions of Kenyans and imposing tangible economic constraints on development. Its prominence as a public health problem is reflected in it being the most commonly diagnosed disease in outpatients. Furthermore, even in the context of growing awareness of the overall need for increases in health information for evidence-based decision-making, malaria is a disease for which the need for accurate quantification is especially acute. As discussed earlier, this particular urgency arises from the current need to determine and obtain donor funding for the switch from inexpensive but rapidly failing antimalarials to more effective, but expensive, alternatives. Secondly, malaria is a disease that exhibits considerable spatial (Craig et al., 1999; Omumbo et al., 2002, 2005) and temporal (Hay et al., 1998a, 1998b) heterogeneity across Kenya and basic attempts to compensate for missing data within the HMIS to estimate treatment burdens do not consider explicitly the influence of space and time and are, therefore, likely to lead to biased results. As such, the decision to focus on malaria presents the opportunity for appropriate spatiotemporal techniques to be applied to a pressing public health problem.

3.3.4 Summary of data for model development

Having described in detail the HMIS outpatient and georeferenced health facility databases that underpin this project, the specific data that were extracted and used can now be summarised as follows. Data consisted of monthly records of diagnoses made at outpatient departments of Ministry of Health facilities across Kenya over an 84-month period (January 1996- December 2002). Each record included the total number of outpatient diagnoses made at a given facility during a given month and the number of these diagnoses that were for malaria. The records available were not structured by age, sex or distinguished as initial or follow-up visits, and malaria diagnoses were generally not laboratory-confirmed. The data, therefore, represent total cases (TC) or presumed

malaria cases (MC) seen as outpatients each month at health facilities identified by a unique georeferenced facility code.

3.4 Chapter summary

This chapter has presented the two principal data sets from Kenya on which this project was based. The first was a routine outpatient database collected by the HMIS from health facilities across the country listing the tally of outpatients treated each month under a variety of diagnostic codes, including malaria. This data set had been linked to a second which contained a comprehensive list of health facilities around the country and included, for the first time in Kenya, extensive georeferencing information. This study focused only on facilities operated by the Ministry of Health and, because the facility data set is being updated constantly, two versions were used in this project. An earlier version was used for model development and testing (Chapters 7 and 8) and an updated version was used in the implementation of the final model (Chapter 9). This chapter has also presented the first detailed description of the HMIS outpatient database, assessing the extent and pattern of under-reporting and describing the broad spatial and temporal heterogeneity of malaria and other important illnesses in the outpatient record.

Chapter 4

Methods

Chapter 4

4. Methods

4.1 Introduction

The purpose of this chapter is to provide a detailed description of the modelling tools that have been used in this project. The discussion in this chapter is limited to established geostatistical concepts and techniques, whilst the incorporation and adaptation of these methods in a series of modelling frameworks to meet the stated project objectives is discussed in subsequent chapters. The following sections introduce the conceptual underpinnings of geostatistics, and the key concepts and tools by which geostatistics can be used to characterise and predict spatial variables. The extension of spatial-only geostatistical techniques to space-time settings is then introduced and some key considerations are discussed along with examples of applications. Finally, a brief review is included of the use of geostatistical methods in public health and malaria settings.

4.2 The geostatistical paradigm

4.2.1 Deterministic and probabilistic modelling

The objectives of this project entail the prediction of presumed malaria cases (MC) at locations where it has not been sampled (government outpatient facilities in Kenya with data missing from the HMIS database). Such prediction requires the use of a model of

how the property of interest behaves at these unsampled locations. Various conceptual approaches exist for the formulation of such a model and a useful categorisation is between deterministic and probabilistic models. In a deterministic model, each unknown value is predicted as a single value with no associated prediction error. Such models can be employed when the physical mechanisms that govern the variable of interest are well understood and established physical equations exist that allow calculation of the unknown value with negligible or no error. As the scope and depth of contemporary scientific knowledge continues to grow, the complexity of systems for which deterministic modelling is feasible increases also. Sophisticated deterministic models have been developed to model processes in fields as diverse as sub-atomic physics, molecular biology, population dynamics, and glaciology. In fields such as the epidemiological and public health sciences, however, the systems of interest are generally of such complexity and magnitude that they retain an inherent unpredictability, even when many of the constituent processes are understood in detail. In the current setting, the variable of interest is the number of cases of malaria diagnosed at a given health facility in a given month. This variable is dependent on a myriad of massively complex and interacting physical, biological, demographic, social, and political systems that drive the prevalence of malaria in the population, the way that malaria sufferers utilise health services to obtain treatment, and the way they are diagnosed and recorded if they present to the formal health system. Given the gulf between our understanding and the complexity of the data-generating system, a deterministic model is neither feasible nor appropriate. What is needed instead is a modelling approach that recognises explicitly our uncertainty and allows the inevitable error associated with our predictions to be assessed. Probabilistic models represent an alternative paradigm to deterministic approaches. In a probabilistic model, the mechanism that generates the sample data and determines the values of the variable at unsampled locations is viewed as a random process. Although the mechanism in question is rarely, if ever, entirely random, the adoption of a probabilistic model provides a framework that can prove extremely useful in both predicting unsampled values and assessing the uncertainty of those predictions. Instead of predicting a single value for each unsampled location with assumed zero error, probabilistic models allow the prediction of a set of possible values with corresponding probabilities of occurrence. Unlike deterministic models, probabilistic models do not necessarily require knowledge of the physical process that generated the sample data. Rather, most of the information used is derived from the data themselves.

4.2.2 Geostatistics and the random function model

Geostatistics has been defined in broad terms as the study of phenomenon that fluctuate in space (Olea, 1991). Developed originally to address problems of spatial prediction in the mining industry (Matheron, 1971), the generality of the approach has led subsequently to its application in a diverse range of settings including geological, atmospheric, environmental, and epidemiological sciences. Geostatistics offers a collection of primarily probabilistic tools that have been developed to aid the understanding and modelling of spatial variability, with the principal motivation of predicting unsampled values dispersed in space. In common with most probabilistic approaches, each unobserved value z is characterised as the outcome of a random variable (RV) Z, defined as a variable whose values are randomly generated according to some probabilistic mechanism (Isaaks and Srivastava, 1989). RVs can be categorical or continuous with the probability of different outcome values being determined by some probability distribution. In spatial settings, each RV Z and outcome z are associated with a certain location $\mathbf{u}_0 = (x,y)$, a vector of spatial coordinates, and are denoted as $Z(\mathbf{u}_0)$ and $z(\mathbf{u}_0)$, respectively. The uncertainty about values of $Z(\mathbf{u}_0)$ can be fully characterised by a univariate cumulative distribution function (cdf) which models the probability that $Z(\mathbf{u}_0)$ does not exceed any given outcome z:

$$F(\mathbf{u}_0; z) = \text{Prob}\{Z(\mathbf{u}_0 \leq z\} \tag{4.1}$$

In the absence of any information about a given RV, all possible outcomes have an equal probability of occurrence and, as such, the cdf model does not increase our ability to infer the value of $z(\mathbf{u}_0)$. If a set of data from n neighbouring locations $\{z(\mathbf{u}_\alpha), \alpha = 1,2,\ldots, n\}$ is available, however, the information provided by these data may allow this prior model of uncertainty to be updated. This posterior model, updated by neighbouring data $z(\mathbf{u}_\alpha)$, is termed a conditional cumulative distribution function (ccdf):

$$F(\mathbf{u}_0; z \mid z(\mathbf{u}_\alpha)) = \text{Prob}\{Z(\mathbf{u}_0 \leq z \mid z(\mathbf{u}_\alpha)\} \tag{4.2}$$

A central theme of geostatistics is the provision of a framework by which sample data $z(\mathbf{u}_\alpha)$ can be used to update prior models of uncertainty for unsampled RVs $Z(\mathbf{u}_0)$ in

order to produce posterior ccdfs from which predictions of the unsampled value $z(\mathbf{u}_0)$ can be derived. Such a framework is provided by the random function (RF) model.

A spatial RF $Z(\mathbf{u})$ is defined as an infinite set of usually dependent RVs Z, one for each possible location $\mathbf{u}$ in the study area $\mathcal{A}$, $\{Z(\mathbf{u}), \forall\ \mathbf{u} \in \mathcal{A}\}$(Goovaerts, 1997). Just as a univariate cdf of an RV Z can be used to represent uncertainty around an unknown outcome value z, a multi-point cdf can be used to represent the joint uncertainty around outcome values $\{z(\mathbf{u}_1),z(\mathbf{u}_2),\ldots, z(\mathbf{u}_N)\}$at any given set of N locations spatially distributed across the study area:

$$F(\mathbf{u}_1,\mathbf{u}_2,\ldots\mathbf{u}_N;z_1,z_2,\ldots,z_N) = \text{Prob}\{Z(\mathbf{u}_1) \leq z_1, Z(\mathbf{u}_2) \leq z_2,\ldots, Z(\mathbf{u}_N) \leq z_N\} \qquad (4.3)$$

The set of all possible N-point cdfs for any value of N ($N \subseteq \mathbb{N}$) and for any choice of locations constitutes the complete spatial law of the RF $Z(\mathbf{u})$. In principle, a RF is only characterised fully by this complete spatial law. In practice, such complete characterisation is both infeasible and unnecessary for the prediction of unobserved values. The approach taken is to characterise the joint relationship between RVs at no more than two locations at a time, say $Z(\mathbf{u}_0)$ and $Z(\mathbf{u}_0')$, by the one- and two-point cdfs and corresponding moments. Of particular importance are the
two-point covariance:

$$C(\mathbf{u}_0, \mathbf{u}_0') = \text{E}\{Z(\mathbf{u}_0) \cdot Z(\mathbf{u}_0')\} - \text{E}\{Z(\mathbf{u}_0)\} \cdot \text{E}\{Z(\mathbf{u}_0')\} \qquad (4.4)$$

and variogram:

$$2\gamma(\mathbf{u}_0, \mathbf{u}_0') = \text{Var}[Z(\mathbf{u}_0) - Z(\mathbf{u}_0')] \qquad (4.5)$$

Under conditions of stationarity, the degree of dependence between two RVs separated by the same lag $\mathbf{h}$ ($\mathbf{h} = \mathbf{u}_0' - \mathbf{u}_0$), a vector of distance and direction) is the same for any such pair $Z(\mathbf{u}_0)$ and $Z(\mathbf{u}_0')$ across the study area. Under these conditions, a number of parameters of the RF exist that summarise this bivariate dependence, dependent only on $\mathbf{h}$ and not on $\mathbf{u}$. These include its covariance function, $C(\mathbf{h})$:

$$C(\mathbf{h}) = \mathrm{E}\{Z(\mathbf{u}) \cdot Z(\mathbf{u}+\mathbf{h})\} - \mathrm{E}\{Z(\mathbf{u})\} \cdot \mathrm{E}\{Z(\mathbf{u}+\mathbf{h})\} \quad (4.6)$$

and its variogram, $\gamma(\mathbf{h})$:

$$2\gamma(\mathbf{h}) = \mathrm{E}\{[Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})]^2\} \quad (4.7)$$

The adoption and parameterisation of a RF model provides a powerful framework for the prediction of unsampled values dispersed in space. The following section describes the principal considerations and tools by which this framework is implemented.

4.3 Spatial prediction with geostatistics

4.3.1 Stationarity

The existence and inference of the covariance function (4.6) or variogram (4.7) described above requires certain assumptions regarding the stationarity of the RF model. Strict stationarity entails that the multivariate cdf of the RF is invariant under translation, such that the N-point cdf of any given set of N RVs $\{Z(\mathbf{u}_1), Z(\mathbf{u}_2), \ldots, Z(\mathbf{u}_N)\}$ is the same of that of any other translated set of N RVs $\{Z(\mathbf{u}_1) + \mathbf{h}, Z(\mathbf{u}_2) + \mathbf{h}, \ldots, Z(\mathbf{u}_N) + \mathbf{h}\}$, regardless of the translational lag $\mathbf{h}$. Inference of the covariance function assumes implicitly that all p RV pairs $\{Z(\mathbf{u}_i), Z(\mathbf{u}_i+\mathbf{h}); i = 1,2,\ldots, p\}$ separated by the same lag $\mathbf{h}$ share the same two-point cdf. Under these conditions the two-point covariance $C(\mathbf{u}_0, \mathbf{u}_0')$ is dependent only on lag and not on location, allowing all RV pairs $(Z(\mathbf{u}_0), Z(\mathbf{u}_0'))$ with that separation vector to be used in inference of the covariance function for that lag, $C(\mathbf{h})$. This location independence of the two-point cdf, in addition to the expectation $\mathrm{E}\{Z(\mathbf{u})\}$, is termed *stationarity of order two* or *second-order stationarity*, defined formally by:

$$\mathrm{E}\{Z(\mathbf{u})\} = m \quad \forall\ \mathbf{u} \in \mathcal{A} \quad (4.8)$$

and:

$$C(\mathbf{h}) = \mathrm{E}\{Z(\mathbf{u}) \cdot Z(\mathbf{u}+\mathbf{h})\} - m^2 \quad \forall\ \mathbf{u} \in \mathcal{A} \quad (4.9)$$

where m is the expectation of the RF (Journel and Huijbregts, 1978).
Second-order stationarity also implies stationarity of the variogram (4.7) and leads to the following relationship between it and the covariance function:

$$\gamma(\mathbf{h}) = C(0) - C(\mathbf{h}) \tag{4.10}$$

where $C(0)$ is the covariance at zero lag, equivalent to the variance of the RF. However, the definition of the variogram does not require second-order stationarity. In addition to condition (4.8) above, it is sufficient that the *increments* of the RF, $[Z(\mathbf{u}) - Z(\mathbf{u} + \mathbf{h})]$, are second-order stationary. This is termed *intrinsic stationarity*.

When defining stationarity, it must be stressed that it is simply a property of the RF model and, hence, a modelling decision which is necessary for statistical inference. Stationarity is not a real-world characteristic of the phenomenon of interest or a hypothesis that can be tested. For a given data set, however, stationarity can be judged subjectively as an appropriate or inappropriate modelling decision and this may depend on the objectives of the study and the nature of sampling as well as the underlying characteristics of the phenomenon. Alternative approaches that can be implemented when a stationarity RF is considered inappropriate are discussed later.

4.3.2 Inferring second-order moments of the random function model

4.3.2.1 Variogram estimation

It is necessary to characterise the dependence between RV pairs with different separations $\mathbf{h}$ for use in prediction algorithms such a kriging, described later. This requirement underpins the rationale for inferring the covariance function or variogram of the RF, as defined above. The most common approach taken is to estimate the variogram with the n sample data using a straightforward method-of-moments approach. For each lag, $\mathbf{h}$, the sample (semi)variogram[2] $\hat{\gamma}(\mathbf{h})$ can be estimated as half the mean squared

[2] Strictly, the term *variogram* refers to the function $2\gamma(\mathbf{h}) = E\{[Z(\mathbf{u}) - Z(\mathbf{u} + \mathbf{h})]^2\}$, and the *semivariogram* refers to this function divided by two, hence, $\gamma(\mathbf{h})$. It is the latter value that is generally estimated and, because of its useful relation to the covariance function (4.10), used in interpolation algorithms. Hereafter, the term variogram is used in place of semivariogram, accepting that strictly they refer to different values.

difference between all $i = 1,2,\ldots, p$ data pairs separated by that lag:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2p(\mathbf{h})}\sum_{i=1}^{p(\mathbf{h})}[z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})]^2 \tag{4.11}$$

Semivariance values can be calculated for every data pair in the data set and compared to the corresponding lags, **h**, by plotting the resulting variogram cloud. An alternative approach is to pool data pairs according to a finite set of regularly spaced lags, with each value of **h** actually representing a defined range of lag separations. The latter approach allows a larger sample, and hence a more stable estimate, for each value of $\hat{\gamma}(\mathbf{h})$. A further issue is the effect of direction on the variogram. Where semivariance is dependent only on the separation distance, $|\mathbf{h}|$, the variogram is deemed *isotropic*. Where the direction of separation also has an effect, the variogram is deemed *anisotropic*. In the latter case, data pairs are generally pooled by both distance and direction, and separate sample variograms estimated for each direction.

4.3.2.2 Variogram modelling

Variogram inference using the sample variogram defined above (4.11) leads to sample values of $\hat{\gamma}(\mathbf{h}_k)$ at a finite number of lags (and possibly directions) $k = 1,2,\ldots,K$. Because interpolation algorithms such as kriging require semivariance values for any possible lag **h**, it is necessary to fit a continuous model $\tilde{\gamma}(\mathbf{h})$ to the K sample values. When choosing variogram models to fit to the sample values, it is imperative that the model chosen is deemed permissible. Of critical importance is that when any given RV, Y, is created as a finite linear combination of $i=1,2,\ldots,n$ RVs $Z(\mathbf{u}_i)$ across the study area, the variance of Y is non-negative. Such a linear combination is expressed formally as:

$$Y = \sum_{i=1}^{n}\lambda_i Z(\mathbf{u}_i) \tag{4.12}$$

The variance of Y is expressed as a linear combination of covariance values:

$$\mathrm{Var}[Y] = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C(\mathbf{u}_i - \mathbf{u}_j) \tag{4.13}$$

and the covariance function $C(\mathbf{h})$ must be chosen such that $\mathrm{Var}[Y] \geq 0$. Covariance functions that fulfil this requirement are deemed *positive definite*. Because variogram models are used ultimately in kriging algorithms to calculate covariances, it follows that the model $\tilde{\gamma}(\mathbf{h})$ must also result in non-negative variances for Y. Accounting for an RF model that is only intrinsically stationary (and, hence, the covariance function does not exist), the variance of Y can be expressed in terms of the variogram as:

$$\mathrm{Var}[Y] = -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \gamma(\mathbf{u}_i - \mathbf{u}_j) \tag{4.14}$$

Variogram models that ensure non-negativity of the variance of Y are termed *conditionally negative definite*. The condition is that the sum of the weights λ_i is zero, which is necessary to remove the covariance term $C(0)$ from the expression. In practice, rather than exhaustively check any given model for conditional negative definiteness, variogram models are selected from a set of established models that are known to meet this condition (e.g. Journel and Huijbregts, 1978, Ch. 3; Goovaerts, 1997, Ch. 4; Deutsch and Journel, 1998, Ch. 2). Of this set, those variogram models considered in this study are:

the spherical model:

$$\tilde{\gamma}(h) = \begin{cases} c\cdot\left[1.5\dfrac{h}{a} - 0.5\left(\dfrac{h}{a}\right)^3\right] & \text{if } h \leq a \\ c & \text{if } h > a \end{cases} \tag{4.15}$$

the exponential model:

$$\tilde{\gamma}(h) = c\cdot\left[1 - \exp\left(-\frac{3h}{a}\right)\right] \tag{4.16}$$

the Gaussian model:

$$\tilde{\gamma}(h) = c \cdot \left[1 - \exp\left(-\frac{(3h)^2}{a^2} \right) \right] \tag{4.17}$$

the power model (where ω is a power $0 < \omega < 2$) :

$$\tilde{\gamma}(h) = c \cdot h^{\omega} \tag{4.18}$$

and the periodic model:

$$\tilde{\gamma}(h) = c \cdot \left[1 - \cos\left(\frac{h}{a} \cdot \pi \right) \right] \tag{4.19}$$

where h is the distance component of the lag vector (h = $|\mathbf{h}|$), c is the structural component or *sill* parameter, and a is the *range* parameter (Deutsch and Journel, 1998, p. 25). These parameters are shown for a hypothetical spherical variogram model in Figure 4.1, and each model type is shown in Figure 4.2.

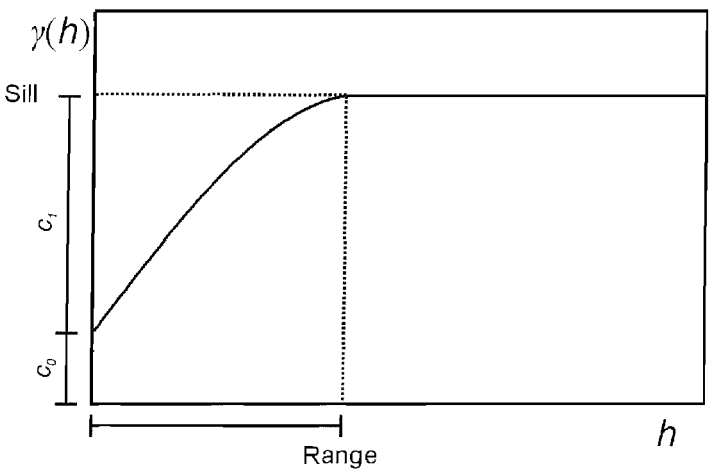


Figure 4.1 A hypothetical variogram model. Variograms plot semivariance (γ) against, in this case, omnidirectional distance, h. The model shown is a nested structure consisting of a nugget effect model with structural component c_0, and a spherical model with structural component c_1. The sill value is the sum of these structural components, and the distance at which this sill is reached is termed the range.

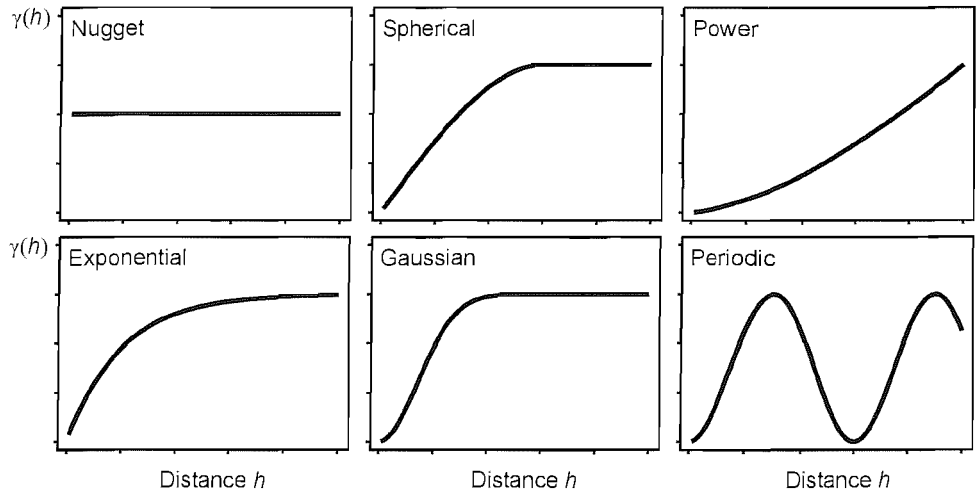


Figure 4.2 Examples of six of the most common permissible variogram models.

The sill parameter is the limiting value $\gamma(\infty)$, equivalent to the *a priori* variance $C(0)$ of a stationary RF. Variogram models that reach a sill are deemed *bounded* or *transitive*. The lag distance h at which the sill is reached is represented by the range parameter. The spherical model reaches its sill at value a, the *actual range*. In the case of the exponential and Gaussian models, however, the sill is reached asymptotically and a *practical range* is therefore defined as the distance at which the variogram reaches 95% of its sill, $\gamma(a) = 0.95 \cdot c$. The range represents the separation distance beyond which pairs of RVs are modelled as independent, that is, no spatial dependence exists. The power model is an example of an *unbounded* model and does not reach a sill. Such models represent RFs with an unlimited capacity for spatial dispersion for which neither the *a priori* variance nor covariance function can be defined. The periodic model is used to represent RFs in which a pattern is repeated regularly through space. In this case, the range parameter defines the distance to the first peak, equivalent to the size of the underlying cyclic feature (Deutsch and Journel, 1998, p. 25). When modelling semivariance through time, a seasonally repeating feature can be modelled using a periodic model with a range of 6 months which results in a period of 12 months. The periodic model is conditionally negative definite in 1-D only.

The value of the variogram at $\mathbf{h} = 0$ is strictly 0. Often, however, sample semivariance values suggest a model should be used that intercepts the ordinate at some positive

value. This discontinuity is termed the *nugget effect*, and is modelled with a nugget model, defined simply as:

$$\tilde{\gamma}(h) = \begin{cases} 0 & \text{if } h = 0 \\ c & \text{if } h > 0 \end{cases} \tag{4.20}$$

The nugget effect occurs when the expected difference $z(\mathbf{u}) - z(\mathbf{u} + \mathbf{h})$ remains positive even when the separation $\mathbf{h}$ tends to zero. This expected difference can be caused by several factors including measurement error, spatial variability over distances smaller than the shortest sampling interval, or non-spatial sources of variability that may operate independently at very close locations. In cases where the sample variogram suggests a complete absence of spatial auto-correlation (a *pure nugget* effect), the nugget effect model can be used in isolation.

In many situations, two or more separate models are used together as a linear combination to form a *nested* variogram model. For example, a nested model $\tilde{\gamma}(\mathbf{h})$ could be constructed using a nugget effect $\tilde{\gamma}_{nug}(\mathbf{h})$, Gaussian $\tilde{\gamma}_{gau}(\mathbf{h})$ and spherical $\tilde{\gamma}_{sph}(\mathbf{h})$ model as $\tilde{\gamma}(\mathbf{h}) = \tilde{\gamma}_{nug}(\mathbf{h}) + \tilde{\gamma}_{gau}(\mathbf{h}) + \tilde{\gamma}_{sph}(\mathbf{h})$. The sill value of a nested variogram is defined by the sum of the structural components of each constituent model. The ratio of the nugget model to this sill is termed the *relative nugget effect* and is indicative of the proportion of the total variance of the RF that is not due to spatial variability.

4.3.3 Kriging

Consider a set of spatial data, $z(\mathbf{u}_\alpha)$, of an attribute z at n locations $\mathbf{u}_\alpha$, $\alpha = 1,2,\ldots, n$ and a set of q unsampled locations, $\mathbf{u}_\beta$, $z^*(\mathbf{u}_\beta)$, $\beta = 1,2,\ldots, q$ for which predictions are required. Kriging is a geostatistical term for a family of generalised linear regression techniques that provides an approach by which the available data $z(\mathbf{u}_\alpha)$ can be used to predict values $z^*(\mathbf{u}_\beta)$ at the unsampled locations (Krige, 1951; Matheron, 1971). Kriging techniques operate within the conceptual framework provided by the RF model and exploit spatial dependence in the phenomenon of interest, as modelled by the covariance function or variogram. Interpreting each datum $z(\mathbf{u}_\alpha)$ as a realisation of the RV $Z(\mathbf{u}_\alpha)$, the kriging

predictor $Z^*(\mathbf{u}_0)$ can be expressed as a basic linear regression predictor:

$$Z^*(\mathbf{u}_0) - m(\mathbf{u}_0) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}_0)[Z(\mathbf{u}_\alpha) - m(\mathbf{u}_\alpha)] \tag{4.21}$$

where λ_α is the weight assigned to the datum corresponding to $Z(\mathbf{u}_\alpha)$ and $m(\mathbf{u}_\alpha)$ and $m(\mathbf{u}_0)$ are the expected values of the RVs $Z(\mathbf{u}_\alpha)$ and $Z(\mathbf{u}_0)$, respectively. The prediction error can also be defined as a random variable $Z^*(\mathbf{u}_0)$ - $Z(\mathbf{u}_0)$ and the objective of kriging is to minimise the variance of this error, termed the error variance $\sigma_E^2(\mathbf{u}_0)$, under the constraint of unbiasedness (i.e. under the constraint that E{ $Z^*(\mathbf{u}_0)$ - $Z(\mathbf{u}_0)$ } = 0). Numerous variations of kriging exist, each targeted at subtly different prediction problems. Two widely used approaches are of relevance in this project, simple kriging (SK) and ordinary kriging (OK), and these are now discussed in more detail.

4.3.3.1 Simple kriging

The RF $Z(\mathbf{u})$ can be decomposed into a trend component $m(\mathbf{u})$ and a residual component $R(\mathbf{u})$: $Z(\mathbf{u}) = R(\mathbf{u}) + m(\mathbf{u})$. In SK, the trend component is modelled as a known stationary mean m which allows the basic predictor (4.21) to be re-expressed as the SK estimator $Z^*_{SK}(\mathbf{u}_0)$:

$$Z^*_{SK}(\mathbf{u}_0) = \sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha(\mathbf{u}_0) Z(\mathbf{u}_\alpha) + \left[1 - \sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha(\mathbf{u}_0)\right] \cdot m \tag{4.22}$$

where λ_α are termed the kriging weights. A minimisation procedure (e.g. see Goovaerts, 1997, p. 128) can be implemented to derive a system of equations in terms of Z-covariances for the set of $n(\mathbf{u}_0)$ kriging weights that minimise the error variance $\sigma_E^2(\mathbf{u}_0)$:

$$\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta(\mathbf{u}_0) C(\mathbf{u}_\alpha - \mathbf{u}_\beta) = C(\mathbf{u}_\alpha - \mathbf{u}_0) \qquad \alpha = 1,2,\ldots,n(\mathbf{u}_0) \tag{4.23}$$

The two covariance terms are $C(\mathbf{u}_\beta$ - $\mathbf{u}_\alpha)$, the covariance between RVs at two data

locations $Z(\mathbf{u}_\alpha)$ and $Z(\mathbf{u}_\beta)$, and $C(\mathbf{u}_\alpha - \mathbf{u}_0)$, the covariance between the RV at a given data location $Z(\mathbf{u}_\alpha)$ and at the prediction location $Z(\mathbf{u}_0)$. These covariance values are calculated using a covariance function or, more commonly from a variogram model fitted to the sample variogram as described previously, which is then converted into a covariance function using relation (4.10).

The system of equations (4.23) is solved using a matrix operation. In matrix notation, these equations are written as:

$$\mathbf{K}_{SK} \,.\, \lambda_{SK}(\mathbf{u}_0) = \mathbf{k}_{SK} \tag{4.24}$$

Where $\mathbf{K}_{SK}$ is a $n(\mathbf{u}) \times n(\mathbf{u})$ matrix of the covariances between RVs at data locations, $\mathbf{k}_{SK}$ is a vector of the $n(\mathbf{u}_0)$ covariances between RVs at data locations and the prediction location, and λ_{SK} is a vector of the $n(\mathbf{u}_0)$ kriging weights:

$$\begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}_1) & \cdots & C(\mathbf{u}_1 - \mathbf{u}_{n(\mathbf{u}_0)}) \\ \vdots & \vdots & \vdots \\ C(\mathbf{u}_{n(\mathbf{u}_0)} - \mathbf{u}_1) & \cdots & C(\mathbf{u}_{n(\mathbf{u}_0)} - \mathbf{u}_{n(\mathbf{u}_0)}) \end{bmatrix} \cdot \begin{bmatrix} \lambda_1(\mathbf{u}) \\ \vdots \\ \lambda_{n(\mathbf{u}_0)}(\mathbf{u}) \end{bmatrix} = \begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}) \\ \vdots \\ C(\mathbf{u}_{n(\mathbf{u}_0)} - \mathbf{u}) \end{bmatrix} \tag{4.25}$$

The vector of kriging weights is obtained by inverting the covariance matrix $\mathbf{K}_{SK}$ and multiplying the resulting matrix $\mathbf{K}^{-1}{}_{SK}$ by the covariance vector $\mathbf{k}_{SK}$:

$$\lambda_{SK}(\mathbf{u}_0) = \mathbf{K}^{-1}{}_{SK} \,.\, \mathbf{k}_{SK} \tag{4.26}$$

The minimised error variance, termed the kriging variance $\sigma^2_{SK}(\mathbf{u}_0)$, is defined as :

$$\sigma^2_{SK}(\mathbf{u}_0) = C(0) - \sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha(\mathbf{u}_0) C(\mathbf{u}_\alpha - \mathbf{u}_0) \tag{4.27}$$

and is also calculated using the matrices defined above by:

$$\sigma^2_{SK}(\mathbf{u}_0) = C(0) - \mathbf{k}^{T}{}_{SK} \,.\, \mathbf{K}^{-1}{}_{SK} \,.\, \mathbf{k}_{SK} \tag{4.28}$$

4.3.3.2 Ordinary kriging

OK differs conceptually from SK in the way the RF trend component $m(\mathbf{u})$ is modelled. Rather than consider $m(\mathbf{u})$ to be a known stationary mean m, OK considers the mean to be unknown and limits its domain of stationarity to a local neighbourhood centred on the location $\mathbf{u}_0$ to be predicted. This approach has the important practical implication that the local mean may vary considerably over the study area which, in practice, is often considered a more appropriate modelling strategy. Under OK, the basic linear predictor (4.22) is expressed as a linear combination of the $n(\mathbf{u}_0)$ RVs $Z(\mathbf{u}_\alpha)$ and the constant local mean $m(\mathbf{u}_0)$:

$$Z^*(\mathbf{u}_0) = \sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha(\mathbf{u}_0) Z(\mathbf{u}_\alpha) + \left[1 - \sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha(\mathbf{u}_0) \right] \cdot m(\mathbf{u}_0) \tag{4.29}$$

In order to remove the unknown local mean $m(\mathbf{u}_0)$ from the expression, and to ensure the unbiasedness of the predictor (i.e. that $E\{Z^*(\mathbf{u}_0) - Z(\mathbf{u}_0)\} = 0$), the sum of weights is constrained to sum to 1, thus removing the second term. This allows the OK predictor $Z^*_{\mathrm{OK}}(\mathbf{u}_0)$ to be expressed as:

$$Z^*_{\mathrm{OK}}(\mathbf{u}_0) = \sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha(\mathbf{u}_0) Z(\mathbf{u}_\alpha) \tag{4.30}$$

with the unbiasedness constraint:

$$\sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha(\mathbf{u}_0) = 1 \tag{4.31}$$

As with SK, a minimisation procedure (e.g. see Isaaks and Srivastava, 1989, p. 286 - 289) can be implemented to derive a system of equations in terms of Z-covariances for the set of $n(\mathbf{u}_0)$ kriging weights that minimise the error variance $\sigma^2_E(\mathbf{u}_0)$ of the above predictor under the constraint of unbiasedness:

$$\begin{cases} \sum_{\beta=1}^{n(\mathbf{u}_0)} \lambda_\beta(\mathbf{u}_0) C(\mathbf{u}_\alpha - \mathbf{u}_\beta) + \mu_{OK}(\mathbf{u}_0) = C(\mathbf{u}_\alpha - \mathbf{u}_0) \qquad \alpha = 1,2,...,n(\mathbf{u}_0) \\ \sum_{\beta=1}^{n(\mathbf{u}_0)} \lambda_\beta(\mathbf{u}_0) = 1 \end{cases} \tag{4.32}$$

The two covariance terms in (4.32) correspond to those in the equivalent SK system (4.23) and, although the mean $m(\mathbf{u}_0)$ is assumed stationary only within local neighbourhoods, the covariance is generally inferred from all data available across the study area. The term $\mu_{OK}(\mathbf{u}_0)$ is the *Lagrange parameter* (e.g. see James, 2001, p. 655) and is introduced as part of the minimisation procedure to maintain the balance of $n(\mathbf{u}_0)$ + 1 equations and $n(\mathbf{u}_0)$ + 1 unknowns that is upset by the addition of the further equation for the constrained weights (4.31).

As for SK, the OK system of equations (4.32) is solved using a matrix operation, written as:

$$\mathbf{K}_{OK} \,.\, \boldsymbol{\lambda}_{OK}(\mathbf{u}_0) = \mathbf{k}_{OK} \tag{4.33}$$

The addition of the Lagrange parameter alters these matrices from the SK case as follows:

$$\begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}_1) & \cdots & C(\mathbf{u}_1 - \mathbf{u}_{n(\mathbf{u}_0)}) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C(\mathbf{u}_{n(\mathbf{u}_0)} - \mathbf{u}_1) & \cdots & C(\mathbf{u}_{n(\mathbf{u}_0)} - \mathbf{u}_{n(\mathbf{u}_0)}) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_1(\mathbf{u}_0) \\ \vdots \\ \lambda_{n(\mathbf{u}_0)}(\mathbf{u}_0) \\ \mu(\mathbf{u}_0) \end{bmatrix} = \begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}_0) \\ \vdots \\ C(\mathbf{u}_{n(\mathbf{u}_0)} - \mathbf{u}_0) \\ 1 \end{bmatrix} \tag{4.34}$$

The vector of kriging weights is obtained in the same way, by inverting the covariance matrix $\mathbf{K}_{OK}$ and multiplying the resulting matrix $\mathbf{K}^{-1}{}_{OK}$ by the covariance vector $\mathbf{k}_{OK}$:

$$\boldsymbol{\lambda}_{OK}(\mathbf{u}_0) = \mathbf{K}^{-1}{}_{OK} \,.\, \mathbf{k}_{OK} \tag{4.35}$$

The ordinary kriging variance $\sigma^2_{\mathrm{OK}}(\mathbf{u}_0)$, is defined as :

$$\sigma^2_{\mathrm{OK}}(\mathbf{u}_0) = C(0) - \sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha(\mathbf{u}_0) C(\mathbf{u}_\alpha - \mathbf{u}_0) - \mu_{OK}(\mathbf{u}_0) \tag{4.36}$$

and is calculated as:

$$\sigma^2_{\mathrm{OK}}(\mathbf{u}_0) = C(0) - \mathbf{k}^{\mathrm{T}}{}_{\mathrm{OK}} \cdot \mathbf{K}^{-1}{}_{\mathrm{OK}} \cdot \mathbf{k}_{\mathrm{OK}} \tag{4.37}$$

4.3.3.3 Features of kriging predictors

Both OK and SK are *exact interpolators* such that all data values $z(\mathbf{u}_\alpha)$ are honoured at their locations, $z^*(\mathbf{u}_0) = z(\mathbf{u}_\alpha) \ \forall \ \mathbf{u}_0 = \mathbf{u}_\alpha$, both produce unbiased predictions in the sense that $\mathrm{E}\{ Z^*(\mathbf{u}_0) - Z(\mathbf{u}_0) \} = 0$, and both minimise the modelled prediction error variance $\sigma^2_E(\mathbf{u}_0) = \mathrm{var}[Z^*(\mathbf{u}_0) - Z(\mathbf{u}_0)]$. The weighting system used by the OK and SK predictors takes into account both the proximity of each datum to the prediction location (via the covariance term $C(\mathbf{u}_\alpha - \mathbf{u}_0)$) and the proximity between data (via the covariance term $C(\mathbf{u}_\alpha - \mathbf{u}_\beta)$), with the latter consideration accounting for redundancy between data.

In addition to the provision of a model for the variogram or covariance function, implementation of either predictor requires various parameters to be set by the user. Of particular importance is the choice of search strategy that defines the number $n(\mathbf{u})$ of local data $\{z(\mathbf{u}_\alpha), \alpha = 1,2,\ldots,n\}$ that are used in each prediction $z^*(\mathbf{u})$. Generally, a local search radius or limit value is implemented such that $n(\mathbf{u})$ is restricted to substantially smaller than the total number of data available across the study area. This practice is motivated by various factors. Firstly, the reliability of covariance estimates for large separation distances $|\mathbf{h}|$ is questionable since the number of data pairs with this separation is often small. Secondly, the use of a local search neighbourhood with OK allows local fluctuations in the mean to be taken into account. Thirdly, the influence of distant data is generally screened by those more proximate, such that their inclusion has little effect on the prediction. A further benefit of reducing the number of data involved in each prediction is that the computational requirements decrease dramatically, with

processing time approximately proportional to $(n(\mathbf{u}))^3$.

OK is preferred to SK in many situations because it requires neither knowledge or stationarity of the mean over the entire study area (Goovaerts, 1997). The difference between OK and SK predictions at a given location is dependent on the departure of the local mean from the global mean. In regions with a small local mean, the OK estimate will be smaller than the SK estimate, with the converse applying for regions with a larger local mean. The difference between OK and SK predictions increases as predictions are made at locations more distant from any data, since the relative influence of the mean increases in these situations.

4.3.3.4 Kriging variance as a measure of prediction uncertainty

The kriging variance serves as a criterion for optimisation of the kriging equations as described above. However, it also provides useful information about each prediction. The kriging variance is dependent on the variogram or covariance model and on the spatial configuration of the data in relation to the prediction location. Because of this dependence, the kriging variance provides a measure of the uncertainty of each prediction, with uncertainty increasing for RFs with large spatial variance, and for predictions that are made at locations more distant from data. The kriging variance is not dependent, however, on the data values, such that any two sets of data with different values but the same spatial configuration would yield a prediction with the same kriging variance. This independence reduces the utility of the kriging variance as an absolute measure of uncertainty since, for example, it is intuitive that a local set of data with large variability will result in a less certain estimate than a less variable set with the same spatial configuration. As such, the use of kriging variance is generally restricted to a relative measure of uncertainty, allowing relative comparison of the uncertainty of individual predictions and different data configurations.

4.3.3.5 Effect of variogram structures on kriging predictions

Having outlined the kriging process, and explained the role of the variogram model, it is appropriate to consider how the characteristics of the variogram model affect the

resulting predictions. If two variograms are considered that differ only in scale, e.g. $\gamma(\mathbf{h})$ and $2\gamma(\mathbf{h})$, the values of the resulting kriging predictions will not differ. This is because the relative influence of different data does not change and, hence, the kriging weights remain unaltered. The kriging variance, however, *is* affected in proportion to the change in scale. The shape of the variogram model, as determined by the choice and parameterisation of the constituent permissible models, can have a substantial effect on both prediction values and kriging variance. Models with a parabolic shape close to the origin, such as the Gaussian model, are best suited to representing very continuous phenomena, and result in much larger influence being attributed to data very close to the prediction location. Models such as the spherical and exponential model have a linear shape close to the origin, which leads to the influence of nearby data declining more evenly with increasing separation from the prediction location than is the case for the Gaussian model. Models with a large nugget effect mean that the relative importance of data proximity in determining influence is small. In the extreme case, with a pure nugget effect model, the influence of proximity is zero and all data are weighted equally. Under these conditions, the prediction is equivalent to the mean of the data. The variogram range reflects the maximum distance over which spatial dependence exists, such that points separated by greater distances are deemed independent.

For a given data set and an RF model with a given *a priori* variance, variogram models with a small nugget and large range values result in relatively more certain kriging predictions than do models with a large relative nugget effect and/or a small range value. The equivalent real-world interpretations are that, in the former scenario, the property of interest varies smoothly through space and is spatially dependent over large distances whilst, in the second scenario, the property varies erratically in space over short distances such that there is only a weak tendency for proximate points to be more similar than those much further apart.

4.3.3.6 Cross-validation

Having implemented a kriging predictor to make a set of predictions $z^*(\mathbf{u}_\beta)$ at β = 1,2,...,q unsampled locations, it is generally necessary to assess the accuracy of these predictions, that is to assess the values of $z^*(\mathbf{u}_\beta) - z(\mathbf{u}_\beta)$. However, in genuine prediction

settings the set of q true values $z(\mathbf{u}_\beta)$ is, by definition, unknown such that the accuracy of predictions cannot be assessed directly. An alternative strategy is provided by cross-validation which allows the prediction method to be tested at the locations of the existing observations. Cross-validation proceeds by the removal of a single datum, $z(\mathbf{u}_\alpha)$. The kriging technique in question is then implemented to obtain a prediction $z^*(\mathbf{u}_\alpha)$ at this point, and the error between datum and prediction $z^*(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha)$ is noted. The datum is then replaced, another removed, and the process begins again, eventually repeating for all $\alpha = 1,2,\ldots, n$ data locations to provide a complete set of predicted values for comparison with the data set.

A series of summary statistics can be calculated from the set of cross-validation predictions to allow, for example, straightforward comparison between different prediction approaches. Summary statistics used in this project include the correlation coefficient between the predicted and actual set:

$$\rho\left[z^*(\mathbf{u}_\alpha), z(\mathbf{u}_\alpha)\right] \tag{4.38}$$

the mean prediction error (ME):

$$\mathrm{ME} = \frac{1}{n}\sum_{\alpha=1}^{n} z^*(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha) \tag{4.39}$$

and the mean absolute prediction error (MAE) (Saito and Goovaerts, 2000):

$$\mathrm{MAE} = \frac{1}{n}\sum_{\alpha=1}^{n}\left|z^*(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha)\right| \tag{4.40}$$

The correlation coefficient provides a straightforward measure of linear association between the data and prediction sets, the ME provides a measure of the bias of the predictor, and the MAE provides a measure of the mean accuracy of individual predictions.

The use of cross-validation as a method of accuracy assessment is limited by a number

of factors. Firstly, although each datum is removed temporarily to generate a cross-validation prediction at that point, the variogram is not recalculated with the datum removed and, hence, each cross-validation prediction is not strictly independent of the datum to which it is compared. Where the number of data is large, however, and no extreme outliers are present, the influence of an individual datum on the sample variogram can be considered negligible in most cases. Secondly, the use of simple arithmetic averages to generate estimates of ME and MAE may result in biased estimates when the data are clustered, and this issue is revisited in later chapters.

4.4 Space-time geostatistics

Geostatistics was conceived as an approach for the investigation and prediction of natural phenomena distributed across space. The restriction of the conceptual approach to the spatial domain was appropriate for the geological settings in which the paradigm became established. The estimation of ore reserves, for example, requires no consideration of the temporal domain since, in the timescales that are likely to be of interest, the variability through time of the property under study can be considered negligible. As the range of disciplines in which geostatistical tools have been applied has expanded, however, prediction scenarios have been increasingly encountered in which variability of a property through time, as well as space, is of interest. Where data on such properties are themselves collected at appreciably different times, and where predictions are required at unsampled points in time as well as space, it is clear that an approach is required in which the temporal domain, as well as the spatial domain, is considered explicitly.

Space-time geostatistics is a broad term that incorporates a diverse set of approaches in which geostatistical concepts and tools developed originally for spatial-only settings have been adapted for the characterisation and prediction of properties that vary, and are investigated, through both time and space. Examples of the use of space-time geostatistical approaches can be found in a wide range of disciplines. Such techniques have been used to model the space-time distribution of pollutants in the atmosphere (Casado et al., 1994; Christakos and Vyas, 1998; Meiring et al., 1998; De Iaco et al., 2002; Host et al., 2004; Nunes and Soares, 2005) and how such pollutants are deposited

on the land surface (Bilonick, 1985; Haas, 1990, 1995; Kyriakidis and Journel, 2001). Similarly, the dispersion of chemicals has been modelled in oceanographic (Lophaven et al., 2006) and groundwater settings (D'Agostino et al., 1998; Douaik et al., 2005; Vanderlinden et al., 2006). Geostatistical space-time interpolation and simulation techniques have been used in hydrological studies (Christakos et al., 2000; Araghinejad and Burn, 2005) and the assessment of soil- and groundwater resources (Rouhani and Myers, 1990; Snepvangers et al., 2003; Jost et al., 2005). Further applications include the modelling of spatiotemporal patterns in air temperature (Bogaert and Christakos, 1997), the evaluation of long-term wind-field strength as a source of renewable energy (Haslett and Raftery, 1989), and the simulation of regional daily precipitation (Kyriakidis et al., 2004).

4.4.1 Approaches to space-time geostatistical modelling

Of the broad swathe of conceptual approaches by which space-time variables can be represented in a geostatistical framework, two distinct strategies have been identified (Kyriakidis and Journel, 1999). The first strategy is to model the variable as either a set of temporally correlated spatial RFs at T points in time (multiple RF model) or a set of spatially correlated time series (TS) located at n locations in space (multiple TS model). The choice between these two sub-strategies is likely to be motivated by the relative abundance of data in the two domains. Where a large number of data have been collected through time at a small number of locations in space, the multiple TS model is likely to be more appropriate. Conversely, where data have been collected densely in space but at only a small number of times, then the multiple RF model may be more suitable. In both cases the spatiotemporal continuity is modelled using the linear model of coregionalisation (LMC). In this model, every T spatial RF or n TS is characterised with, respectively, a spatial or temporal variogram or covariance function. Temporal continuity between all T RFs is characterised by $T(T - 1)/2$ cross-variograms or cross-covariance functions, whilst spatial continuity between n TS is characterised by $n(n - 1)/2$ such functions. In this LMC approach, predictions are made using cokriging (e.g. see Goovaerts, 1997, p. 203 - 258) and can be made only at space time locations within each spatial RF or TS. This restriction represents a limitation to the use of an LMC model in situations where predictions are required at any space-time location within the

spatiotemporal domain of interest. A further limitation is that the number of auto- and cross- variograms or covariance functions that must be estimated and modelled, ($T(T + 1)/2$ or $n(n + 1)/2$), can become impractical if both T and n are large.

A second strategy that overcomes the limitations described above is to represent the space-time phenomenon of interest using a single space-time RF $Z(\mathbf{u}, t)$ where $\mathbf{u}$ is a vector of spatial coordinates and t is an instant in time. This approach extends the RF concept introduced earlier for a spatial-only setting to include time as an additional dimension. RVs $Z((\mathbf{u}, t)_0)$ exist for all possible space-time locations $(\mathbf{u}_0, t_0)$ in the spatiotemporal study domain, each characterised by their cdf:

$$F((\mathbf{u}, t)_0; z) = \text{Prob}\{Z((\mathbf{u}, t)_0) \quad z\} \tag{4.41}$$

Any set of N space-time RVs are characterised by the corresponding N-point cdf:

$$F((\mathbf{u}, t)_1,\ldots,(\mathbf{u}, t)_N; z_1,\ldots,z_N) = \text{Prob}\{ Z((\mathbf{u}, t)_1) \quad z_1 ,\ldots, Z((\mathbf{u}, t)_N) \quad z_N \} \tag{4.42}$$

The set of all possible N-point cdfs for any value of N ($N \subseteq \mathbb{N}$) and for any choice of space-time locations constitutes the complete spatial law of the RF $Z(\mathbf{u}, t)$. Separations between space-time locations are defined by a space-time lag $(\mathbf{h}_s, h_t)$ where $\mathbf{h}_s$ is the spatial lag vector, as defined previously, and h_t is the scalar separation in time. Unlike space, there is no concept of anisotropy in time.

Mirroring the spatial-only case, the continuity of the space-time RF can be characterised by the space-time covariance function, $C(\mathbf{h}_s, h_t)$:

$$C(\mathbf{h}_s, h_t) = \text{E}\{Z(\mathbf{u}, t) \cdot Z(\mathbf{u} + \mathbf{h}_s, t + h_t)\} - \text{E}\{Z(\mathbf{u}, t)\} \cdot \text{E}\{Z(\mathbf{u} + \mathbf{h}_s, t + h_t)\} \tag{4.43}$$

and the space-time variogram, $\gamma(\mathbf{h}_s, h_t)$:

$$2\gamma(\mathbf{h}_s, h_t) = \text{E}\{[Z(\mathbf{u}, t) - Z(\mathbf{u} + \mathbf{h}_s, t + h_t)]^2\} \tag{4.44}$$

The space-time RF allows predictions to be made at any given space-time location

within the study domain. As in the spatial-only case, this process requires estimation and modelling of one of the above structural functions, and the implementation of a kriging procedure. The extension of these processes to the space-time case is now summarised.

4.4.2 Space-time variogram estimation and kriging

Consider a set of space-time data $z((\mathbf{u}, t)_\alpha)$ of the attribute z at n space-time locations $\alpha = 1,2,\ldots,n$, and a set of predictions $z^*((\mathbf{u}, t)_\beta)$ required at q unsampled space-time locations $\beta = 1,2,\ldots,q$. Mirroring the spatial-only case, an accepted geostatistical approach is to infer the space-time autocorrelation structure of the RF by using the n data to estimate a sample space-time variogram $\hat{\gamma}_{st}(\mathbf{h}_s, h_t)$ between all $i = 1,2,\ldots, p$, data pairs at a series of regular space-time lags:

$$\hat{\gamma}_{st}(\mathbf{h}_s, h_t) = \frac{1}{2p(\mathbf{h}_s, h_t)} \sum_{i=1}^{p(\mathbf{h}_s, h_t)} [z((\mathbf{u}, t)_i) - z((\mathbf{u}, t)_i + (\mathbf{h}_s, h_t))]^2 \tag{4.45}$$

A continuous 2-D space-time variogram model, $\tilde{\gamma}_{st}(\mathbf{h}_s, h_t)$, can then be fitted to this variogram surface allowing semivariance values to be estimated at any lag for input into a space-time kriging system.

The extension of spatial-only OK to space-time OK (STOK) results in an equivalent predictor $Z^*_{\text{STOK}}((\mathbf{u}, t)_0)$, a linear combination of $n(\mathbf{u}, t)$ RVs at data locations local in space and time to the prediction location:

$$Z^*_{\text{STOK}}((\mathbf{u}, t)_0) = \sum_{\alpha=1}^{n((\mathbf{u}, t)_0)} \lambda_\alpha((\mathbf{u}, t)_0) Z((\mathbf{u}, t)_\alpha) \tag{4.46}$$

with the equivalent unbiasedness constraint:

$$\sum_{\alpha=1}^{n((\mathbf{u}, t)_0)} \lambda_\alpha((\mathbf{u}, t)_0) = 1 \tag{4.47}$$

Again, the utility of this approach lies in its capability to determine the weight, $\lambda_\alpha((\mathbf{u},t)_0)$, assigned to each neighbouring datum such as to minimise the prediction variance:

$$\sigma^2_{\mathrm{STOK}}((\mathbf{u},t)_0) = Var[z^*((\mathbf{u},t)_0) - z((\mathbf{u},t)_0)] \tag{4.48}$$

4.4.3 Models for space-time covariance structures

A critical stage in the process described above is the choice of model for the variogram or covariance function and the estimation of model parameters. As in the spatial-only case, the principal concerns when modelling space-time autocorrelation structures are to ensure that the model chosen is valid (i.e. that conditional negative definiteness or positive-definiteness is ensured for variogram or covariance function models, respectively) and that the model is sufficiently flexible to allow fitting to the data though careful estimation of model parameters. Whilst a well established set of models exists for spatial-only variograms (Deutsch and Journel, 1998), a more diverse range of models have been proposed for the modelling of space-time autocorrelation structures (Kyriakidis and Journel, 1999; De Cesare et al., 2001). These include the product model (Rodriguez-Iturbe and Mejia, 1974), the metric model (Dimitrakopoulos and Luo, 1994), the integrated product model (Cressie and Huang, 1999), and the product-sum model (De Cesare et al., 2001; 2002). In this study, this last class of model was adopted because: (a) it offers a large class of flexible models that impose less constraints of symmetry between the spatial and temporal correlation components than other classes, (b) it does not require an arbitrary space-time metric to be imposed, and (c) the model can be fitted to data using relatively straightforward techniques similar to those established for spatial-only variograms.

The product-sum space-time variogram model, $\tilde{\gamma}_{st}(\mathbf{h}_s, h_t)$, is defined in terms of the separate spatial and temporal variograms, $\tilde{\gamma}_s$ and $\tilde{\gamma}_t$, and the corresponding spatial and temporal sills, $C_s(0)$ and $C_t(0)$:

$$\tilde{\gamma}_{st}(\mathbf{h}_s, h_t) = (k_1 C_s(0) + k_3)\tilde{\gamma}_t(h_t) + (k_1 C_t(0) + k_2)\tilde{\gamma}_s(\mathbf{h}_s) - k_1 \tilde{\gamma}_s(\mathbf{h}_s)\tilde{\gamma}_t(h_t) \tag{4.49}$$

The parameters k_1, k_2, and k_3 are defined as:

$$k_1 = \left[C_s(0) + C_t(0) - C_{st}(0,0)\right] / C_s(0) C_t(0) \tag{4.50}$$

$$k_2 = \left[C_{st}(0,0) - C_t(0)\right] / C_s(0) \tag{4.51}$$

$$k_3 = \left[C_{st}(0,0) - C_s(0)\right] / C_t(0) \tag{4.52}$$

where $C_{st}(0,0)$ is the sill of the space-time variogram, i.e. the limit value at large space-time lags. Various constraints are placed on these parameters to ensure model validity (see De Cesare et al., 2001). A key advantage of the product-sum model is that $\tilde{\gamma}_{st}(\mathbf{h}_s, h_t)$ is defined entirely by parameters of the sample space-marginal and time-marginal variograms and the space-time sill, $C_{st}(0,0)$, which can all be estimated from the sample space-time variogram surface (4.45). The space-marginal variogram plots the semivariance at each spatial lag for temporal lags of zero, i.e. $\hat{\gamma}_{st}(\mathbf{h}_s, 0)$, and is equivalent to the mean of all spatial-only variograms for all values of h_t. Conversely, the time-marginal variogram plots semivariance at each temporal lag for spatial lags of zero, $\hat{\gamma}_{st}(0, h_t)$, and is equivalent to the mean of all temporal-only variograms for all values of h_s.

4.5 Geostatistics and public health

The robust conceptual framework offered by the RF model and its utility in exploring and predicting heterogeneous spatial and space-time properties has resulted in the approach being applied to an increasingly diverse range of problems in many disciplines. Of particular relevance to this project are the growing number of studies in which geostatistical concepts and tools have been applied to public health problems. The most common motivation for the application of geostatistical techniques in a public health context has been the production of continuous maps of a given public health variable that has been sampled through space (and/or time). A straightforward example is provided by Carrat and Valleron (1992) who used kriging to produce maps from data on influenza morbidity in France.

4.5.1 Geostatistics and areal public health data

Frequently, data relating to public health status are available in an aggregated form over finite spatial units. The areal data might represent, for example, the incidence rate (e.g. number of cases per head of population) of a given condition within an administrative region (e.g. census enumeration district, ward, district, province) during a given period. A common requirement is the production of a smoothed risk map from such areal data that allows assessment of the spatial variability in the risk of morbidity or mortality due to the condition in question which can then be used by policy makers to identify areas of highest public health need, and to highlight potential causative factors. Examples of the use of geostatistical tools in this context include the modelling of the risk of sudden infant death syndrome (Berke, 2004), the mapping and analysis of rates of sexually transmitted diseases (Law et al., 2004), and the space-time mapping of breast cancer incidence (Christakos and Lai, 1997).

An important problem in the mapping of incidence rate or relative risk from areal-level data is that the variance of these values derived from different areal units is non-stationary because the population size will vary between units. Numerous studies have used geostatistical tools to address this problem. In a series of studies focusing on childhood cancer in the West Midlands, England, geostatistical strategies were developed to account for spatial heterogeneities in the population of children in order to produce more stable characterisation of cancer risk (Oliver et al., 1992, 1998; Webster et al., 1994). Starting with the sample variogram of the rudimentary incidence rate within electoral wards, they were able to modify the variogram to incorporate information on the number of children in each ward in order to estimate a variogram of risk. This allowed assessment of the spatial autocorrelation of the underlying risk, and prediction of this variable was carried out using cokriging based on the incidence and risk variograms. Similar approaches to this problem were suggested by Goovaerts (2005a, 2005b), Goovaerts and Jacquez (2004), and Goovaerts et al. (2005) who developed a population-weighted semivariogram estimator in order to reduce the influence of cancer incidence data based on small population sizes. In later work, Goovaerts (2005c, 2006) developed a kriging and simulation approach based on the Poisson distribution to account explicitly for the handling of count data in predictions and simulations of uncertainty.

4.5.2 Geostatistics and malaria

Just as the use of geostatistics has increased in recent years within the field of public health as a whole, so too has its application to problems specific to malaria. In an early example, Ribeiro et al. (1996) used kriging to map the distribution of mosquitoes around villages in Ethiopia in order to guide control measures. Kleinschmidt et al. (2000, 2001) used kriging to interpolate the residuals of a logistic regression model that predicted childhood malaria prevalence in West Africa using climatic, population and topographic variables. Diggle et al. (2002) used geostatistics to predict the presence or absence of malaria parasites in children in the Gambia. They analysed the influence of a range of social, environmental, and behavioural factors relating to each child and their village of residence and were able to reveal underlying spatial heterogeneities in risk. Gemperli et al. (2004) used a Bayesian hierarchical geostatistical logistic model to model the risk of infant mortality in Mali and were able to relate spatial patterns in this risk to known foci of intense malaria transmission. In a later study, Gemperli et al. (2006) used a similar approach in conjunction with a deterministic model of transmission intensity to model the spatial distribution of the entomological inoculation rate (EIR), which is the expected number of infective bites from malarial mosquitoes sustained per person in a given time period.

4.6 Chapter summary

This chapter has presented a review of the most important established geostatistical concepts and methods that have been incorporated in this study. The central concept of the random function has been introduced as a probabilistic model for the data-generating mechanism of spatial data. Variogram estimation and modelling and kriging prediction have been described as the fundamental tools by which the utility of the random function can be exploited to explore and predict spatial variables. The extension of these concepts and tools to space-time settings has been described, along with example applications and consideration of the main conceptual approaches by which this extension can be achieved. Examples of the application of geostatistics in public health and malaria studies have been given. No examples have been found of the application of geostatistical techniques to routine outpatient data or to address problems of predicting

national-level treatment burdens when such data are incomplete. Subsequent chapters present a conceptual framework and a series of analyses carried out in this project by which this problem is addressed.

Chapter 5

Conceptual Framework

Chapter 5

5. Conceptual Framework

5.1 Introduction

Having explained the background and motivation behind this project, described the principal data sets involved, and presented the main geostatistical concepts and tools of relevance, the purpose of this chapter is to provide an overview of the conceptual framework developed in this project to meet the stated aims. In Chapter 1, the overall aim was stated as being to provide reliable national and sub-national estimates of the annual outpatient treatment burden for malaria at health facilities in the formal government health sector in Kenya. Specifically, this entails the prediction of missing MC values within the HMIS database, where MC (malaria cases) is defined as the monthly count of diagnoses for malaria at each facility. In the following section, the MC variable is considered in more detail, examining the factors that are likely to determine its value at a given facility and month and the way in which these may vary through space and time. The implications of these spatial and temporal dependencies for modelling MC are then discussed and two distinct modelling strategies are identified that form the basis for the remainder of the thesis.

5.2 Conceptual exploration of the MC variable

This project is centred on the need to predict, and therefore model, the number of malaria diagnoses that are made at facilities each month as represented by the MC variable in the HMIS database. Regardless of whether a deterministic or probabilistic modelling approach is ultimately adopted, a useful preliminary exercise is to consider what *a priori* knowledge exists about the variable of interest independently of the data,

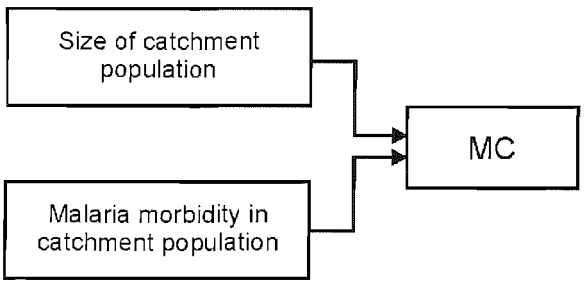


Figure 5.1 Simple conceptual model of factors determining MC.

and to construct conceptual relationships between the various contributory factors that determine its value at any given facility and month. A simple conceptual model is that, for a given facility, the value of MC is determined by the catchment population of the facility and by the level of morbidity due to malaria in that population (Figure 5.1). Both of these determining factors are now explored in more detail.

5.2.1 Factors determining malaria morbidity

The aetiology of malaria in a given population is driven by a complex array of interacting factors (Figure 5.2) and some of the most important are discussed in this section. A useful categorisation is between factors that determine the suitability of environmental conditions for the malaria parasite and vector and those that determine the susceptibility of the human population to infection and subsequent illness (Mouchet et al., 1998). The presence and intensity of malaria in a given region is determined partly by the presence and abundance of female *Anopheles* mosquitoes and of the *Plasmodium* parasite. Both are strongly dependent on minimum, maximum, and prevailing temperatures and humidity (Beier et al., 1990; Patz et al., 1998; Craig et al., 1999; Koenraadt et al., 2003; Hoshen and Morse, 2004). Furthermore, female *Anopheles* require surface water in which to lay their eggs, with a preference for temporary and turbid water bodies in which the risks of predation are small (Snow and Gilles, 2002). These dependencies mean that the suitability of the environment for malarial conditions is determined over large spatial scales by altitude and macro-climatic conditions. In many regions, marked seasonal variations in rainfall and temperature mean that conditions become suitable for some months in each year and are unsuitable for the

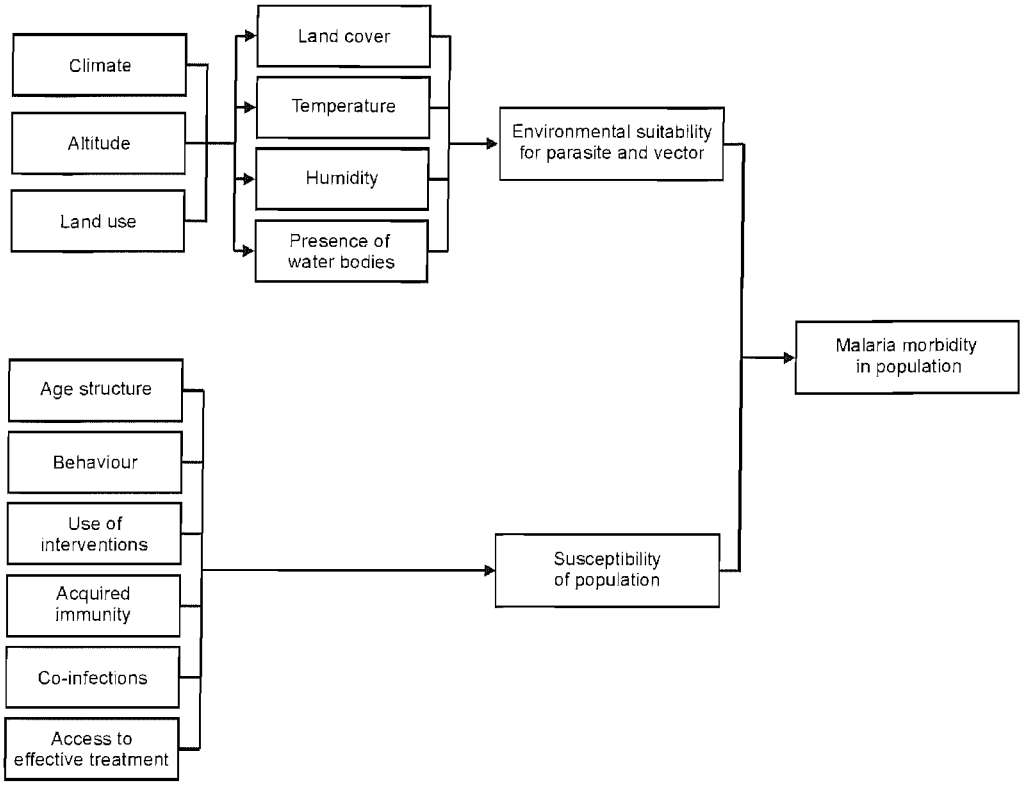


Figure 5.2 Key environmental and human determinants of malaria morbidity in a given population.

remainder. Smaller scale spatial and temporal variation (Greenwood, 1989; Snow et al., 1993; Schellenberg et al., 1998; Brooker et al., 2004) may be driven by regional and local climate and by the nature of the land surface (Afrane et al., 2005; Minakawa et al., 2005; Patz and Olson, 2006). Changes to land use and land cover can play an important role in creating breeding habitats, often caused by agricultural practices, the building of irrigation and drainage channels, and the disturbance of land due to clearance or construction (Keiser et al., 2002; Ijumba et al., 2002; Afrane et al., 2006; Munga et al., 2006). These micro-scale factors may vary considerably over short distances, and be modified over short timescales.

Given the presence of infective vectors in a region, a further myriad set of interacting factors determines the extent to which the resident population is susceptible to infection and illness from malaria. Susceptibility can be decreased, for example, by protecting against mosquito bites, particularly through the use of ITNs and the spraying of

insecticide residues inside homes (Nevill et al., 1996; Mbogo et al., 1996; Lengeler et al., 1997; Howard et al., 2000; Guyatt et al., 2002; Ter Kuile et al., 2003; Hawley et al., 2003; Gimnig et al., 2003; Muller et al., 2006), although the availability and uptake of such interventions varies according to levels of education, poverty, and other factors (Brinkmann and Brinkmann, 1995; Makemba et al., 1995; Lengeler and Snow, 1996; Marsh et al., 1996; Binka and Adongo, 1997; Cham et al., 1997; Snow et al., 1999b; Abdulla et al., 2005). Control measures can be implemented to reduce mosquito populations through the use of insecticides and larvicides targeted at breeding sites, and by environmental management to reduce the abundance of these sites (Beales and Gilles, 2002).

Because human hosts represent an integral part of the parasite life cycle, human populations are themselves a component of the ecosystem that supports parasite populations, and not simply a passive recipient of infection. As such, the distinction between environmental and human determinants of morbidity neglects the interactions between the two. If a malarious population has access to rapid treatment with effective anti-malarial drugs, for example, this not only reduces the morbidity of infected individuals directly but, by reducing or eradicating the presence of the parasite in the bloodstream, the parasite prevalence in the vector and risk of further transmission to uninfected humans is also reduced. Of particular importance in determining the pattern of morbidity is the role of acquired immunity. Individuals that are repeatedly infected with, and recover from, malaria develop a functional immunity such that the risk of morbidity and mortality are reduced for subsequent infections. In areas of intense transmission, this immunity generally develops in early childhood meaning that a disproportionate share of illness and death occurs in the very young. In areas of less intense, seasonal, or sporadic transmission, immunity may not develop or develop much more slowly such that the risk of morbidity and mortality is shared more evenly across different age groups (Snow and Gilles, 2002).

The social, biological, economic, and behavioural factors that determine malaria susceptibility are likely to vary with less spatial continuity than the climatic and habitat factors that determine overall environmental suitability. Neighbouring homesteads, for example, may differ widely in their use of interventions or ability to pay for effective treatment. A degree of spatial dependence in these factors may be expected, however,

driven by underlying regional differences in socio-economic status, levels of education, and control measures. The degree of functional immunity within different populations, in particular, may be expected to vary continuously through space, mirroring approximately the suitability of environmental conditions.

5.2.2 Factors determining the size of a facility catchment population

Various approaches exist for defining the catchment population of a health facility. When the catchment population of a given facility is discussed as a determinant of MC, the quantity of interest may be defined as the number of people who would hypothetically attend that facility to seek treatment for malaria. A useful way to consider what determines the size of a given catchment population is to start with the largest possible set of people and progressively refine this set by considering factors that act to exclude certain groups (Figure 5.3). The suitable set of people to consider initially is simply all those in the environs of the facility, which will be determined for a given region by the population density. Of this set, many may chose not to utilise any formal health facility to seek treatment (Mwenesi et al., 1995; McCombie, 1996, 2002; Amin et al., 2003). The issue of low utilisation rates of formal health services is an important one in low-income settings, and was introduced in Chapter 2. The decision to seek or not seek formal care is influenced by various cultural, social, and economic factors including the availability and perception of alternatives such as traditional or faith healers, or self treatment with home remedies or drugs purchased from the informal retail sector (Snow et al., 1992a; Ruebush et al., 1995; Goodman et al., 2004; Marsh et al., 2004; Guyatt and Snow, 2004; Amin and Snow, 2005). A further factor is the ability to pay for formal care, for the transport needed to reach the facility, or for the time taken away from work. Furthermore, the social hierarchy in place in a given community may mean that female or junior community members cannot obtain permission to leave to seek care for themselves or for children in their care (Molyneux et al., 1999, 2002). Of the subset who do choose to attend a formal health facility, not everyone in the region will have physical access to the facility in question. Factors such as the distance to the facility, the quality of the transport infrastructure, and the availability and cost of public transport mean that attendance at the facility may not be feasible in many cases (Noor et al., 2003; Tanser et al., 2006). Of those who do choose to attend a formal health facility,

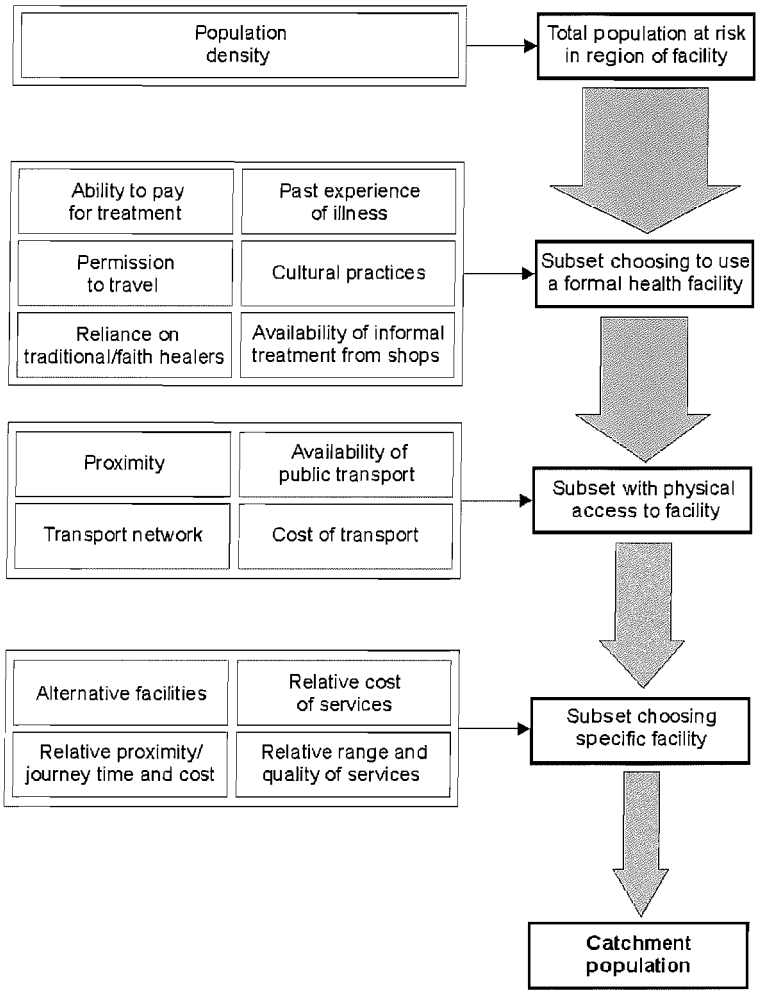


Figure 5.3 Schematic diagram showing factors determining the size of the catchment population for a given facility. Thin arrows show causation, thick arrows illustrate the progressively smaller subset of the total population that make up the catchment population. Although each set of reasons is shown as operating independently, these are likely to overlap in reality.

and who have access to the facility in question, many may chose instead to attend alternative facilities. Individuals may choose between a set of formal health facilities based on a range of factors including the distance and cost of journeys to each, the range and relative quality of services offered, and the cost of these services.

Of the various factors discussed above, many can be considered spatially independent in that they may vary substantially between facilities regardless of their proximity in space. Whilst socio-economic and cultural factors that determine care-seeking choices may display a degree of spatial dependence, other factors are entirely facility-specific. The

physical accessibility of two adjacent facilities, for example, may be nearly identical but each facility may offer very different services at different levels of quality and cost and they are therefore likely to have very different catchment populations.

5.2.3 The influence of misdiagnosis

The basic conceptual model for MC as a function of catchment population size and malaria morbidity (Figure 5.1) is likely to be too simplistic, notwithstanding the complexity of these two factors themselves. A particularly important source of uncertainty is the role of misdiagnosis which leads to a disparity between the number of outpatients who attend a given facility due to an episode of malarial illness, and the number that are incorporated in the MC variable (Figure 5.4). The causes and extent of misdiagnosis for malaria were discussed in Chapter 2, and two distinct scenarios can be identified. The non-diagnosis of true episodes of malaria (false-negative diagnosis) leads to an unknown proportion of malaria outpatient visits not contributing to MC. The incorrect diagnosis of non-malarial illness as malaria leads to an unknown number of false-positive diagnoses contributing to MC. This latter consideration means that MC is partly determined by the level of morbidity due to non-malaria conditions in the catchment population, which will be determined by a wide range of illness aetiologies. Furthermore, the facility catchment population as defined in the previous section may not be the same for malaria as for other, non-malaria, conditions since it is determined in part by the response of individuals to becoming ill with a specific condition. It is worth noting that the effects of false-negative and false-positive diagnoses on MC are opposing, such that they counteract one another. Without available data, however, the net effect is impossible

The spatial pattern of misdiagnosis is difficult to infer. It is likely that both spatial and non-spatial factors operate to determine the extent of misdiagnosis at different facilities. Misdiagnosis may be determined by factors such as consultation practices, the type of medical staff available, levels of training, and the availability of diagnostic equipment and laboratory facilities. Whilst many of these factors are facility-specific and, therefore, non-spatial, it is plausible that many of these factors are relatively uniform for a given facility type and that levels of misdiagnosis will be similar. Since no comprehensive data

are available on misdiagnosis rates across Kenya, its effect on MC cannot be quantified. As noted in Chapter 2, however, it is not the case that estimates of MC are required in which the effects of misdiagnosis have been removed because such effects contribute to defining the treatment burden.

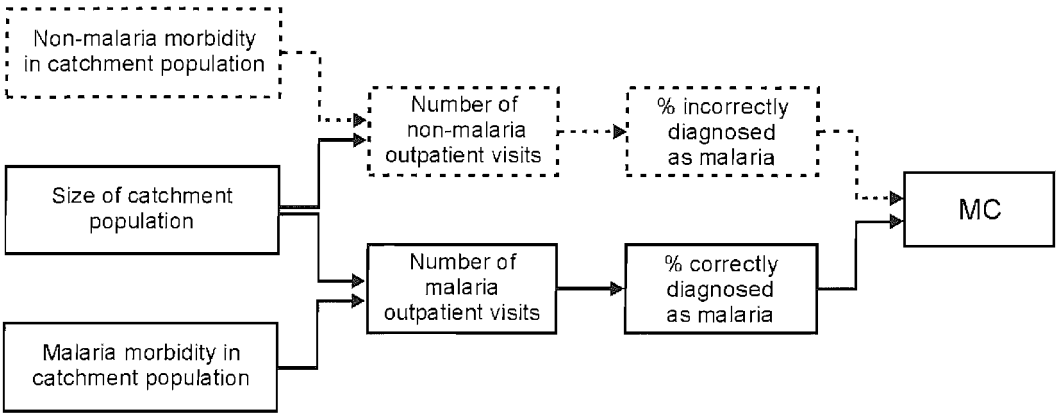


Figure 5.4 A conceptual model for the MC variable including the influence of misdiagnosis. The dashed components are the resulting non-malaria factors that may affect MC.

5.2.4 Implications for modelling MC

Having explored the factors that may determine the value of MC at any given facility and month and how these may vary in time and space, the implications for modelling unknown MC values can now be discussed. Returning to the original conceptual model presented in Figure 5.1, a reasonable expectation is that spatial variability in the first determining factor, the level of underlying malaria morbidity, is driven mainly by spatially-dependent processes operating at a series of spatial scales. In contrast, the expectation is that spatial variability in the size of each catchment population is driven by both spatially independent facility-specific effects and spatially dependent effects operating regionally and locally.

The most straightforward geostatistical modelling approach is to use the available MC data directly to predict unsampled MC values. An implicit assumption of this approach is that spatial dependence exists between MC values at different facilities. Under these conditions, the spatial structure can be characterised by estimating and modelling a

variogram, which can then be used to estimate the covariances needed for kriging predictions. A more refined approach, however, is to attempt to account for the sources of non-spatial variation in MC so that these effects can be removed. If this can be achieved, the resulting variable will vary more smoothly through space meaning that a greater proportion of its variability is spatially autocorrelated. These features will be reflected in a variogram with a smaller relative nugget effect and will, ultimately, allow kriging predictions to be made with greater accuracy. Because malaria morbidity is expected to be spatially dependent but catchment population size is expected to be largely spatially independent, such an approach may amount to standardising the raw MC data by measures of the non-spatial catchment and facility-specific factors that confound the spatial structure inherent in the underlying pattern of malaria morbidity.

Standardisation of disease incidence data is a common procedure in epidemiological and public health studies. Such data are rarely used for inference or prediction in their raw format but are usually divided by a denominator that quantifies the population that generated the incidences in order to reveal underlying spatial structure in disease risk. If the incidence data relate to an administrative unit used during a national census, for example, the population of that unit could be used as a denominator to convert the incidence count into an incidence rate (Lawson 2001). In the current setting, however, the population that generates MC values at each health facility cannot be defined easily within a discrete areal unit. Rather, the population of interest is the catchment population of each facility. For the vast majority of health facilities in Kenya, however, no direct information exists about the size of the catchment population, which means that straightforward standardisation of the MC data is not possible.

The issue of standardising MC data between different facilities to account for non-spatial variation caused by different catchment population sizes and other facility-specific factors is a central theme of this thesis. Various approaches have been explored in this project for the development of a modelling framework that incorporates such standardisation, and these are introduced in the remainder of this chapter.

5.3 Approaches to standardising MC data

Two distinct strategies were explored in this project for the standardisation of raw MC data. The first strategy was to develop approaches for estimating catchment populations from census-derived population data using geographic information system (GIS) functionality and novel spatial modelling techniques. The second strategy was to use TC data (the total number of monthly outpatient diagnoses at each facility and month) as a proxy measure of the catchment population, and to incorporate these data in the prediction framework for MC. The motivation for these two strategies is now discussed.

5.3.1 Modelling facility catchment populations

In the absence of existing information about the size of facility catchment populations across Kenya, a straightforward response is to attempt to estimate these values using a predictive model. Data on the distribution of the Kenyan population are available at fine spatial resolution from the decennial census and, in principle, the job of estimating catchment populations amounts to identifying the population subset that would attend each facility according to the factors identified in section 5.2.2. For most facilities in Kenya, however, data on many or all of these factors do not exist or are insufficient, meaning that a model that incorporates all of these factors is infeasible. The recent construction of the NHSD database, however, means that the type and location of each facility is known. Furthermore, data are available from a small number of Kenyan districts on the spatial factors that affect the way people choose to utilise formal health services. These factors have led in this project to attempts to develop methods for modelling the spatial aspects of facility catchments, with the ultimate aim of producing estimates of catchment population size that can act as a standardising denominator to the raw MC data. This work is presented in Chapter 6.

5.3.2 Incorporating TC data

As explained in Chapter 3, every MC datum in the HMIS database was accompanied by a corresponding TC datum detailing the total number of all-cause diagnoses for each facility and month. As an alternative to deriving catchment population estimates, a

strategy was devised that used these TC data as a way of standardising the raw MC values. The rationale was that TC reflects the overall level of use of each facility as driven by its type, size and utilisation and therefore acts as a useful proxy measure of catchment population size. The obvious limitation of this use of TC data is that values are only available for the same facilities and months for which MC data also exist. This co-location means that TC values are not available for those points at which predictions of MC are required and, as such, these TC values must themselves be predicted at these points. This problem has led to the development of two different modelling frameworks that incorporate several geostatistical prediction components to allow TC data to be used to standardise the raw MC data. These modelling frameworks are now presented.

5.4 Modelling frameworks for predicting MC

Without any standardisation, the prediction of MC can be carried out directly using established geostatistical techniques. This straightforward approach is termed Model 1 in this project, and is represented schematically in Figure 5.5 (a). Model 1 can be thought of as representing the null approach and can be defined more formally as the prediction of values of MC at the q unsampled facility-months $z^*_{MC}((\mathbf{u}, t)_\beta)$, $\beta = 1,2,\ldots, q$ directly from the n MC data $z_{MC}((\mathbf{u}, t)_\alpha)$, $\alpha = 1,2,\ldots, n$.

Two further modelling frameworks were proposed that incorporate TC data, and these were termed Model 2 and Model 3. In Model 2 (Figure 5.5 (b)), MC data, $z_{MC}((\mathbf{u}, t)_\alpha)$, are divided by the corresponding TC data, $z_{TC}((\mathbf{u}, t)_\alpha)$, at each sampled facility-month, to create a new variable termed malaria proportion (MP), $z_{MP}((\mathbf{u}, t)_\alpha) = z_{MC}((\mathbf{u}, t)_\alpha) / z_{TC}((\mathbf{u}, t)_\alpha)$. Geostatistical prediction can then be implemented using $z_{MP}((\mathbf{u}, t)_\alpha)$ to obtain predictions $z^*_{MP}((\mathbf{u}, t)_\beta)$ at unsampled facility-months. The back-conversion of these predictions to MC requires corresponding predictions of TC. As such, the TC data $z_{TC}((\mathbf{u}, t)_\alpha)$ are used in a separate prediction exercise to predict $z^*_{TC}((\mathbf{u}, t)_\beta)$. MC can then be predicted as $z^*_{MC}((\mathbf{u}, t)_\beta) = z^*_{MP}((\mathbf{u}, t)_\beta) \times z^*_{TC}((\mathbf{u}, t)_\beta)$.

Model 3 (Figure 5.5 (c)) uses TC data in a different way from Model 2. Instead of using individual TC values as denominators for every facility-month, a single denominator is defined for each facility, referenced by the $k = 1,2,\ldots,K$ facility spatial locations ($\mathbf{u}_k$).

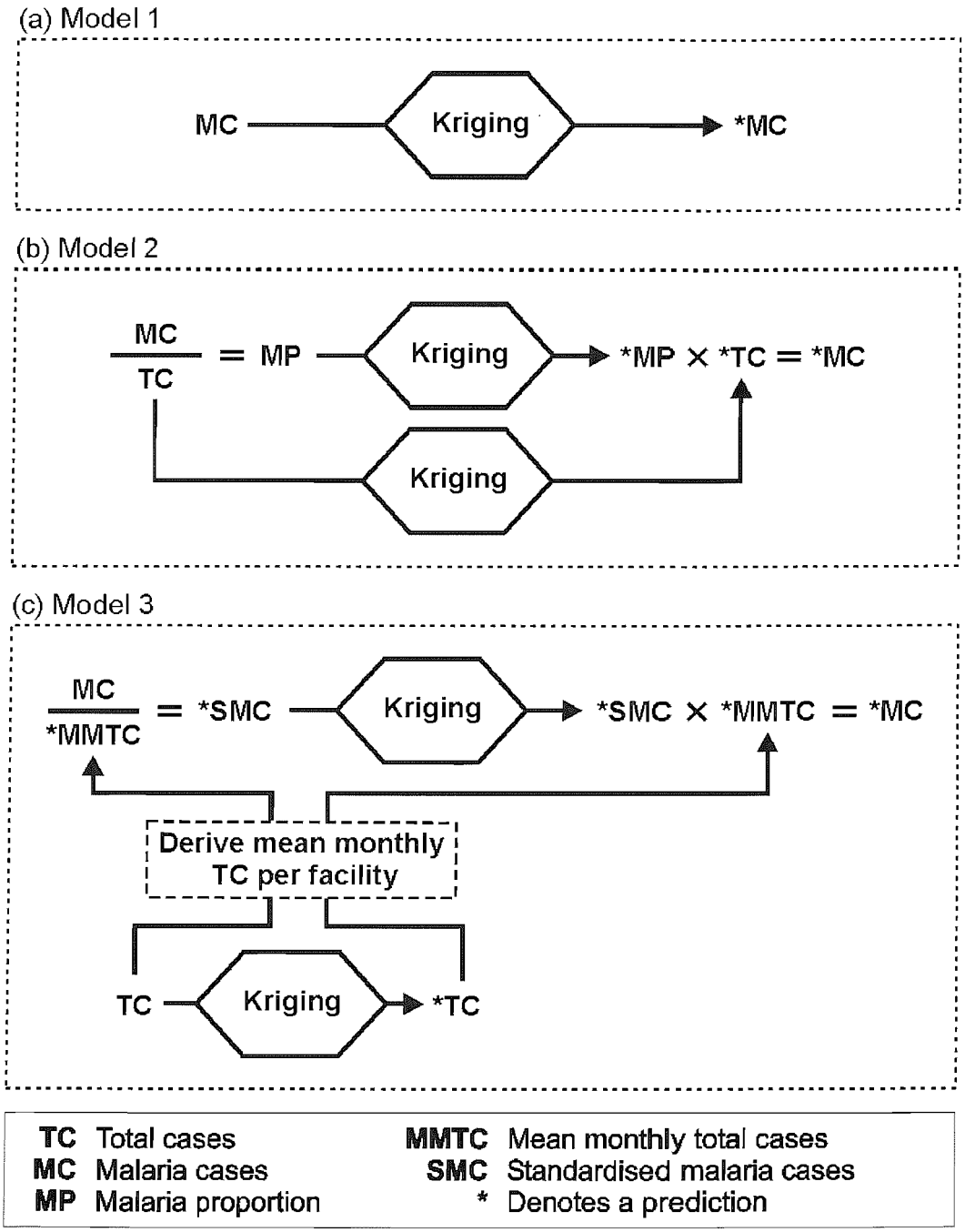


Figure 5.5 Schematic diagrams of three proposed modelling frameworks for predicting malaria cases at unsampled facility-months.

This value is the mean monthly total cases (MMTC) per facility, $z^*_{MMTC}(\mathbf{u}_k)$. Firstly, as with Model 2, geostatistical prediction is implemented with the TC data, $z_{TC}((\mathbf{u}, t)_\alpha)$, to predict $z^*_{TC}((\mathbf{u}, t)_\beta)$ at unsampled facility-months. 84 monthly TC values are now available for each facility, consisting of $d = 1,2,\ldots, D$ data and $p = 1,2,\ldots, P$ predictions, where $D + P = 84$. The MMTC denominator, $z^*_{MMTC}(\mathbf{u}_k)$, is then calculated for each facility as the temporal mean of these combined data and prediction sets:

$$z^*_{MMTC}(\mathbf{u}_k) = \frac{1}{D+P}\left[\sum_{d=1}^{D} z_{TC}((\mathbf{u}_k, t)_d) + \sum_{p=1}^{P} z^*_{TC}((\mathbf{u}_k, t)_p)\right] \quad (5.1)$$

Each MC datum, $z_{MC}((\mathbf{u}, t)_\alpha)$, is then divided by the MMTC value for the facility in question, $z^*_{MMTC}(\mathbf{u}_k)$ to create a new variable termed 'standardised malaria cases' (SMC):

$$z_{SMC}((\mathbf{u}, t)_\alpha) = \frac{z_{MC}((\mathbf{u}, t)_\alpha)}{z^*_{MMTC}(\mathbf{u}_k)} \quad (5.2)$$

where $\mathbf{u}_k$ has the same spatial coordinate as $(\mathbf{u}, t)_\alpha$. Geostatistical prediction can then be implemented using $z_{SMC}((\mathbf{u}, t)_\alpha)$ to obtain predictions, $z^*_{SMC}((\mathbf{u}, t)_\beta)$, at unsampled facility-months. The existing $z^*_{MMTC}(\mathbf{u}_k)$ values are then used to back-transform SMC predictions to MC, $z^*_{MC}((\mathbf{u}, t)_\beta) = z^*_{SMC}((\mathbf{u}, t)_\beta) \times z^*_{MMTC}(\mathbf{u}_k)$, where $\mathbf{u}_k$ has the same spatial coordinate as $(\mathbf{u}, t)_\beta$.

5.4.1 Model development and testing

Having presented three conceptual modelling frameworks for the prediction of MC, it was necessary to develop each framework into a functional approach for obtaining predictions of MC, and to compare each approach in terms of prediction accuracy. The set of three modelling frameworks consisted of four different prediction exercises, to predict MC, TC, MP, and SMC. The first task was to establish a geostatistical methodology for carrying out these predictions. Rather than simply adopt the most established methods such as OK, the approach taken in this project was to implement and develop less widely-used or novel methods that may be better suited to the

characteristics of the HMIS data set and to quantitatively compare the predictive performance of these different methods. This work is presented in Chapter 7. Having identified the most appropriate prediction method, the second task was to implement this approach within each modelling framework to produce MC predictions and to compare the accuracy of these predictions to identify the best-performing modelling framework. This work is presented in Chapter 8.

5.5 Chapter summary

This chapter has provided an exploration of the factors that determine the MC variable and has presented a simple conceptual model that states that MC is largely a function of catchment population size and the level of malaria morbidity within that population. Because of the environmental drivers, the level of malaria morbidity is likely to display substantial spatial dependence but this structure is likely to be confounded in the MC variable by non-spatial variability in the size of different catchment populations. As such, it may result in more accurate predictions of MC if the variable can be standardised to reduce the effect of this non-spatial variation. Two strategies have been presented to tackle this problem: the development of spatial models that allow catchment populations to be predicted; and the use of TC data on total monthly outpatient use at each facility. Models developed for the first strategy are presented in Chapter 6. To implement the second strategy, two modelling frameworks have been proposed that incorporate TC data in the prediction of unsampled MC values. Geostatistical prediction techniques within these modelling frameworks are developed and tested in Chapter 7, and the modelling frameworks are developed and tested in Chapter 8.

Chapter 6

Catchment Modelling

Chapter 6

6. Catchment Modelling

6.1 Introduction

The need for estimates of the size of health facility catchment populations that can be used as denominator values for the raw MC data has been presented in the preceding chapter. This requirement led to a series of modelling studies carried out as part of a wider project led by Dr. Abdisalan Noor at the Malaria Public Health & Epidemiology Group, Centre for Geographic Medicine in Nairobi, part of the KEMRI-University of Oxford-Wellcome Trust Collaborative Programme. This collaborative work is described in this chapter, with work led by the current author (Gething et al., 2004) presented in full and work led by Dr. Noor (Noor et al., 2006) described in summary.

The problem faced in this project was to develop ways of estimating facility catchment populations based only on a census-based GIS population map and data on the type and location of each facility. Under these circumstances the most straightforward and widely used approach is to define catchment boundaries based on Thiessen polygons. This approach relies on several implicit assumptions about the way care-seekers utilise different health facilities. In this project, a series of novel spatial modelling techniques were devised that allowed these assumptions to be tested using data from a patient-use study carried out in four Kenyan districts (Zurovac et al., 2002). This work is presented in the following section. A further limitation of many existing catchment modelling techniques is the representation of space using Euclidean distance. This modelling approach neglects the heterogeneity of the land surface and how this may affect the decisions made by care-seekers choosing between health facilities. Collaborative work led by Dr. Noor to develop GIS-based catchment models that incorporate more realistic

representations of space is described in the second main section of this chapter. The chapter concludes with a discussion of the success of this work in modelling catchment populations, and the implications for the overall project aim of defining treatment burdens for malaria.

6.2 Assessment of a simple Thiessen polygon model

For the majority of the government health facilities in Kenya, the only information available to assist in estimating catchment populations is their location and type (e.g. hospital, health centre, dispensary), as provided by the NHSD. Given this information, a simple and intuitive means of partitioning a population between a series of facilities is provided by Thiessen polygons. A Thiessen polygon (also called a Dirichlet tile) is defined in this case as the region that incorporates all points in space that are closer to a given facility than any other. The use of Thiessen polygons in this context is well established (Twigg, 1990; Zwarenstein et al., 1991; Albert et al., 2000; Noor et al., 2003) and is based on two key assumptions:

(1) that all patients choose to utilise the facility nearest to them, regardless of its type, and hence the spatial extent of a facility catchment is determined solely by the proximity of its neighbours; and

(2) that the proportion of care-seekers who utilise a given facility (the utilisation rate) is constant throughout a catchment, and does not decline, for example, with distance away from the facility.

Previous studies of patient behaviour have allowed inferences to be made about the validity of one or both of these assumptions in various settings and these are discussed in the following section. This study presents a series of new spatial analytical methods by which the validity of these assumptions can be tested directly and hence the suitability of a Thiessen polygon catchment model assessed explicitly. These methods were applied to paediatric outpatient origin data from a sample of 81 government health facilities in four districts of Kenya, and the observed patient-use patterns reported. The extent to which

the methods presented allow the validity of the Thiessen polygon assumptions to be assessed is then discussed.

6.2.1 Background

6.2.1.1 Patient choice

The actual partitioning of a population between two neighbouring facilities is determined by choices made by care-seeking members of that population. As discussed in Chapter 5, these choices can be based on a wide range of considerations including social, cultural, economic and behavioural factors, as well as the characteristics of the facilities in question (Stock, 1983; Müller et al., 1998; Onokerhoraye, 1999; Deressa et al., 2003). If the two facilities are perceived to be of equal standing by the population then it would be reasonable to expect care-seekers to base their choice of facility on the relative distance to each. In this idealised case, a theoretical catchment boundary would exist that is equidistant to both facilities. If one facility was perceived as a more attractive option, however, then one might expect care-seekers to be willing to travel relatively further to reach it than its less-favoured neighbour. In these circumstances, the location of the catchment boundary would be shifted towards the latter facility. Various studies into health facility utilisation patterns in developing countries have observed differences in the attraction or draw of different facility types. A study in rural Nigeria reported that the perceived lower quality of service available from dispensaries meant that they were less likely to attract patients over longer distances than were the higher-order facilities (Stock, 1983). Similar patterns were also noted in later studies in Nigeria (Onokerhoraye, 1999) and Papua New Guinea (Müller et al., 1998).

Few studies have attempted to test directly actual patient-use patterns in relation to the theoretical patterns defined by Thiessen polygons. A simple means of quantifying this pattern is to determine the proportion of people who have utilised their nearest facility. Previous analysis of the 81-facility patient origin data used in this study has shown that this proportion ranges from 56% to 83% over the four districts (Noor et al., 2003). An earlier study in rural South Africa stated that 81% of homesteads utilise their nearest facility (Tanser et al., 2001). These values suggest that, although Thiessen polygons may

provide a reasonable approximation of patient behaviour, there is a proportion of patients that base their choice of facility on factors other than distance. Tanser et al. (2001) also compared actual to predicted (Thiessen) catchments and concluded that there was overall agreement between predicted and actual catchments but that large inter-catchment variation existed.

6.2.1.2 Utilisation Rate

The simple allocation of a population into a series of contiguous facility catchments such as Thiessen polygons assumes a uniform utilisation rate throughout that population. This implies that, within a catchment, a patient's likelihood of visiting the facility is not affected by their distance from it. The concept of distance as a primary influence on health facility utilisation is well established (Shannon et al., 1969, 1973; Kohli et al., 1995). Previous studies have investigated the relationship between utilisation rate and distance in a wide range of settings and a variety of different trends have been observed. Several studies in rural areas of Ethiopia, for example, have reported distance effects on care-seeking behaviour with steep distance-decay gradients in utilisation rate and under-utilisation of more rural health services (Kloos, 1990; Deressa et al., 2003). The studies by Stock (1983) and Tanser et al. (2001) both describe an exponential decay in utilisation rate with distance and this model has commonly been presented as a reasonable approximation of the utilisation-distance relationship in both developed and developing world settings (Morrill and Earickson, 1968; Ingram et al., 1978). A study in rural Papua New Guinea reported that although utilisation rate showed a general decline with distance, this decline was not evident until some distance away from facilities and a Gaussian curve was therefore proposed as being a more representative model (Muller et al., 2006). A study in the Kilifi District of Kenya found a decrease in admission rates to the district hospital with distance such that the rate in populations located more than 25 km from the hospital was one fifth of that within 5 km (Schellenberg et al., 1998). In contrast, other studies have found distance to have no systematic effect on utilisation rate even in rural settings (Girt, 1973; Slack et al., 2002).

6.2.2 Data and study area

This study was based on data acquired by the Government of Kenya (Ministry of Health – Division of Malaria Control) and the Kenya Medical Research Institute-Wellcome Trust Collaborative Programme (Zurovac et al., 2002). The four Kenyan districts of Bondo, Greater Kisii (now composed of Kisii Central and Gucha district), Kwale and Makueni (Figure 6.1) were chosen as encompassing a broad range of the most prevalent environmental, demographic, and socio-economic conditions found across Kenya. Greater Kisii and Bondo exhibit relatively evenly distributed and high density population whereas Makueni and Kwale include areas of very low population density. This difference is reflected in the density of health facilities within the districts. The districts are described in more detail in Zurovac et al. (2002), Noor et al. (2003) and Amin et al. (2003).

A total of 81 government facilities consisting of hospitals, health centres and dispensaries were sampled from the four districts during 2001-2002. Each facility was sampled over two days during which time the place of origin was determined for all children who were attending with a fever. The smallest Kenyan census unit is the enumeration area (EA), normally consisting of not more than 100 households, and these were the spatial units by which each child was located. EA population and out-patient data were compiled into a GIS polygon layer in ArcView 3.2 (ESRI Inc., USA) along with a point coverage of all GoK health facilities. For a full description of out-patient and population data acquisition and digitisation see Noor et al. (2003).

6.2.3 Methodology

6.2.3.1 Overview of approach

The various studies described above reported differing draws from different facility types, significant proportions of patients attending facilities other than their nearest, and decay in utilisation rate with distance. Whilst these findings enable an assessment of the suitability of the Thiessen polygon model, they are less able to suggest how such a model could be modified to represent more accurately the patient behaviour observed.

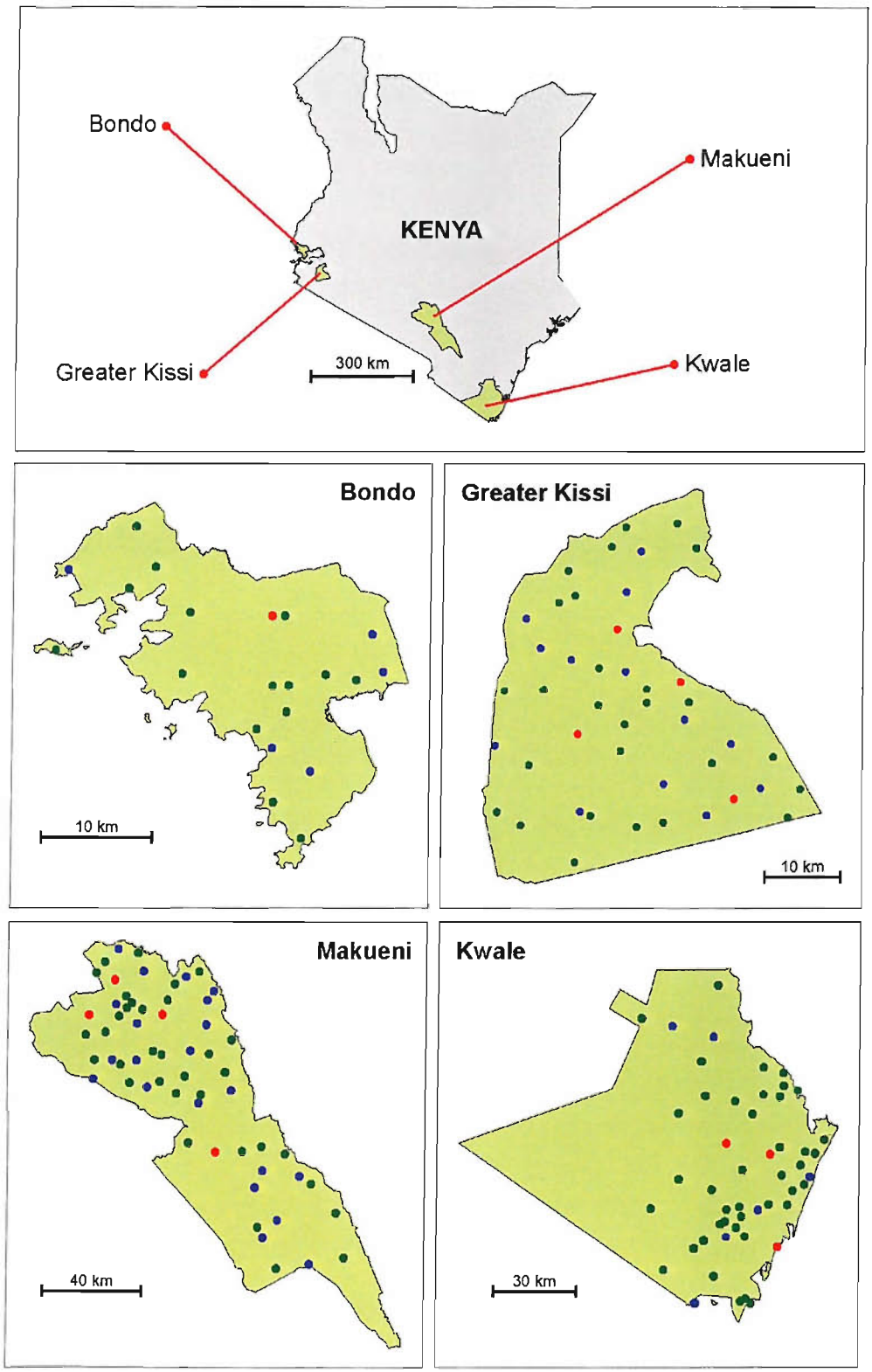


Figure 6.1 Location of four study districts in Kenya (top) and maps of each district showing location of all government hospitals (red dots), health centres (blue dots) , and dispensaries (green dots). North is to the top in all maps.

In stage 1 of this study, a GIS was used to predict the location of the catchment boundary along a direct transect between each pair of neighbouring facilities (i.e. along the imaginary straight line connecting the two facilities) in the sample set based on patient choice patterns (Figure 6.2). Once the location of the catchment boundary was established, this was compared to the location predicted by a Thiessen polygon.

When considering a distance decay effect in utilisation rate it may be difficult to disentangle the influence of neighbouring facilities, especially where they are in relatively close proximity. This can lead to the incorrect conclusion that distance limits access to a facility when, in reality, patients at the periphery of a catchment are simply choosing to utilise a neighbouring facility. These effects can be disentangled, however, if the pattern of patient choice between the two facilities is analysed prior to assessing the utilisation rate gradient, and this is the approach taken in Stage 2 of this study. If a clear patient choice boundary can be identified then it is reasonable to interpret any reduction in utilisation rate within this boundary as being primarily a distance effect. The approach taken was to use a Thiessen polygon to define the boundary of each catchment, but to limit analysis of utilisation rate to a smaller area within this catchment by excluding a buffered area around the periphery. A suitable width for these buffers that could be considered sufficient to remove the effect of neighbouring facilities was determined based on the spatial patterns of patient choice found in Stage 1.

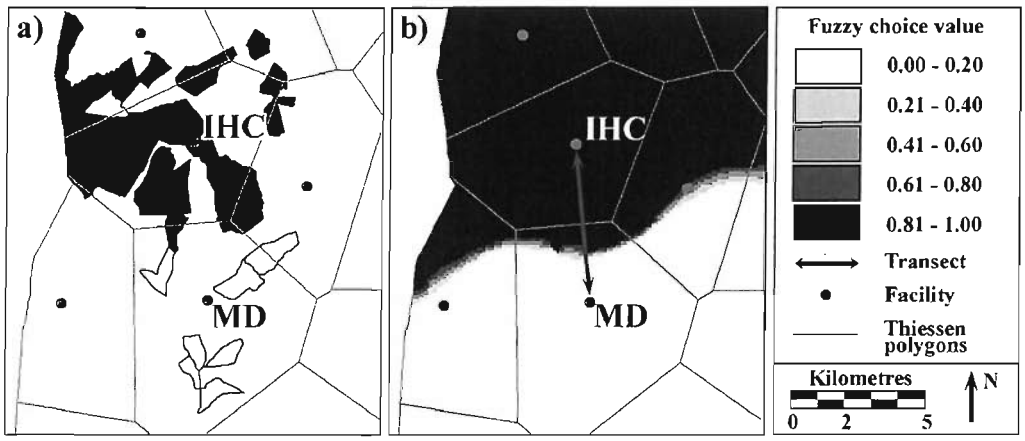


Figure 6.2 Creation of a fuzzy choice surface, as defined in the text. This example shows the case of Iyabe health centre (IHC) and Misesi dispensary (MD) in the Greater Kisii district. All enumeration areas contributing one or more patients to either facility were allocated a fuzzy choice value corresponding to the relative proportion attending Iyabe health centre (a). The polygon coverage was then rasterised into a 100 m grid and interpolated using an inverse distance weighting algorithm to predict a choice surface (b). Thiessen polygon boundaries are also shown for reference.

6.2.3.2 Stage 1: Analysis of spatial patterns of patient choice

A total of 174 GoK facilities were located in the four districts. Thiessen polygons were created around all these facilities and all cases were identified in which two of the 81 sampled facilities were immediately adjacent (i.e. they shared a Thiessen boundary). A total of 78 such pairs were identified across the four districts. Each pair was considered in turn and, for each, analysis was performed along the transect between the two facilities. A *fuzzy choice* value was assigned to every EA that contributed one or more patients to either facility in the pair. This value was simply the relative proportion of patients attending each facility from a given EA. The two facilities in each pair were labelled A and B such that values ranged from one (all patients went to A) to zero (all patients went to B). Facilities were assigned as A or B in a consistent manner depending on the type of facilities in question. This meant that each pair fell into one of five transect classes: health centre-to-dispensary (HC-D); dispensary-to-hospital (D-H); health centre-to-hospital (HC-H); health centre to health centre (HC-HC); or dispensary-to-dispensary (D-D). The opposite relationships (i.e. D-HC, H-D, H-HC) did not need to be considered separately as they were simply the inverse of those considered. For pairs of matching type, facilities were assigned as A or B arbitrarily. Hospital-to-hospital transects were not considered as hospitals did not neighbour one another. The EA fuzzy choice values were assigned to the EA polygon coverage (Figure 6.2 (a)) and these vector layers were converted into 100 m by 100 m raster grids and interpolated using an inverse-distance weighting algorithm. The result was a *fuzzy choice surface* (Figure 6.2 (b)) which represented a continuous prediction of patient choice behaviour between the two facilities in question.

For each facility pair, the fuzzy choice surface was analysed along the transect between the two facilities in question. Each transect was divided into 100 equally spaced points and the fuzzy choice value recorded at each point. This process was implemented using the ArcView *X-Section Utility v1.0* extension. The catchment 'choice boundary' was taken to be located at the point where the fuzzy choice value was equal to 0.5. For each of the five transect classes an 'average' transect was created by calculating the mean fuzzy choice value over all such transects for each of the 100 divisions. Relative distances along the transect were considered because the split of patients between neighbouring facilities was of interest regardless of the absolute distance between them.

In addition to creating a mean transect for each transect class, the relative location of the choice boundary was recorded for each individual transect. A Thiessen polygon boundary is located at the exact mid-point of a given transect (i.e. at 50%, since transects ran from zero to 100%). Actual boundary locations less than 50% were closer to facility A, while those greater than 50% were closer to facility B. Overall mean and district mean transect location was calculated for each transect class. Single sample *t*-tests were carried out on the overall mean location for each class. For the D-D and HC-HC classes a two-tailed test was applied to test for a significant difference from the Thiessen boundary (i.e. from a mean value of 50%). For the remaining three transect classes a one-tailed test was used. Hospitals were the highest-order facility followed by health centres and then dispensaries. The expectation was that any deviation from the Thiessen boundary is due to patients choosing to make a longer journey to reach a higher-order facility, resulting in a displacement from the Thiessen boundary towards the lower-order facility.

6.2.3.3 Stage 2: Analysis of spatial patterns of utilisation rate

To isolate the effect of distance on utilisation rate for each facility it was necessary to define each catchment such that the influence of neighbouring facilities could be considered minor. This was achieved by shrinking the Thiessen polygon boundaries of each catchment such that their radii were reduced by approximately 25%. This value exceeds the largest mean deviation from a Thiessen boundary position found in the analysis of patient choice in Stage 1 (see Table 6.1). This strategy was implemented by creating an exclusion buffer, the width of which was calculated as a function of the area of each polygon. If polygons can be assumed to be approximately square then the width W of buffer required to achieve a reduction in radius of 25% can be defined in terms of the polygon area A as:

$$W = 0.25\frac{\sqrt{A}}{2} \tag{6.1}$$

and buffers were created at this width for each catchment polygon (Figure 6.3).

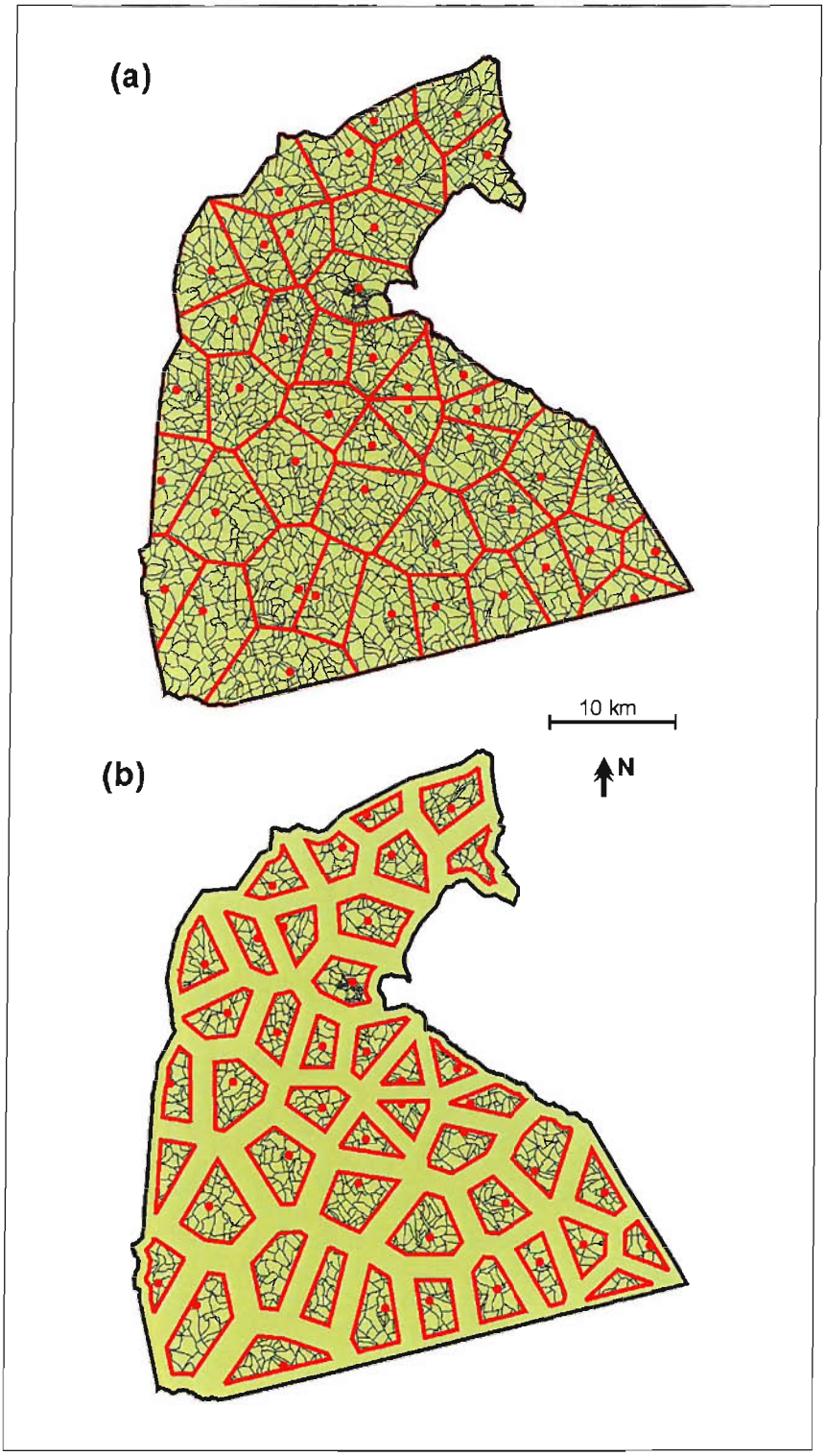


Figure 6.3 Usc of exclusion buffers for assessing within-catchment utilisation rate, the example of Greater Kisii district. Map (a) shows enumeration areas (fine black lines), government facilitics (red dots), and facility catchment boundaries based on Thiessen polygons (red lines). The bottom map (b) shows the shrunken catchments used for analysis of utilisation ratc following application of exclusion buffers, as described in the text.

For each sampled facility, a utilisation rate was calculated within each EA contributing one or more patients (i.e. excluding those contributing zero patients). This rate was calculated simply by dividing the number of patients from a given EA who attended the facility in question by the total population of that EA. Such an approach to defining a utilisation rate is sub-optimal since the appropriate denominator is the total number of children in each EA who suffered a fever during the sample period. Facility-based surveys cannot capture this population-based phenomenon, however, and the use of total EA population as a denominator was the only viable option given the data available. It was reasonable to assume, however, that there were not substantial systematic differences in the incidence of fever over the small spatial regions of interest. To obtain utilisation rate values that were more comparable between the set of EAs considered for each facility, each rate value was standardised into a relative utilisation rate (RUR) by dividing it by the largest value in the set. These RUR values were linked back to the EA polygon coverage and rasterised into a 100 m by 100 m grid. The study catchment for each facility was then delineated using the exclusion buffers.

For each of the 81 rasterised study catchments, the RUR value of every grid cell was output along with its six-digit latitude and longitude. The straight line distance between each facility and the centroid of each non-zero RUR cell in its study catchment was calculated. RUR values were then grouped by distance from facility and a mean value was calculated for every successive 100 m. An overall mean RUR plot was created along with one for each district. These plots illustrate the influence of distance from facility on RUR. In contrast to the analysis of patient choice, utilisation rate was considered with reference to absolute distance.

6.2.4 Results

Mean fuzzy choice transects are shown for the three classes of differing facility type that were present: HC-D, D-H, and HC-H (Figure 6.4). In each case, the position of the 0.5 fuzzy value, taken to represent the choice boundary, was located nearer the lower-order facility. Table 6.1 lists the overall and district mean boundary locations for all five transect classes. The overall mean boundary locations were 51% for the mean D-D transect, 50% for HC-HC, 59% for HC-D, 40% for D-H and 39% for HC-H. Two-tailed

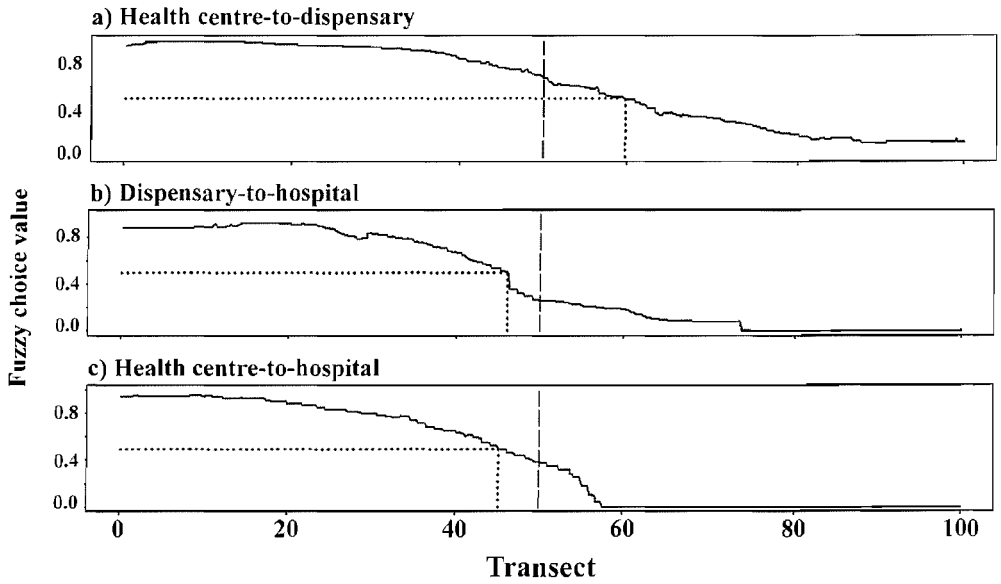


Figure 6.4 Mean fuzzy choice transects for all neighbouring facility pairs of class health centre-to-dispensary (a), dispensary-to-hospital (b) and health centre-to-hospital (c) illustrating the relative draw of different facility types. The location of the theoretical Thiessen boundary is marked at the mid-point (dashed line) along with the location of the observed 0.5 fuzzy choice value (dotted line).

single sample *t*-tests for the D-D and HC-HC classes both revealed no significant difference from the Thiessen boundary location of 50% (P=0.77 and P=0.98 respectively). One-tailed single sample *t*-tests for the remaining three classes revealed that boundary locations were significantly nearer the lower-order facility in each case. There was substantial variation between districts with, for example, the mean HC-D boundary location ranging from 62% in Kwale to 54% in Makueni, and the mean D-H boundary location ranging from 37% in Bondo to 47% in Makueni. Caution should be

Table 6.1 Mean position of catchment boundaries for each transect class (%). Values of less than 50% are closer to the first facility in the pair while values greater than fifty are closer to the second. A theoretical Thiessen boundary is equidistant to both facilities and would therefore be located at exactly 50%.

Transect class	District means				Overall means
	Bondo	Grt. Kisii	Kwale	Makueni	
Dispensary-to-dispensary	57.30	49.40	42.40		**51.41 (P = 0.7701)[a]**
Health centre-to-health centre		52.87	20.20	56.60	**49.80 (P = 0.9798)[a]**
Health centre-to-dispensary	61.31	55.25	62.17	54.40	**58.50 (P = 0.0077)[b]**
Dispensary-to-hospital	37.30	39.05	38.60	46.70	**39.88 (P = 0.0041)[b]**
Health centre-to-hospital		32.75		50.40	**38.63 (P = 0.0656)[b]**

[a] two-tailed single sample *t*-test for significant difference from 50%
[b] one-tailed single sample *t*-test for significant shift from 50% towards lower-order facility

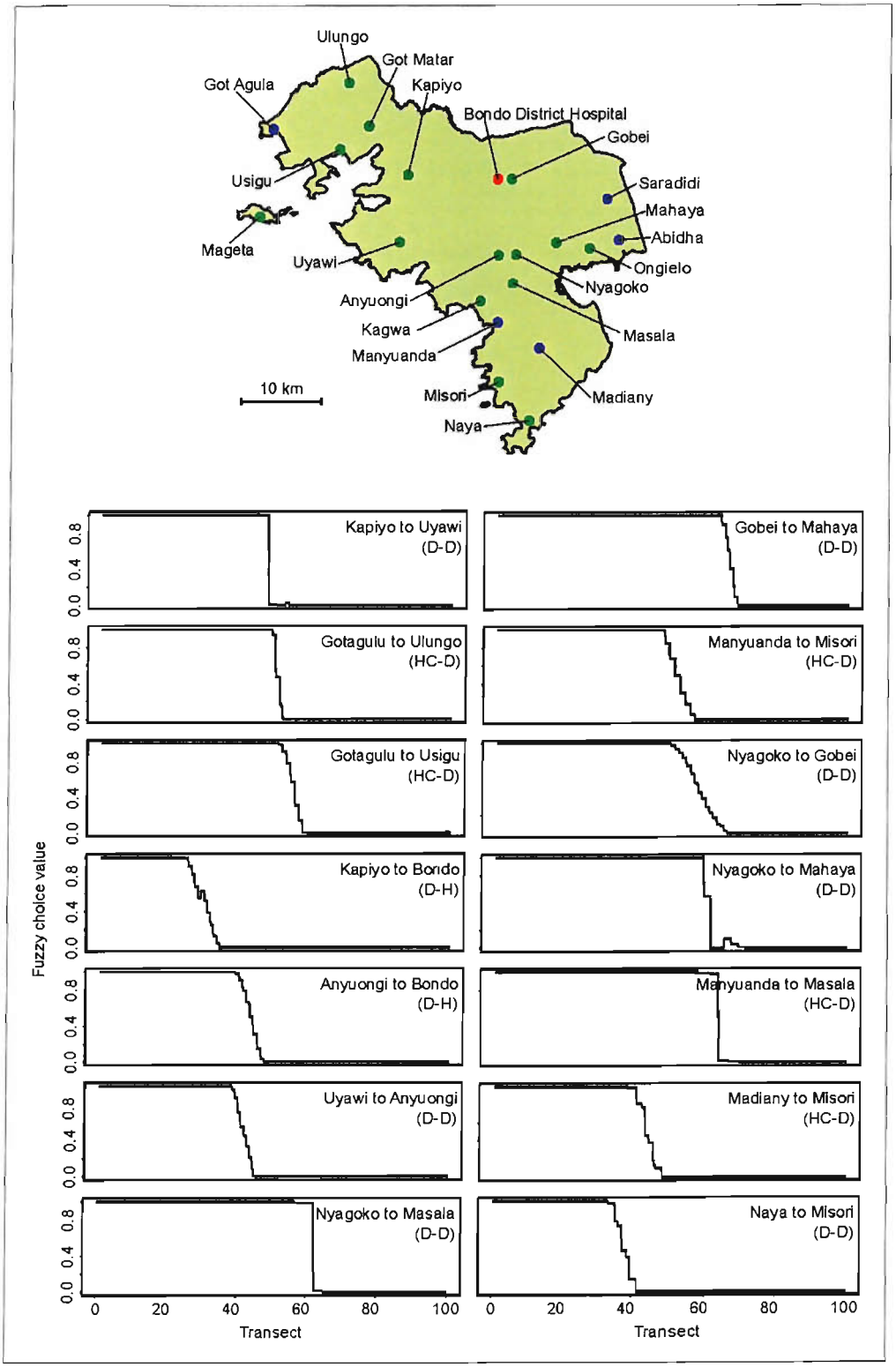


Figure 6.5 Sample of individual fuzzy choice transects from Bondo district. Map shows location and names of government hospitals (red dots), health centres (blue dots) and dispensaries (green dots) in Bondo district. Labels on each transect detail the two facilities in question. Labels in parentheses refer to the type of facilities involved: H = hospital, HC = health centre, D = dispensary.

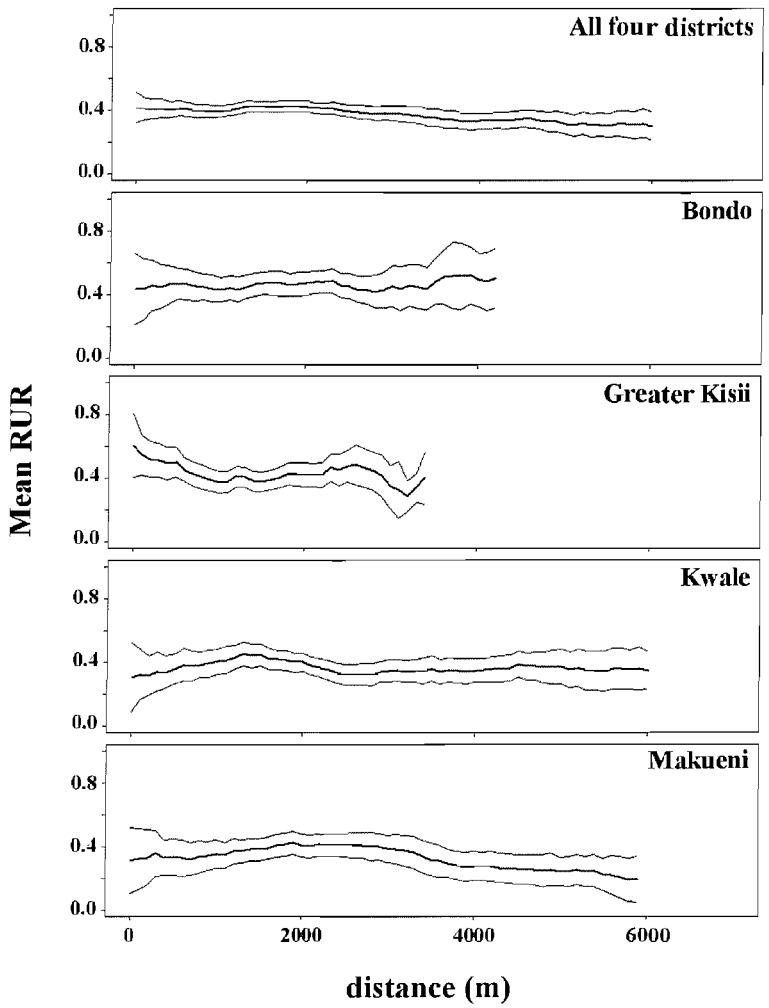


Figure 6.6 Mean overall and district relative utilisation rate (RUR) plots. Thick lines show the mean RUR value of all sampled 100 m by 100 m grid cells within every study catchment at each distance. 95% confidence intervals are also shown (fine lines).

exercised when interpreting these results, however, because the number of pairs for a given transect class in a single district was often small meaning that sampling variation was large. The boundary position for the HC-HC class in Kwale, for example, was based on a single pair. Whilst it is likely that disparities in facility and population density between the four districts lead to differences in the spatial patterns of patient choice, the sparsity of data at the district level mean that the results are best interpreted with data from the four districts combined. Figure 6.5 shows a selection of choice transects from individual facility pairs in Bondo district. Whilst the mean transects (Figure 6.4) illustrate the overall draw of different facility types, these individual plots illustrate the transition in patient choice between two specific facilities.

Relative utilisation plots are shown in Figure 6.6. These include a mean plot for each district as well as an overall mean. Mean plots are accompanied by 95% confidence intervals. Each of the district plots extends to a different length which corresponds to the most distant non-zero pixels found in any of the study catchments in each district. The districts of Bondo and Greater Kisii are characterised by a relatively dense network of facilities corresponding to a higher population density. Catchments are, therefore, smaller than some of those found in the more rural Kwale and Makueni districts (Figure 6.1). For Bondo, RUR fluctuates but exhibits no systematic trend with distance. For Greater Kisii RUR decreases with distance. For both Kwale and Makueni RUR increases up to around 2 km, and then levels off (Kwale) or steadily declines (Makueni). Overall there exists a slight, but steady decrease in RUR with distance up to 6 km.

6.2.5 Discussion

6.2.5.1 Patient choice

The construction of fuzzy patient choice surfaces is presented as a robust means of assessing patient behaviour for two neighbouring health facilities and identifying the location and nature of the choice boundary between them. This method represents the conversion of two separate facility-based variables (attendance per EA) into a single facility-pair-based variable (fuzzy choice) that describes the spatial partitioning of patients between the two facilities in question. By analysing patient choice along a transect between two neighbouring facilities, the influence of other facilities is minimized. The mean fuzzy choice transects for each transect class (Figure 6.4) suggest a smooth gradient of choice between the two facility types in question. This is not, however, representative of the shape of most of the 78 individual choice transects (see examples in Figure 6.5). These exhibited a much sharper transition from high to low choice values indicating a crisper boundary. Although this characteristic of the individual plots is smoothed in the averaging process, the mean transects are useful for illustrating the relative drawing power of the different facility types as a whole, especially with reference to the Thiessen boundary. The calculation of mean boundary locations, along with the use of appropriate significance tests (Table 6.1), provides a

means of comparing directly the observed behaviour patterns to those assumed in a Thiessen polygon catchment model.

6.2.5.2 Utilisation rate

The method presented allows the effect of distance on utilisation rate to be studied in isolation from the possible influence of surrounding facilities. Although one consequence of the use of exclusion buffers is that the maximum distance over which this relationship can be studied for any given set of sample facilities is inevitably reduced, it provides a means of elucidating the influence of distance alone. The degree of confidence associated with each of the mean plots in Figure 6.6 follows a similar pattern – wide confidence intervals at small distances which narrow through medium distances before widening once more at the larger distances. This consistent pattern can be explained largely by changes in sample size. Only those EAs that contributed patients to the sample could be included in the analysis and this represented a relatively sparse sample (between 13% and 28% of EAs across the four districts). The successive 100 m distance bands (over which RUR values were averaged) can be considered as a series of concentric bands of equal width and, as such, their area increases linearly with distance from the facility. Distance bands close to the facility are, therefore, smaller and less likely to contain as many non-zero RUR pixels as those further away, with a corresponding effect on sample size. When considering the largest distances in each district the sample size is likely to be small since there are few examples of catchments that extend to this distance.

Individual district plots display considerable variation, particularly at distance up to around 1500m. It is likely that much of this inter-district variation can be attributed to sampling variation due to the relatively sparse number of contributing EAs at these short distances in each district. When the individual district plots were combined, a general decrease in RUR is evident over the distances studied (up to around 6000m). This suggests that, for the data set studied, the assumption of uniform within-catchment utilisation rate is inappropriate. The observed decline with distance is consistent with most other low-income country studies (Stock, 1983; Kloos, 1990; Schellenberg et al., 1998; Tanser et al., 2001; Deressa et al., 2003; Muller et al., 2006) although the

observed decline is far less pronounced than many of those reported. A reasonable explanation for this difference is that the influence of neighbouring facilities is often manifest as a reduction in RUR towards the periphery of a catchment and this effect has not been removed adequately in many studies leading to the over-reporting of decline in RUR with distance.

Euclidean distance, simply the straight-line distance between two points, was the distance metric used in this study. When considering distances on the ground, an alternative metric is the financial or time cost of making the journey between two points. Various studies have investigated the correspondence between Euclidean distance and journey cost (Perry and Gesler, 2000; Costa et al., 2003) and the extent to which the two concur is dependent on factors such as the density and quality of transport networks, the nature of topography, and features such as rivers and swamps. Whilst Thiessen polygons are defined in terms of Euclidean distance, their validation using georeferenced patient-use data does not explicitly use any distance metric as it relies solely on the positions of patients in Cartesian space and their choice of health facility in relation to the Thiessen polygon boundaries. A consideration of the discrepancy between journey cost and Euclidean distance becomes important, however, when attempting to explain the reasons for the observed shifts in boundary positions towards lower-order facilities. A situation may exist, for example, whereby the quality of transport networks around hospitals allows more efficient journeys than that around health centres and dispensaries. In this case, if care-seekers based their choice of facility entirely on journey cost then one would expect a relative increase in the spatial extent of hospital catchments. This explanation has some grounding in that hospitals are generally located in urban areas, where one might expect transport networks to be most efficient. However, given that the majority of patients made the journey to seek treatment on foot, and that populated non-urban areas generally have comprehensive road, track, and footpath networks, it is unlikely that this effect is responsible for all of the observed shift in boundary location. These issues do, however, underline the importance of considering journey cost when attempting to predict patient behaviour.

6.2.6 Conclusions

The methods presented in this study allow the two key assumptions inherent in a Thiessen polygon approach to defining catchment boundaries to be tested directly using patient-use data and, where these assumptions are found to be invalid, provide guidance as to how the approach can be refined to better represent the patient behaviour observed. For the four Kenyan districts studied, the two assumptions were found to be invalid. In the 78 cases of neighbouring facilities considered, mean boundary locations were found to be significantly closer to lower-order facilities than predicted by Thiessen polygons. This implies that the relative draw of facilities of different types is different and as such patients are willing to travel some distance further to reach a higher-order facility than a neighbouring lower-order facility. Analysis of mean within-catchment utilisation rate revealed that, for distances of up to six km from a facility, a steady decline in utilisation rate with distance was present. This implies that it is sub-optimal to model utilisation rate as uniform within a catchment.

6.3 Incorporating the effects of journey-time

In the previous section it was established that the use of Thiessen polygons to define facility catchment boundaries is likely to be inappropriate for catchments across Kenya. A further limitation of this and other approaches is the use of straight-line Euclidean measures to represent the distance between care-seekers and facilities. Factors such as topography, the presence of natural or human-made barriers, and the nature of the transport network, mean that the effort, time, and expenditure required to reach a facility from a given location is not necessarily well represented by Euclidean distance. Rather, a metric such as journey-time provides a more useful way of defining access to facilities.

One component of this project has been to contribute to a collaborative study led by Dr. Noor and the KEMRI-University of Oxford-Wellcome Trust Collaborative Programme team which addressed the problem of incorporating a journey-time metric in catchment boundary models. This work followed directly from the previous Thiessen polygon study and the findings played an important role in shaping the development of the modelling strategy used ultimately in this project to predict MC. As such, this work is

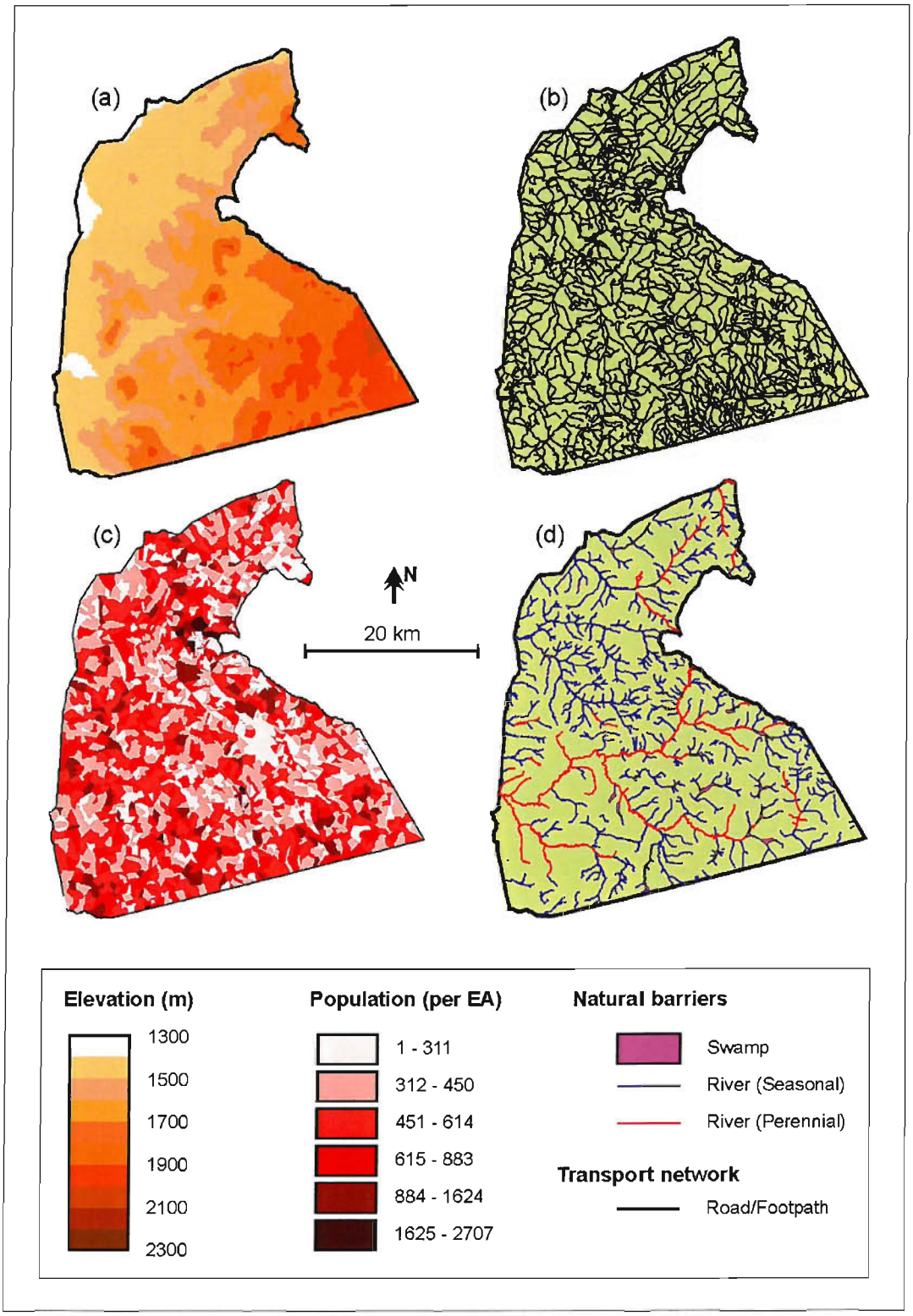


Figure 6.7 Examples of GIS data used for input into a journey-time based catchment model. Data shown are for Greater Kisii district and consist of (a) a raster coverage digital elevation model (DEM), (b) a vector coverage of the transport network including roads, tracks, and footpaths, (c) a vector coverage of population in each enumeration area (EA) derived from the 1999 national census, and (d) a vector coverage of natural barriers including, in this case, rivers and swamps.

briefly summarised in this section, and the reader is pointed to the original accounts in Noor (2005) and Noor et al. (2006).

6.3.1 GIS and patient-use data

The study was based in the same four Kenyan districts as the Thiessen-polygon study, namely Bondo, Greater Kisii, Kwale and Makueni. For each district, data were collected by the KEMRI team from a range of sources to create a series of GIS layers for later model development. In addition to the data on facility type and location contained within the NHSD, population data were obtained from the 1999 census at EA level and a digital national road network map was obtained and augmented for the four districts by digitizing footpath networks obtained from paper maps. In addition, a digital elevation model (DEM) was derived from contour maps, and vector data on natural barriers such as rivers and coastlines were obtained, along with data on human-made barriers such as national parks and other sanctuaries (Figure 6.7).

Patient-use data were collected in the four districts as part of a household survey conducted by KEMRI in 2001 (Amin et al., 2003; Guyatt et al., 2004). A stratified random sample of approximately 230 EAs was included, covering between 25,040 and 25,928 people in each district. During this survey, each sampled homestead was georeferenced in the field using a geographic positioning system (GPS) and, where a child had suffered a fever in the previous fourteen days, their care-seeking behaviour was documented, including which formal health facility they had attended, if any. These data were incorporated in the GIS as a point coverage of homesteads accompanied by data on their choice of health facility.

6.3.2 Model development

A facility catchment model was developed that accounted for the journey-time required by care-seekers to reach different facilities. Pedestrian journey-time was used because this was the mode of transport used by the overwhelming majority of patients in the household survey. A raster cost-surface was developed that estimated the journey time in minutes from each grid cell to the nearest health facility, taking into account the layout

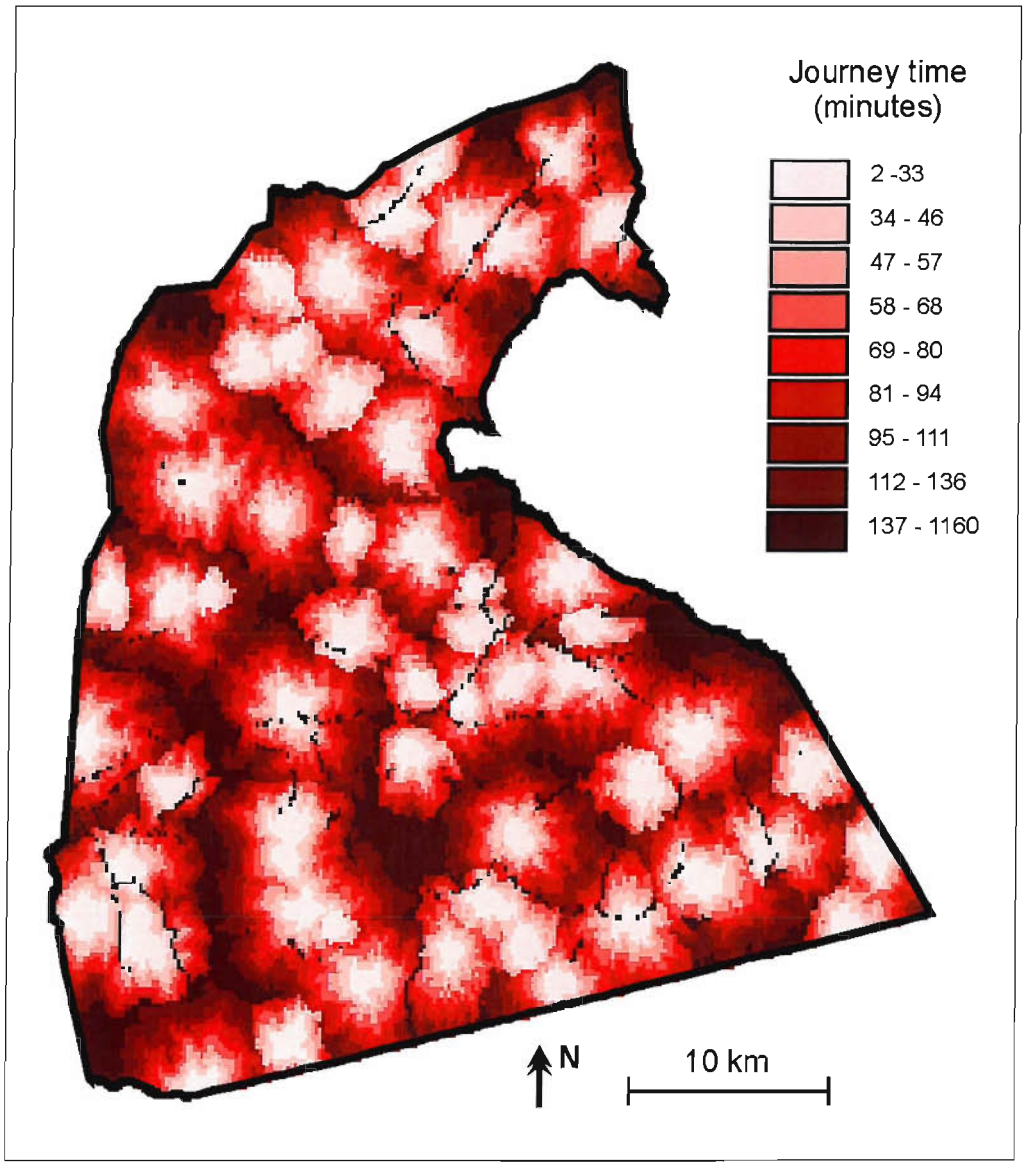
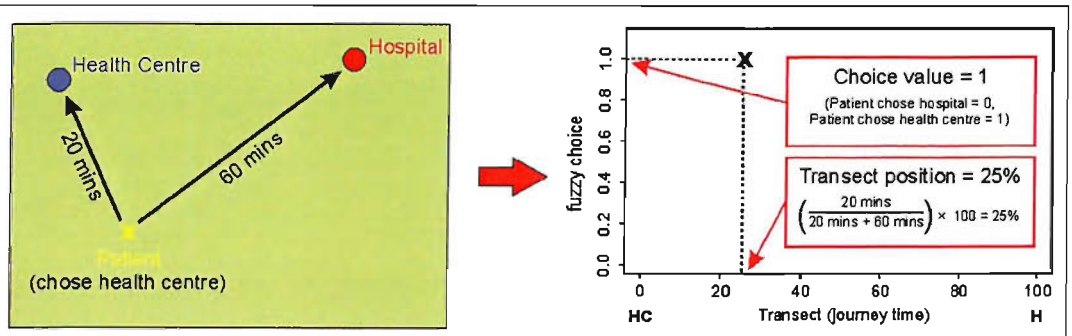
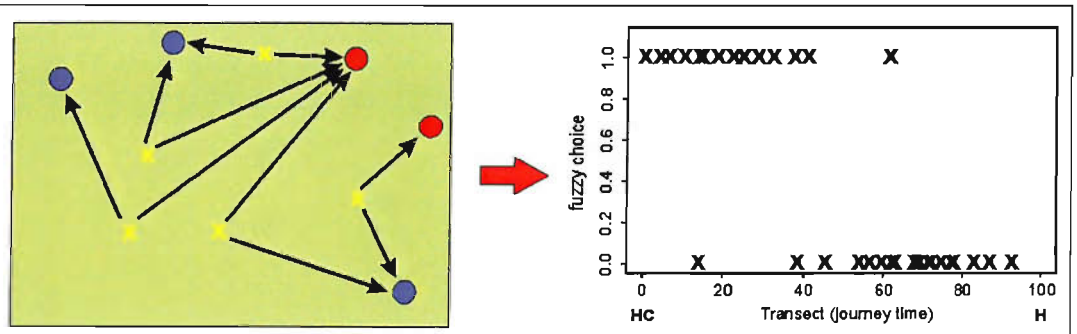


Figure 6.8 Example cost-surface for Greater Kisii district showing he estimated pedestrian journey-time in minutes from each 100m by 100m grid cell to the nearest government health facility. The journey-time algorithm took into account factors such as the road and footpath network, gradient, and natural barriers.

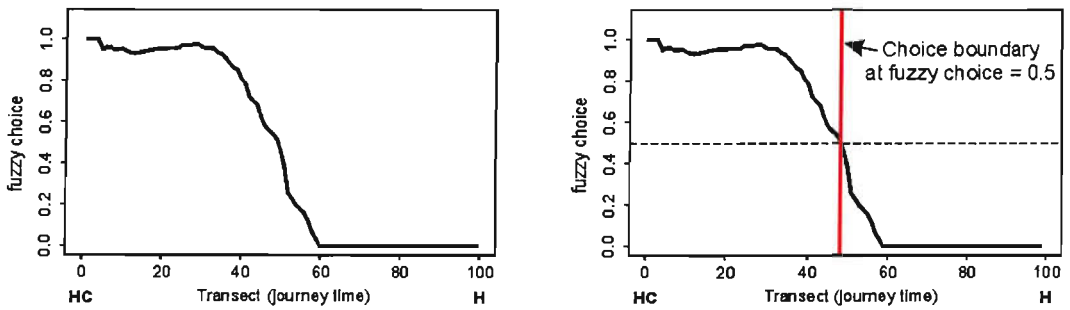
of the road and footpath network, the gradient at each point (including the direction so that uphill and downhill effects were distinguished), and the presence of impassable features such as rivers, coastline, and enclosed areas such as National Parks and other sanctuaries (Figure 6.8). This cost-surface was generated using a novel region-growing algorithm coded in ANSI C. The raster cost-surface was then used to predict catchment boundaries that were the journey-time equivalent of Thiessen polygons in that each



Step 1. For each patient, journey time in minutes was calculated to nearest health centre and hospital (left diagram) using the journey-time algorithm. The patient's choice of facility was marked on a relative journey-time transect (right diagram) with the *x*-axis position representing their relative location (in terms of journey-time) between the two facilities, and the *y*-axis value representing their choice between the two facility types.



Step 2. Step 1 was repeated for all patients who used either a health centre or a hospital in all four districts. In each case, the patient was added to the HC-H transect according to their relative location (in terms of journey-time) between the two, and their choice of facility was represented as either a 1 (went to the health centre) or 0 (went to the hospital).



Step 3. When all patients that had used either a hospital or a health centre had been marked on the HC-H transect, the resulting plot was smoothed using a moving window to give a smooth plot showing the mean transition in choice between the two facility types with relative journey-time between them (left plot). This transition was summarised by the location of the 0.5 fuzzy choice value, deemed the choice boundary (right plot).

Figure 6.9 Schematic diagram outlining the procedure by which the spatial patterns of patient choice were assessed using journey-time transects. In this example, patients' choices between health centres and hospitals is considered using the health centre – to – hospital (HC-H) transect.

location was predicted as utilising the health facility that was located the shortest journey-time away.

Mirroring the Thiessen polygon study, a method was devised for assessing the relative draw of different facility types. As before, mean choice transects were developed for the HC-D, D-H and HC-H facility classes. This time, however, the concept of a straight-line transect between neighbouring facilities was not useful, since no straightforward journey-time equivalent exists. Instead, an alternative approach was developed that placed each sampled homestead on a 'journey-time transect' that represented the relative journey-time between it and facilities of each different type. In this approach, a homestead that was located 20 minutes from the nearest health centre, and 60 minutes from the nearest hospital, for example, would be placed 25% of the way along the HC-H transect (20/(20+60) = 0.25). With all sampled homesteads positioned along the various mean transects in this way, the proportion utilising each facility type was calculated at 100 points along each transect. As before, the point along each transect at which patients were found to be equally likely to utilise each facility was deemed the 'choice boundary'. This process is illustrated schematically in Figure 6.9. Deviations in these mean choice boundary locations from the centre of each transect indicated that patients were willing to make longer journeys in order to reach certain facility types than others. The results of these journey-time transects were then incorporated in a further catchment model. Rather than defining catchment boundaries based on nearest facility journey-times, the differential draw of different facility types was incorporated and the equidistant boundaries were adjusted accordingly. The three catchment models described (the Thiessen polygon model, the journey-time model, and the journey-time model adjusted for patient choice) were tested empirically by comparing the proportion of homesteads whose choice of facility was correctly predicted by each model.

6.3.3 Key findings

Of the various findings of this study, the most important in the context of the current project is the comparison of the different catchment modelling approaches. The basic Thiessen polygon model was found to predict the correct choice of facility for 72% of homesteads. Replacing the Euclidean distance metric with journey-time resulted in 74%

of homesteads being correctly modelled, and when this model was adjusted to incorporate the effects of patient choice, this value increased to 84%.

6.4 Implications of catchment modelling studies

The two studies presented in the chapter resulted in a series of findings that are important in their own right. In the context of the current project, however, these studies were required as ways of developing catchment models that could provide estimates of catchment populations for health facilities across Kenya that could, in turn, be used to standardise the raw MC data in the HMIS. In this context, the most important finding of these studies is that the most basic catchment models are based on unfounded assumptions, and that more complex models that incorporate the effects of journey-time and the differential draw of different facility types provide more realistic predictions of catchment boundaries. This finding is significant because data with which to implement these more complex models are not available for the vast majority of facilities across Kenya. As such, the only feasible approach to define catchments nationwide would be the use of Thiessen polygons and this has been shown here to be sub-optimal. A further consideration is that even these more complex catchment models only deliver predictions of the boundaries between catchments. Whilst this provides useful information, the issue of within-catchment utilisation rate remains largely unaddressed, due largely to the complexity of factors involved and the absence of reliable and comprehensive data.

In summary, the studies presented here have contributed significantly to the understanding of the spatial patterns of care-seeking behaviour in Kenya, and have developed a series of novel techniques for assessing these patterns and generating refined catchment boundary models. However, the goal of most significance to the current project, the estimation of catchment populations for all health facilities across Kenya, remains elusive due largely to the unavailability of data on a national scale. The unavailability of these catchment population estimates has led in this project to the development of an alternative set of approaches to standardising MC data and these are discussed in the proceeding chapters of this thesis.

6.5 Chapter summary

This chapter has presented two studies that have addressed the issue of defining spatial models to predict facility catchment populations. In the first study, the most rudimentary and widely-used catchment model, based on a Thiessen polygon approach, was investigated. Novel tests were developed using polygon-based patient use data within a GIS that determined that the two principal assumptions implicit in a Thiessen polygon model were inappropriate: that all patients use their nearest facility and that utilisation rate within each catchment is constant. In the second study, the use of Euclidean distance as a journey metric was replaced with a more realistic measure that used information on transport networks, gradient, and natural barriers to estimate the journey-time between care-seekers and different facilities. This approach allowed patient-use patterns to be assessed more realistically and a catchment model was produced that predicted patient's choice of facility more accurately. By incorporating the fact that different facility types are able to draw in patients from greater distances, another catchment model was produced that further improved prediction accuracy.

These studies have indicated that, in order to predict most accurately which facilities patients will use when seeking care, the more complex catchment models must be implemented. Because the data do not exist currently to support this implementation for health facilities across Kenya (being unavailable outside the four study districts), and because even these models do not encapsulate the complexities that determine within-catchment utilisation patterns, it has not been possible to produce the accurate estimates of facility catchment populations that are required to act as denominators with which to standardise the raw MC data in the HMIS database. Because of this, an alternative strategy has been developed in this project that uses TC data on the total number of monthly outpatients at facilities as a proxy measure of catchment populations. The development and testing of this approach is described in the remaining chapters of this thesis.

Chapter 7

Model Development 1: Development and Evaluation of Kriging Approaches

Chapter 7

7. Model Development 1: Development and Evaluation of Kriging Approaches

7.1 Introduction

Chapter 5 described the need to standardise raw MC data to account for non-spatial facility-specific factors that are likely to confound the inherent spatial structure driven by the underlying pattern of malaria morbidity in Kenya. The preceding chapter presented a series of catchment modelling studies and highlighted that reliable estimates of facility catchment populations for use as denominator values remain unavailable for facilities across Kenya. In response to this, an alternative strategy was developed in this project in which TC data detailing the total number of all-cause diagnoses for each facility and month were used as a way of standardising the MC data. This strategy was presented in Chapter 5 along with two modelling frameworks, Model 2 and Model 3, that incorporated this approach. The basic non-standardised model in which MC is predicted directly was also presented and this was labelled Model 1. The task of developing these frameworks and identifying which approach is likely to provide the most accurate predictions of MC was split into two stages. The set of three modelling frameworks (presented in Figure 5.5) consisted of four different prediction exercises to predict MC, TC, MP and SMC. The first stage was to establish the most appropriate geostatistical methodology for carrying out these prediction tasks and this is the subject of the current chapter. The second stage was to adopt the chosen prediction methodology and implement it within each of the three models to obtain predictions that could be

compared to identify which predicts MC most accurately, and this is the subject of Chapter 8.

The approach taken in this chapter has been to focus on one of the four variables listed above and to compare the performance of three alternative geostatistical prediction approaches in a cross-validation setting. The malaria proportion variable, MP, was chosen as the test variable because, of the two standardised variables (MP and SMC), MP is the more straightforward to obtain and interpret as it is simply the MC value at each facility-month divided by the corresponding TC value.

7.2 Background

As described in Chapter 4, geostatistical prediction techniques were originally developed for, and remain principally targeted at, spatial-only settings (Matheron, 1971; Goovaerts, 1997; Chilès and Delfiner, 1999). When sampled and unsampled locations are distributed through time as well as space, however, the replacement of spatial-only with space-time geostatistical approaches can offer several benefits including more data to support parameter estimation and prediction and, if present, the exploitation of temporal as well as spatial autocorrelation in observed values. Both spatial-only and space-time geostatistical prediction techniques generally rely on the adoption of a stationary RF model parameterised with a stationary variogram. Where a property of interest displays heterogeneous first and second-order characteristics, however, alternative non-stationary models may be more appropriate and yield more accurate predictions (Haas, 1995).

In this chapter, the MP variable has been taken as a test variable and three different geostatistical prediction methodologies are developed and implemented. The objective is to examine the effect on prediction accuracy of (a) the extension of a spatial-only to a space-time prediction approach, and (b) the replacement of a stationary space-time RF model which requires a single global space-time variogram with a locally-varying space-time RF model which allows the space-time variogram to vary across the study domain.

7.3 Methodology

The dataset used in this chapter is the 1765-facility HMIS test set which is described in detail in Chapter 3. MC and TC data were available at 63,542 facility-months and were converted to MP using simply MP = MC/TC. When a disease count (MC) is converted to a proportion (MP) based on a background denominator value (TC), the uncertainty of that proportion can be highly sensitive to the magnitude of the denominator. As a preliminary analysis, the effect of TC on MP variance was checked visually (not shown) and found to be minimal, with variance approximately constant for all values of TC. This can be explained by the consistently large TC values (less than 0.2% of TC values were <30 cases) and the fact that malaria is the most common diagnosis meaning that MC values were generally a substantial proportion of TC. It was decided, therefore, that no aggregation of the monthly MP values was necessary prior to their use in the subsequent prediction exercises.

Three alternative methodologies were used to obtain predictions of MP at individual facility-months in three separate cross-validation procedures. These were OK, STOK, and local space-time ordinary kriging (LSTOK). OK and STOK have been described in detail in sections 4.3.3.2, and 4.4.2, respectively. The procedure used for LSTOK is developed and described in full in this chapter. The cross-validation procedure is explained in section 4.3.3.6 in a spatial-only setting, and its extension to a space-time setting is straightforward.

7.3.1 Spatial-only prediction of MP

The full set of n = 63,542 MP data $\{z((\mathbf{u},t)_\alpha);\alpha=1,...,n\}$ was divided by month into $\{j = 1,...,m\}$ spatial-only sets $\{z_j(\mathbf{u}_\delta);\delta=1,...,p(j)\}$ where m = 84 months, and the size of each set, $p(j)$, varied between months. For each spatial-only set, OK was carried out in the following steps to obtain a set of $p(j)$ cross-validation predictions $\{z_j^*(\mathbf{u}_\delta);\delta=1,...,p(j)\}$.

(1) An omnidirectional sample spatial variogram (4.11) was estimated from the data using the established method-of-moments approach (Deutsch and Journel, 1998, p.53).

(2) A suitable model was fitted by eye to the omnidirectional variogram from the set of five models listed in section 4.3.2.2, the spherical, exponential, Gaussian, power, and periodic models. Due to the large number of variograms involved, a parsimonious model structure was adopted for each consisting of a single structured model component. The spherical model was selected as offering the best fit to the estimated semivariance values. More importance was attached to ensuring a good fit near the ordinate as values of the variogram at smaller lag separations have more influence in the subsequent kriging. In addition to a spherical component, each model included a nugget component (4.20) to model a discontinuity in semivariance at the ordinate.

(3) OK was implemented with the variogram model parameters from (2) to obtain cross-validation predictions $z_j^*(\mathbf{u}_\delta)$ using the GSLIB *kt3d* routine (Deutsch and Journel, 1998). The search neighbourhood for each prediction consisted of the 50 data closest (using Euclidean distance) to the prediction point. A single space-time set of n cross-validation predictions, $\{z_{\mathrm{OK}}^*((\mathbf{u},t)_\alpha);\alpha=1,\ldots,n\}$, (subscripted $_{\mathrm{OK}}$ to denote prediction using spatial-only OK) was then created by joining each of the m spatial-only sets of cross-validation predictions, $z_{\mathrm{OK}}^*((\mathbf{u},t)_\alpha)=\bigcup_{j=1}^{m} z_j^*(\mathbf{u}_\delta)$.

7.3.2 Space-time prediction of MP

STOK was carried out using the full space-time set of n = 63,542 MP data $\{z((\mathbf{u},t)_\alpha);\alpha=1,\ldots,n\}$ to obtain a set of n cross-validation predictions $\{z_{\mathrm{STOK}}^*((\mathbf{u},t)_\alpha);\alpha=1,\ldots,n\}$ to compare to the n data in the following steps:

(1) A sample space-time variogram surface $\hat{\gamma}_{st}(\mathbf{h}_s,h_t)$ was calculated from the data (4.48) using a modified space-time GSLIB *gamv* routine (De Cesare et al., 2002) (see Figure 7.4 (a)). Steps 2-4, below, were then implemented to use this surface to estimate parameters of the product-sum space-time variogram model described in section 4.4.3 (De Cesare et al., 2001, 2002).

(2) Space- and time-marginal variograms were estimated from the space-time variogram surface as $\hat{\gamma}_{st}(\mathbf{h}_s,0)$ and $\hat{\gamma}_{st}(0,h_t)$ by setting h_t = 0 and $\mathbf{h}_s$ = 0, respectively (see De

Cesare et al., 2001, p12). The space-marginal variogram is equivalent to the mean of the 84 monthly spatial-only variograms, whilst the time-marginal variogram is equivalent to the mean of temporal variograms for each of the 1765 facility locations.

(3) Variogram models were fitted by eye to the sample space- and time-marginal variograms. As for the spatial-only variograms described in the previous section, greater emphasis was placed on ensuring a good fit at smaller lags. Since manual model fitting was required for only one spatial and one temporal variogram, a more complex model structure could be adopted, allowing the use of multiple nested structured components from the list described above to provide a closer fit. The space-marginal variogram was fitted with a nested model consisting of a nugget, an exponential, and a spherical component and the time-marginal variogram was fitted with a nested model consisting of a nugget, an exponential, a periodic, and a spherical component (see Table 7.1 for model parameters).

(4) The space-time sill, $C_{st}(0,0)$, was estimated directly from the space-time variogram surface.

(5) The space-time sill and parameters from the space- and time-marginal variogram models were used to define a product-sum space-time variogram model (4.52) (see Figure 7.4 (d).

(6) This variogram model, $\tilde{\gamma}_{st}(\mathbf{h}_s, h_t)$, was then used as input in a STOK procedure to obtain cross-validation predictions $z^*_{\text{STOK}}((\mathbf{u},t)_\alpha)$ using a modified space-time GSLIB *kt3d* routine (De Cesare et al., 2002).

7.3.2.1 Parameterising a space-time search criteria

As in the spatial-only case, the search neighbourhood for each prediction consisted of the 50 data 'closest' to the prediction point. Unlike the spatial-only case, however, the metric by which this closeness can be assessed is not straightforward. In the approach adopted, a space-time metric is defined that can be used to quantify and compare the relative space-time distances between candidate data points and the prediction location.

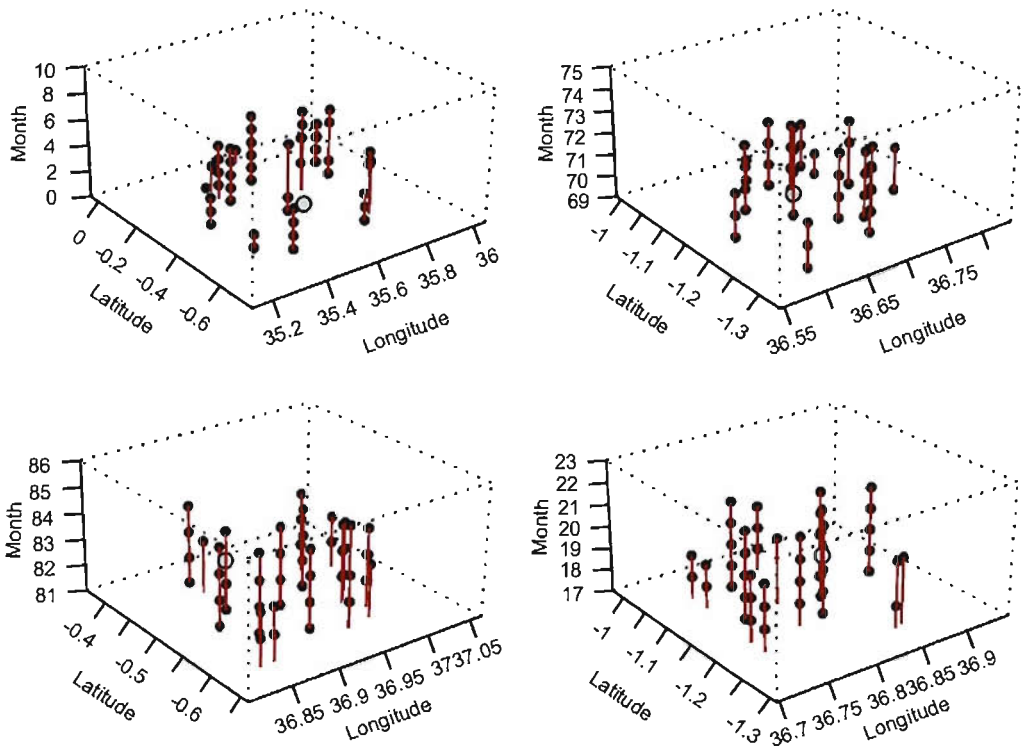


Figure 7.1 Four examples of space-time kriging search neighbourhoods resulting from the parameterisation of the space-time search criteria. In each case, the prediction location is shown (grey dot) along with the data included in its prediction (black dots). Vertical red bars are included to provide perspective, linking data from different months at the same spatial location. Spatial axes are shown in decimal degrees.

If a prediction location has the space-time coordinates (x, y, t) and a candidate datum has the space-time coordinates (x', y', t'), then the spatial separation $|\mathbf{h}|$ is found simply by $|\mathbf{h}| = \sqrt{(x-x')^2 + (y-y)^2}$ and the temporal separation h is found by $h = t - t'$. In order to obtain a single space-time metric, these absolute measures of spatial and temporal separation are represented as relative proportions of the maximum spatial and temporal search radii, S_{max} and T_{max}, as set by the user. The space-time metric h_{ST} is then defined as:

$$h_{st} = \sqrt{\left(\frac{(x-x')}{S_{max}}\right)^2 + \left(\frac{(y-y')}{S_{max}}\right)^2 + \left(\frac{(t-t')}{T_{max}}\right)^2} \qquad (7.1)$$

The choice of values for the S_{max} and T_{max} parameters was determined heuristically to

allow the influence of spatial and temporal separation to be approximately equivalent, and to provide approximately 'spherical' search neighbourhoods defined in space-time. Values of S_{max} = 450 km and T_{max} = 84 months were chosen and Figure 7.1 shows examples of four search neighbourhoods resulting from these parameter settings. This parameterisation is inevitably somewhat arbitrary, and the decision to use a relatively large number of data (50) was made in part to provide a set that would contain sufficient data distributed in both space and time to allow appropriate influence to be assigned via the kriging weights.

It is important to emphasise that this space-time metric is used in the STOK procedure only to define those data that are used in each prediction, and not in the subsequent kriging algorithm. The calculation of covariance values between points separated in space and time respected the absolute spatial and temporal lags between data and between data and predictions.

7.3.3 Local space-time prediction of MP

The use of STOK, as with OK, implies the adoption of an RF model with stationary mean and variogram. Where first-order heterogeneities exist, the effect on prediction accuracy is often attenuated in practice because each prediction is derived from $n((\mathbf{u}, t)_0)$ observations within a limited local space-time neighbourhood $W((\mathbf{u}, t)_0)$ centred on the prediction location $((\mathbf{u}, t)_0)$ rather than from all n observations throughout the global study domain (as explained in the previous section). As such, the required domain of stationarity for each prediction is reduced to the neighbourhood $W((\mathbf{u}, t)_0)$. In the standard form, however, STOK, as with OK, has no such mechanism to attenuate the effects of covariance heterogeneities since it is reliant on the global sample space-time variogram, $\hat{\gamma}_{st}(\mathbf{h}_s, h_t)$, which is estimated from all n data under the assumption of stationarity. An alternative approach is to adopt a RF model that is globally non-stationary, that is, stationarity is considered to exist only within local neighbourhoods (Journel and Huijbregts, 1978; Haas, 1990). This approach was implemented here in a space-time context (denoted local space-time ordinary kriging, LSTOK) to obtain a set of n local cross-validation predictions $\{z^*_{\mathrm{LSTOK}}((\mathbf{u},t)_\alpha);\alpha=1,\ldots,n\}$ to compare to the n data in the following steps:

(1) The space-time set of n = 63,542 MP data $\{z((\mathbf{u},t)_\alpha);\alpha=1,...,n\}$ were distributed at l spatial locations $\{\mathbf{u}_\delta;\ \delta = 1,\ldots,l\}$ where l = 1765, the number of health facilities in the data set. For each of the l spatial locations $\mathbf{u}_\delta$ where one or more of the n cross-validated predictions, $z^*((\mathbf{u},\ t)_\alpha)$, was required, a space-time 'cylinder' (Haas, 1995) was defined in which to estimate a spatially-local space-time sample variogram, $\hat{\gamma}_{st}(\mathbf{h}_s, h_t; \mathbf{u}_\delta)$. Each cylinder consisted of a subset of c=1,2,...,$n(\mathbf{u}_\delta)$ data, $\{z_\delta((\mathbf{u},t)_c);c=1,...,n(\mathbf{u}_\delta)\}$. Each subset was identified as all data located within the nearest l_c = 100 locations in space to the prediction location $\mathbf{u}_\delta$, and at any month. The 'radius' of each cylinder was therefore equal to the distance from the prediction location $\mathbf{u}_\delta$ to its 100th nearest observation in space, and its 'height' was m = 84 months. This approach meant local neighbourhoods were restricted spatially but not temporally. A balance had to be struck between neighbourhood size (with smaller neighbourhoods considered more appropriate to model as being stationary) and the resulting sample size within each neighbourhood, $n(\mathbf{u}_\delta)$, with which to estimate each local sample variogram (with smaller subsets resulting in larger uncertainty in the sample variogram). Exploratory analysis of time-series of MP at different spatial locations (not shown) and of the sample time-marginal MP variogram (see Figure 7.4 (b)) did not suggest the presence of second-order heterogeneity through time. As such, it was decided to include all data through time within each cylinder in order to maximise the sample size $n(\mathbf{u}_\delta)$ for a given spatially-limited neighbourhood.

(2) Spatially-local space-time sample variograms were calculated for each spatial location $\mathbf{u}_\delta$ using the same procedure as for section 7.3.2(1) but applied only to the subset within each spatially-local cylinder, $\{z_\delta(\mathbf{u}_c,t_c);c=1,...,n(\mathbf{u}_\delta)\}$. After assessing the stability of semivariance estimates at the larger lags, it was decided to model spatial lags up to a maximum of 80% of the diameter of each cylinder and temporal lags up to a maximum of 20 months.

(3) A fitted product-sum space-time variogram model was required for each of the 1765 local variograms. This large number prohibited use of the manual procedure detailed in section 7.3.2(3-5) and an automated procedure was developed to replicate these steps. Although estimated and modelled variograms could not be inspected at all 1765 locations, it was necessary to sample the results of the automatic procedure in order to make modelling decisions. As such, a set of 50 prediction locations was selected at

random and manually checked at each stage. The automatic procedure operated as follows for each local variogram.

(3.i) Sample space- and time-marginal variograms were estimated from the sample space-time variogram surface as $\hat{\gamma}_{s,t}(\mathbf{h}_s,0)$ and $\hat{\gamma}_{s,t}(0,h_t)$ by setting $h_t = 0$ and $\mathbf{h}_s = 0$, respectively.

(3.ii) Separate 1-D models were fitted to the sample space- and time-marginal variograms using a weighted least-squares (WLS) procedure (for brevity, the following description focuses on the space-marginal variogram, although the equivalent procedure was also applied to the time-marginal variogram). In order to minimise the computational requirements of parameter estimation, and following examination of the 50 monitored local sample variograms, a parsimonious 1-D model consisting of a nugget component and a single spherical component was selected for fitting to all space-marginal variograms. As such the required parameter set, $\boldsymbol{\theta}$, to be estimated for each 1-D model consisted of three parameters, ($\boldsymbol{\theta} = \{c_0, a_{\text{sph}}, c_{\text{sph}},\}$), where a_{sph} is the range parameter of the spherical component and c_0 and c_{sph} are the sill parameters of the nugget and spherical components, respectively (Deutsch and Journel, 1998). $\boldsymbol{\theta}$ was estimated using a nested grid-search algorithm written in ANSI C. The three-parameter 1-D variogram model described above was fitted manually to the sample space-marginal variogram estimated from the global space-time sample variogram as described earlier in section 7.3.2(1-2) and the resulting parameter set was used as starting values to initialise the algorithm.

The nested grid-search approach consisted of calculating an objective function, $F(\boldsymbol{\theta})$, described below, at a set of evenly-spaced locations in the 3-d parameter space around the starting values. In the first iteration, $j = 50$ values of each parameter were evaluated, meaning objective functions were calculated for $j^3 = 1.25\times10^5$ different parameter sets. The range of parameter values to test in the first iteration was determined heuristically to include a broad swathe of parameter space around the starting values. The range of parameter values was constrained such that impossible values (i.e. $c_0 < 0$, $a_{\text{sph}} < 0$, $c_{\text{sph}} < 0$) were not permitted. The parameter set that minimised $F(\boldsymbol{\theta})$ was identified and became the starting set for the next iteration. Each subsequent iteration evaluated j^3 evenly-

spaced parameters over a progressively smaller region of the parameter space, each time identifying the parameter set that minimised $F(\boldsymbol{\theta})$. The extent to which each iteration converged on progressively smaller regions, and the total number of iterations carried out were again determined heuristically, by examining the fit of the resulting models for the 50 monitored variograms.

The objective function $F(\boldsymbol{\theta})$ (Pardo-Igúzquiza, 1999) evaluated for each parameter set was calculated as a weighted sum of squared differences between the sample space-marginal variogram, $\hat{\gamma}(i)$, at each i=1,2,...,n lags and the value of the variogram model under this parameter set, $\tilde{\gamma}(i;\boldsymbol{\theta})$:

$$F(\boldsymbol{\theta}) = \sum_{i=1}^{n} w(i) \cdot [\hat{\gamma}(i) - \tilde{\gamma}(i;\boldsymbol{\theta})]^2 \tag{7.2}$$

The weighting scheme used to determine $w(i)$ was defined as:

$$w(i) = \frac{m(i)}{[\tilde{\gamma}(i;\boldsymbol{\theta})]^2} \tag{7.3}$$

where $m(i)$ is the number of data pairs used to estimate $\hat{\gamma}(i)$. In this scheme, each variogram estimate $\hat{\gamma}(i)$ is weighted in approximately inverse proportion to its estimation variance (Cressie, 1985).

(3.iii) Having estimated the parameter sets for the space- and time-marginal variograms, $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_t$, the remaining parameter required for the definition of the space-time variogram model was the space-time sill, $C_{st}(0,0)$. A starting value for $C_{st}(0,0)$ was estimated from a manual fit of the global space-time variogram where all the other parameters were provided by $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_t$ and held constant. The WLS procedure described above was then implemented in the 1-D parameter space to estimate the value of $C_{st}(0,0)$.

(4) LSTOK was then implemented to obtain n cross-validation predictions $\{z^*_{\mathrm{LSTOK}}((\mathbf{u},t)_\alpha);\alpha=1,...,n\}$. The kriging algorithm was identical to that used for the global

STOK described in section 7.3.2 except that, for each prediction, the relevant spatially-local space-time variogram model replaced the global model.

7.3.4 Comparison of prediction accuracies

The OK, STOK, and LSTOK prediction methodologies described above each resulted in a set of n = 63,542 cross-validation predictions of MP $\{z^*((\mathbf{u},t)_\alpha);\alpha=1,...,n\}$ to compare to the n MP data $\{z((\mathbf{u},t)_\alpha);\alpha=1,...,n\}$. To compare the performance of the different methods, three summary statistics were calculated for each. These were the correlation coefficient between the predicted and actual set, the ME, and the MAE (defined in section 4.3.3.6 for the spatial-only case). 2-D histograms were produced to display graphically the bivariate distribution of the data and corresponding predicted values. These plots are more informative than scatter-plots when the number of data-prediction pairs is large. Univariate histograms were also produced for each set of prediction errors, $\{z^*((\mathbf{u},t)_\alpha) - z((\mathbf{u},t)_\alpha);\alpha = 1,...,n\}$.

As mentioned in Chapter 4, The use of cross-validation as a method of accuracy assessment is limited by a number of factors. The use of simple arithmetic averages to generate estimates of ME and MAE may result in biased estimates when the data are clustered in space and/or time. In the current case, however, it is important to distinguish between spatial clustering of the set of facilities and clustering of the data themselves in relation to this background pattern. When an arithmetic average of an attribute at the data locations is used to estimate the mean of that attribute at the unsampled locations, the spatial or spatiotemporal arrangement of the combined set of sampled and unsampled points has no effect on the estimate. Rather, it is the arrangement of the sampled points *within* this combined set that may introduce bias if they are highly clustered. Although the set of facilities are highly spatially clustered (see Figure 3.1), reflecting approximately the spatial distribution of the Kenyan population, the spatiotemporal pattern of sampled points within the set of all points did not display substantial clustering either spatially or temporally. The use of cross-validation statistics simply as relative measures of the accuracy of different prediction methods further mitigates the effect of the limitations described above, since such effects are consistent between methods.

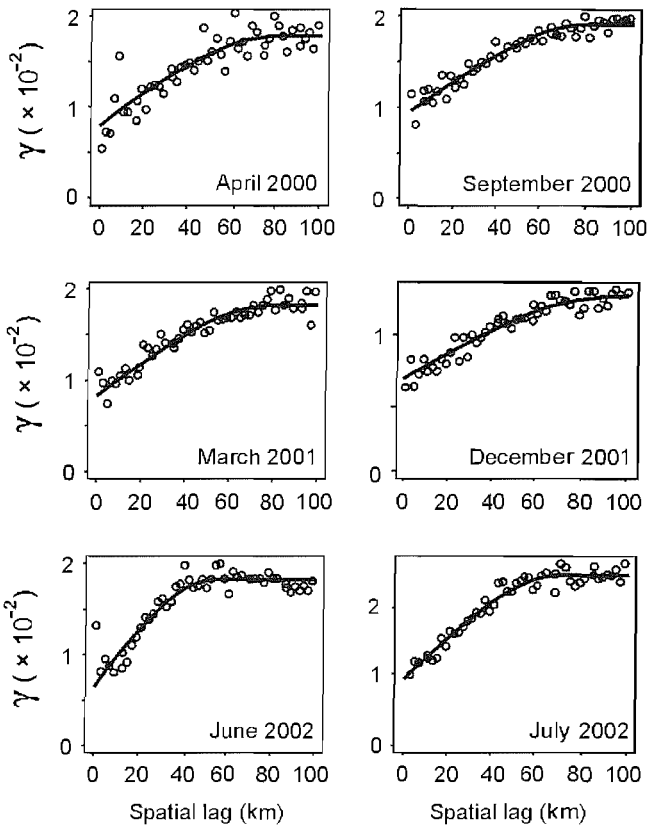


Figure 7.2 Sample spatial variograms (circles) and fitted variogram models (line) for malaria proportion in six different months during 2000, 2001, and 2002. A total of 84 such variograms were estimated and modelled, one for each month of the study period January 1996 – December 2002.

7.4 Results

7.4.1 Variography

Figure 7.2 shows spatial variograms that were estimated from spatial-only data for each of the 84 months in the data set, and the corresponding manually-fitted variogram models. The model parameters of these variograms are listed in Table 7.1. A selection of six of these variograms are presented in a larger format in Figure 7.3. Sample variogram structure was consistent across the different monthly sample variograms, which supported the use of the same class of variogram model (with a nugget and single spherical component) throughout. The estimated range, sill and nugget parameter values, however, displayed considerable variation between months although no clear patterns

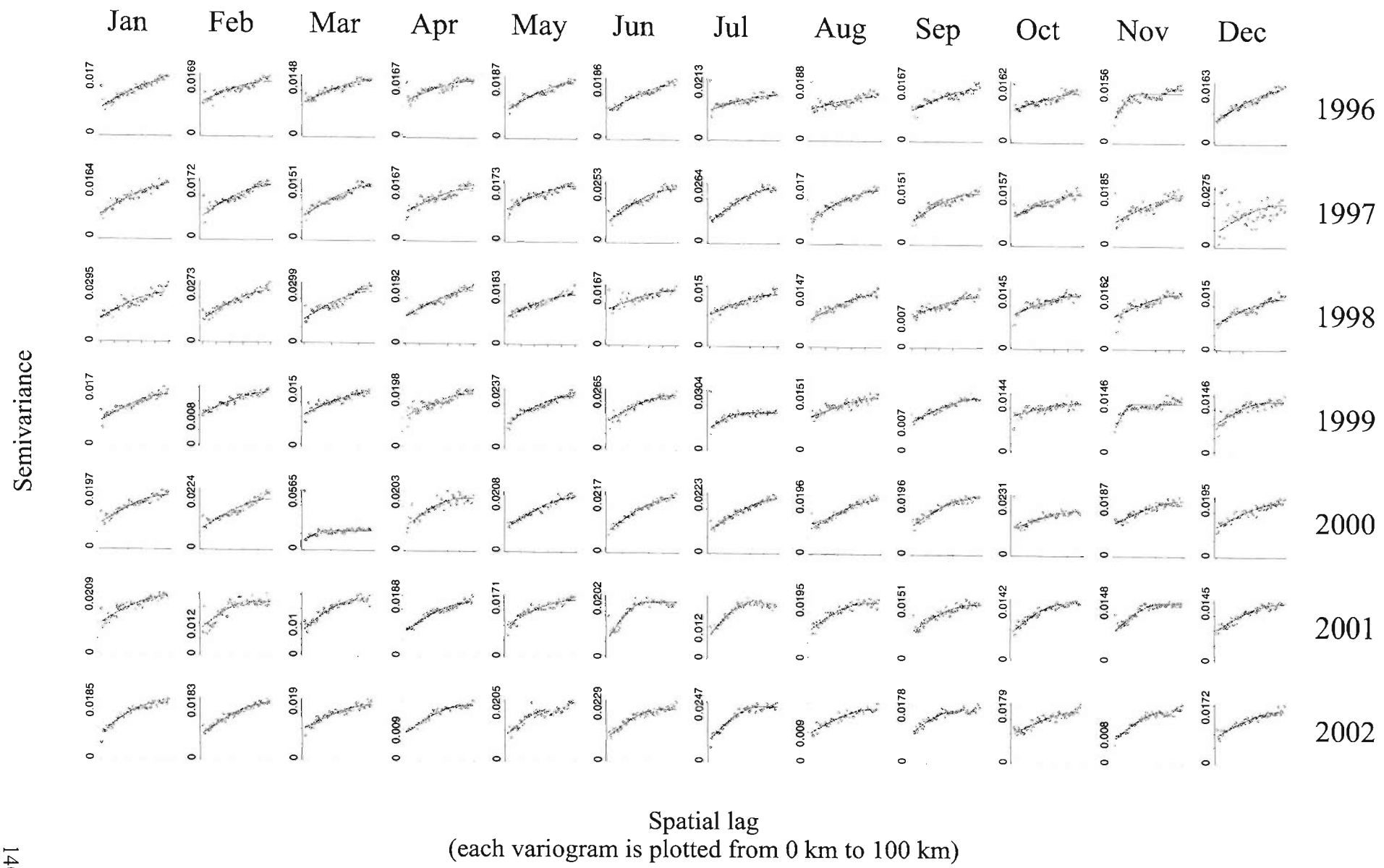


Figure 7.3 Sample spatial variograms (circles) and fitted variogram models (lines) for malaria proportion for all 84 months in the study period January 1996 – December 2002.

Month	a_{SPH}	c_{SPH}	c_{NUG}	Month	a_{SPH}	c_{SPH}	c_{NUG}
1	98.9900	0.00825	0.00736	43	49.2315	0.00835	0.01076
2	98.5600	0.00568	0.00957	44	94.4500	0.00532	0.00849
3	97.6200	0.00554	0.00826	45	94.4700	0.00596	0.00690
4	97.0300	0.00585	0.00990	46	91.2986	0.00312	0.00897
5	96.8700	0.00873	0.00843	47	20.8277	0.00599	0.00602
6	98.5600	0.00858	0.00832	48	75.9037	0.00555	0.00754
7	95.9700	0.00567	0.01020	49	96.9700	0.00951	0.00839
8	95.5000	0.00428	0.00953	50	96.0000	0.01143	0.00755
9	99.1900	0.00640	0.00834	51	33.2939	0.00974	0.00745
10	94.0100	0.00486	0.00849	52	82.0123	0.01105	0.00887
11	27.8136	0.00655	0.00617	53	97.4300	0.01088	0.00812
12	97.2100	0.00860	0.00601	54	99.6800	0.01264	0.00774
13	95.2600	0.00815	0.00669	55	98.7300	0.01226	0.00813
14	94.5200	0.00901	0.00661	56	98.8800	0.00994	0.00794
15	99.3100	0.00768	0.00605	57	80.8380	0.00903	0.00875
16	97.2400	0.00711	0.00783	58	92.0428	0.00677	0.01036
17	98.3200	0.00703	0.00930	59	89.1899	0.00700	0.01014
18	99.3200	0.01473	0.00852	60	98.7500	0.00864	0.00940
19	88.8497	0.01592	0.00880	61	91.8470	0.00911	0.01042
20	96.4900	0.00866	0.00685	62	71.0508	0.01026	0.01045
21	86.9156	0.00728	0.00591	63	81.1953	0.00998	0.00830
22	96.0400	0.00605	0.00766	64	95.5400	0.00915	0.00753
23	95.6200	0.00832	0.00736	65	95.0744	0.00744	0.00860
24	89.5156	0.01199	0.01250	66	53.0273	0.01357	0.00756
25	99.3400	0.01313	0.01155	67	55.0918	0.01285	0.00885
26	95.8600	0.01369	0.01034	68	75.5966	0.00870	0.00933
27	98.6800	0.01450	0.01182	69	84.5026	0.00696	0.00681
28	98.4400	0.00872	0.00871	70	74.1880	0.00666	0.00670
29	97.6100	0.00702	0.00849	71	65.7891	0.00689	0.00721
30	96.5000	0.00554	0.01017	72	93.2243	0.00668	0.00732
31	95.1100	0.00538	0.00796	73	75.5810	0.00956	0.00759
32	97.2200	0.00624	0.00695	74	94.5700	0.00963	0.00774
33	99.3700	0.00479	0.00732	75	97.5527	0.00785	0.00964
34	92.3486	0.00483	0.00853	76	83.7580	0.00857	0.00796
35	94.3500	0.00623	0.00904	77	58.7546	0.00865	0.00905
36	98.8400	0.00657	0.00660	78	75.3251	0.01068	0.00954
37	98.4900	0.00794	0.00688	79	69.0179	0.01413	0.00897
38	95.5600	0.00626	0.00854	80	90.2984	0.00744	0.00887
39	98.3200	0.00597	0.00799	81	58.9361	0.00688	0.00896
40	94.8500	0.00796	0.01033	82	99.7900	0.00709	0.00881
41	95.3700	0.01153	0.00979	83	69.6865	0.00772	0.00642
42	89.5156	0.01199	0.01250	84	94.3455	0.00758	0.00780

Table 7.1 Parameters of the spherical variogram models fitted to each monthly spatial-only sample variogram of MP. Each model consisted of a single spherical component with range a_{SPH} (km) and sill c_{SPH}, and a nugget component c_{NUG}.

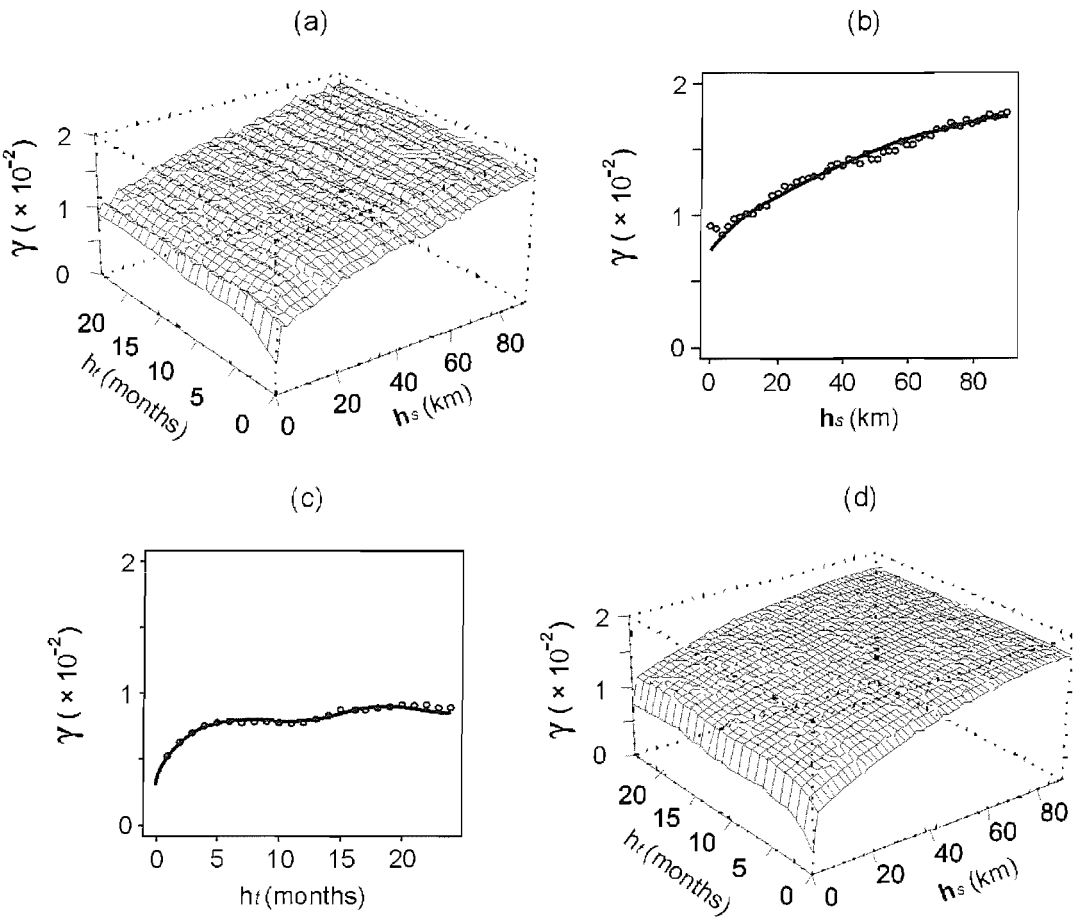


Figure 7.4 Space-time variography for malaria proportion. Plots shown are (a) the sample space-time variogram surface, (b) the sample space-marginal variogram (circles) with fitted 1-D model (line), (c) the sample time-marginal variogram (circles) with fitted 1-D model (line), and (d) the 2-D product-sum space-time variogram model. Each vertical axis measures semivariance, γ, and horizontal axes measure either spatial lag ($\mathbf{h}_s$) or temporal lag (h_t).

	Model type	*a*	*c*
Spatial Model	Nugget	-	0.0075
	Exponential	30	0.0016
	Spherical	85	0.0084
	Nugget	-	0.0030
Temporal Model	Exponential	3.3	0.0035
	Hole	6	0.0003
	Spherical	20	0.0020
Space-time sill	-	-	0.0178

Table 7.2 Parameters of the product-sum space-time variogram model for MP. Values of the range parameter *a* are given in kilometres for spatial model components and in months for the temporal model components. *c* refers to the sill parameter of each component.

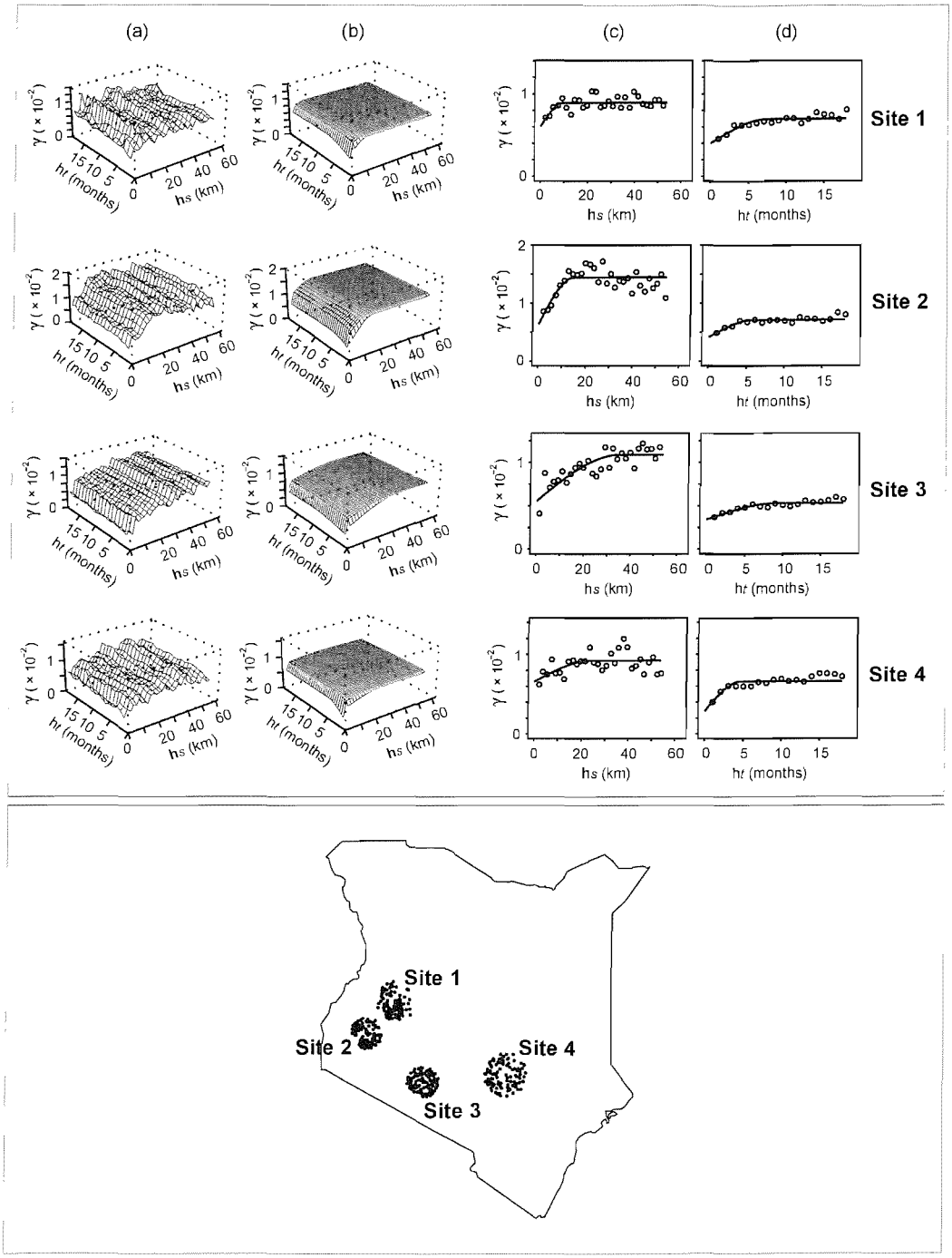


Figure 7.5 Examples of local space-time variography for four different locations (rows). Variography was carried out automatically in a local neighbourhood around each of the 1765 spatial locations where predictions were made. Plots shown for each location (upper box) are the sample space-time variogram surface (column (a)), the fitted 2-D product-sum space-time variogram model (column (b)), the sample space-marginal variogram (circles) with fitted 1-D model (line) (column (c)), and the sample time-marginal variogram (circles) with fitted 1-D model (line) (column (d)). Each vertical axis measures semivariance, γ, and horizontal axes measure either spatial lag ($\mathbf{h}_s$) or temporal lag (h_t). The local neighbourhoods in which each of these four local variograms were calculated are shown in the map (lower box). Each cluster (black dots) represents the spatial locations (facilities) that contributed data in each case.

could be discerned. Figure 7.4 shows the global sample space-time variogram surface and fitted product-sum model. Also shown are the sample space- and time-marginal variograms that were estimated using the sample surface, and the corresponding 1-D variogram models. The corresponding table of model parameters are shown in Table 7.2. The time-marginal variogram differed substantially in structure from the space-marginal variogram, with both a smaller modelled sill value and a smaller relative nugget effect indicative of greater autocorrelation through time than across space. The spatial variogram shows a small upturn in semivariance for the smallest lags. This effect can be attributed to the nature of facility-pairs at these separations. A disproportionate number of these pairs are cross-type: health facilities of the same type are rarely built so close together and it is more commonly the case that large facilities such as hospitals, for example, are surrounded closely by a number of smaller facilities such as health centres or dispensaries. The different facility types are more likely to have different MP values than their spatial separation would otherwise suggest, resulting in a relatively larger semivariance at these short lags. Figure 7.5 shows examples for four different locations of the automatic variography procedure implemented to estimate and model local sample space-time variograms for each of the 1765 spatially-local neighbourhoods. These four examples illustrate the spatial heterogeneity of the observed space-time autocorrelation structure, with space- and time-marginal variogram model parameters varying considerably between the four locations.

7.4.2 Comparison of prediction accuracies

Cross-validation summary statistics for OK, STOK, and LSTOK are shown in Table 7.3. Both space-time approaches, STOK and LSTOK, resulted in substantially larger values of the correlation coefficient ρ than OK (13.1% and 14.8% larger ρ, respectively), indicating larger linear correlation between data and prediction sets. ME was small (indicating small overall bias) for all three approaches, although differences between sets were considerable. The value for OK showed the largest bias and those for STOK and LSTOK and were substantially smaller (98.4% and 87.5% reductions in ME, respectively, relative to OK). The largest MAE was produced by OK predictions, indicating the largest average prediction inaccuracy, with STOK and LSTOK producing more accurate predictions (14.8% and 18.3% reductions in MAE, respectively, relative

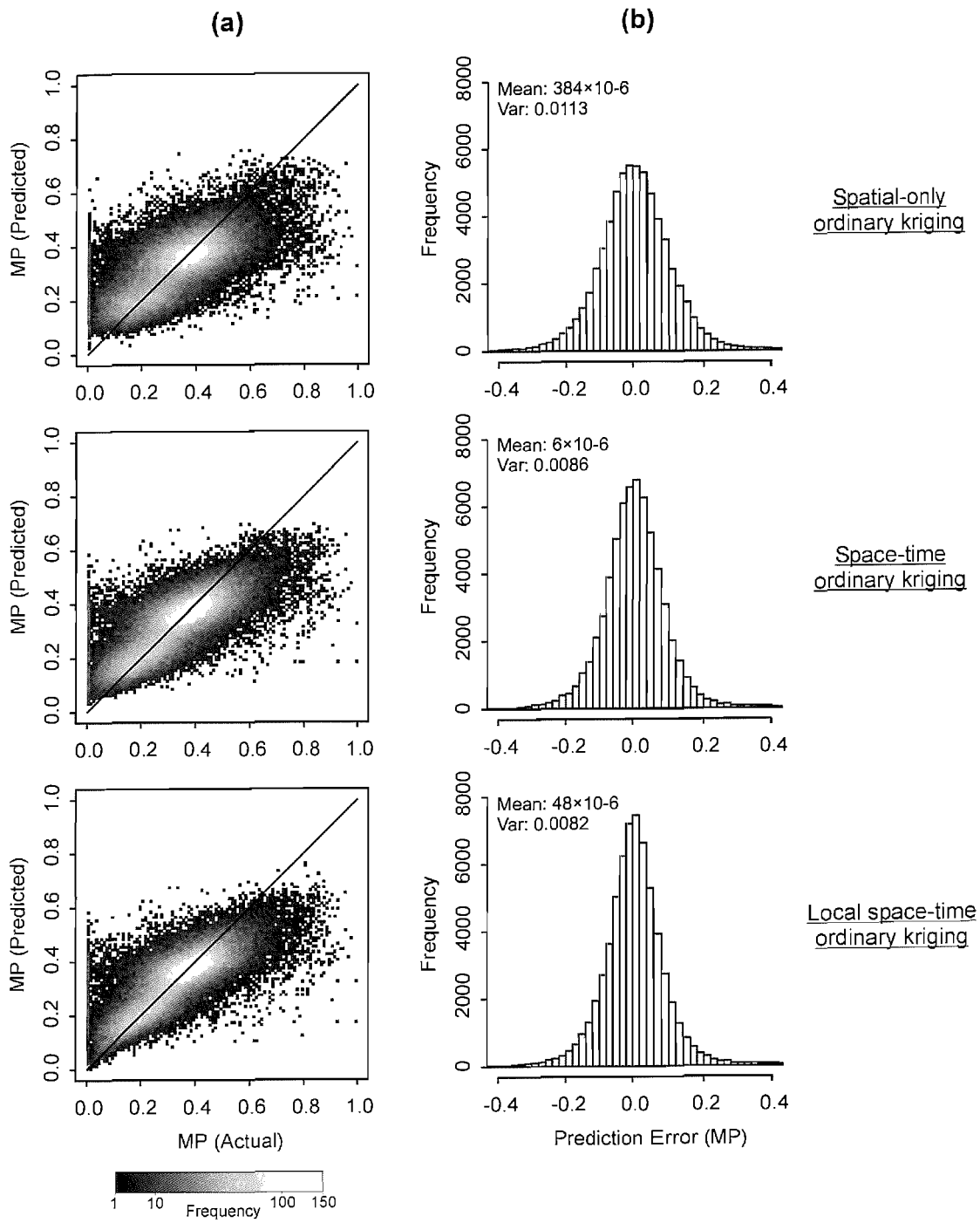


Figure 7.6 2-D histograms (column (a)) showing bivariate distribution of predicted against actual values for cross-validation predictions of malaria proportion (MP) using three different prediction approaches; spatial-only ordinary kriging, space-time ordinary kriging, and local space-time ordinary kriging. Whiter shading represents a higher frequency of values (note non-linear scale). The 1:1 line is also provided (diagonal black line) for each plot. Univariate histograms (column (b)) show the distribution of prediction error values for each prediction methodology. Error mean (Mean) and variance (Var) are also given.

Modelling Approach	ρ	ME	MAE
Spatial-only ordinary kriging (OK)	0.6764	0.000384	0.0796
Space-time ordinary kriging (STOK)	0.7651	0.000006	0.0678
Local space-time ordinary kriging (LSTOK)	0.7768	0.000048	0.0650

Table 7.3 Comparison of summary statistics for cross-validation predictions of malaria proportion using three different prediction approaches. The statistics shown are the correlation coefficient, ρ, the mean error (ME) and mean absolute error (MAE).

to OK). The overall pattern was that the space-time techniques offered less biased and more precise predictions than OK. Of the two space-time approaches, LSTOK provided more precise predictions than STOK but was slightly more biased overall, although bias was small in both cases.

Figure 7.6 (a) shows, for each prediction methodology, a 2-D cross-validation histogram illustrating the bivariate distribution of data and prediction sets. The patterns displayed support the summary statistic findings presented in Table 7.3 and discussed above. A 2-D cross-validation histogram for an accurate prediction exercise would show a high frequency of corresponding data and prediction values along a central region (indicating small imprecision), centred along the 1:1 line (indicating small bias). The 2-D histograms for OK, STOK, and LSTOK display progressively tighter central regions, with a greater frequency of values indicated by the whiter shading. Differences in bias are less noticeable, although the progressively smaller bias for OK, STOK, and LSTOK for small data values (e.g. <0.1) is clear if the bottom-left corner of each plot is compared. Univariate histograms showing the distribution of error values for each prediction are shown in Figure 7.6 (b). Errors are approximately Gaussian in each case and the progressively smaller error variances for OK, STOK, and LSTOK again correspond to respectively more precise predictions.

7.5 Discussion

7.5.1 Comparison of spatial-only and global space-time prediction

When predicting a space-time data set, a potential advantage of the spatial-only approach (e.g. OK) over the global space-time approach (e.g. STOK) is that the spatial variogram is able to vary through time since each month is modelled separately. In contrast, the global space-time variogram averages these individual spatial variograms and month-to-month variability is not represented in the model. This potential advantage of the spatial-only approach is offset by the need to partition the full space-time data set into monthly slices, which may each have insufficient data to obtain a stable estimate of the spatial variogram. A more serious limitation of the spatial-only approach is that any temporal structure present in the data is ignored. The results presented in the previous section showed that STOK yielded more accurate predictions than OK. The global sample space-time variogram (Figure 7.4) displayed substantial temporal autocorrelation and it is intuitive that prediction accuracy should be enhanced by exploiting this temporal structure, allowing predictions to be influenced by observations proximate in time as well as space. A further advantage of STOK over OK in the current context is that the former is significantly less labour-intensive, requiring the estimation and modelling of a single space-time variogram rather than 84 separate spatial variograms. The optimal choice between the two approaches will differ between settings contingent on a range of factors including the space-time distribution of the data and prediction points, and the relative magnitudes of spatial and temporal autocorrelation.

7.5.2 Comparison of global and local space-time prediction

The results described in the previous section showed that more precise predictions were obtained in the space-time prediction exercise when a single global space-time variogram (STOK) was replaced by local space-time variograms that were estimated and modelled for each prediction location using a spatially-local subset of data (LSTOK). As with the preceding comparison between OK and STOK, the relative costs and benefits of LSTOK over STOK in the current case may differ in another setting. Where predictions are to be made over a large region displaying second-order heterogeneity, and where

data exist at a sufficient density to support stable estimation of variograms within local neighbourhoods, the use of LSTOK offers the potential to provide greater prediction accuracy than STOK, as the current case illustrates. Furthermore, the adoption of a RF model with stationarity of order-two or intrinsic stationarity is likely to be more appropriate when these characteristics are considered to exist only within each local neighbourhood rather than throughout the study region.

The principal drawbacks of LSTOK are the difficulties involved in its implementation. Firstly, the calculation of a single sample space-time variogram is computationally expensive (if a spatial variogram is to be estimated at $n(\mathbf{h}_s)$ lags, and a temporal variogram at $n(h_t)$ lags, then the equivalent space-time sample variogram requires estimates at $n(\mathbf{h}_s)\times n(h_t)$ lags). Secondly, where local variograms must be estimated at a large number of locations, automatic variogram model fitting becomes necessary. Although procedures such as WLS allow the implementation of objective criteria for parameterisation, manual fitting is still widely favoured by practitioners of geostatistics as it allows the incorporation of prior knowledge of the property of interest in the variogram model. Algorithms to implement automatic fitting are, again, computationally expensive and can be unreliable, often meaning variogram models must be parametrically simple, with less nested components, than the equivalent manually-fitted models. The net effect of using many simple local variogram models compared to a single complex model will clearly depend on several factors including the nature of the global and local spatiotemporal autocorrelation structures being considered and the number of data available with which to estimate local variograms. In the current case, the use of LSTOK over 1765 spatially-local neighbourhoods has been shown to offer a

modest increase in prediction accuracy over STOK, although at a substantial additional cost in terms of dynamic memory requirements and CPU time.

7.6 Conclusion and implications

Three different geostatistical approaches that predict values of the standardised MP variable have been implemented to examine their relative prediction accuracies. The extension of the established spatial-only approach to a space-time approach yielded

substantially more accurate predictions. The further extension of this globally-stationary space-time approach to a locally-stationary space-time approach whereby space-time variograms were re-estimated for each prediction location within a spatially-local neighbourhood yielded a further increase in prediction precision, although was marginally more biased.

It was decided that the most appropriate choice of prediction technique for implementation of the three modelling frameworks was STOK. Although LSTOK provided a modest increase in prediction precision over STOK, this advantage was offset by a small increase in bias. A further reason for this decision was the dependence of the feasibility of the LSTOK procedure on the form of the spatial and temporal autocorrelation structures of the variable of interest. Whilst the sample space- and time-marginal variograms for MP had a form that could be reasonably represented with simple two-component (nested nugget and spherical model) variogram models, examination of the equivalent sample variograms for MP, SMC, and TC (presented in the next chapter) suggested more complex structures. As such, the automated procedure for fitting the large number of local variograms required by LSTOK was considered infeasible, and the advantages of manually fitting a complex model to a single global variogram were expected to outweigh any potential benefits of incorporating spatial heterogeneities in variogram form.

7.7 Chapter summary

The purpose of this chapter was to develop and test three different kriging approaches in order to identify the most suitable approach to implement in the three modelling frameworks presented in Chapter 5. Using MP as the test variable, the three different kriging methodologies were implemented to make cross-validation predictions of MP in order to test the effect on prediction accuracy of (a) the extension of a spatial-only to a space-time prediction approach, and (b) the replacement of a globally-stationary with a locally-varying random function model. Space-time kriging was found to produce predictions with 98.4% less mean bias and 14.8% smaller mean imprecision than conventional spatial-only kriging. A modification of space-time kriging that allowed space-time variograms to be recalculated for every prediction location within a spatially-

local neighbourhood resulted in a larger decrease in mean imprecision over ordinary kriging (18.3%) although mean bias was reduced less (87.5%). These results have led to the decision to use the STOK approach to implement the three modelling frameworks. In the next chapter this implementation is carried out, and the three modelling frameworks are compared.

Chapter 8

Model Development 2: Evaluation of Modelling Frameworks and Development of an Uncertainty Model

Chapter 8

8. Model Development 2: Evaluation of Modelling Frameworks and Development of an Uncertainty Model

8.1 Introduction

In Chapter 5, three modelling frameworks were presented for predicting unsampled MC values within the HMIS database. Model 1 represented the null model, with MC being predicted directly using the raw MC data. Models 2 and 3 represented two alternative approaches by which TC data could be incorporated as a way of standardising the raw MC data to mitigate the effect of non-spatial facility-specific factors that may confound the spatial structure that would otherwise be present in MC. The modelling frameworks all consisted of one or more geostatistical prediction exercises. In Chapter 7, a series of different kriging methods was tested to identify the most suitable approach to use for these predictions and STOK was chosen as the most appropriate technique. In this chapter, the three modelling frameworks are implemented to obtain predictions of MC. These predictions are then compared and the modelling framework that provides the most accurate predictions is identified.

If the chosen model is to be implemented to deliver predictions of the total treatment burden for malaria, then it is critically important that such predictions are accompanied by measures of their uncertainty. In this chapter, the established geostatistical technique of stochastic simulation is adapted to a space-time setting and used to represent the different sources of prediction

uncertainty within the framework of the chosen model. The resulting uncertainty model was tested by applying it in a cross-validation sense, such that each uncertainty estimate could be compared to a known prediction error.

8.2 Methodology 1: Comparison of modelling frameworks

8.2.1 Implementation of modelling frameworks

As in Chapter 7, the HMIS test set from 1765 facilities was used in this chapter, consisting of co-located data on MC and TC from 63,542 facility-months originating from 126 hospitals (4682 records), 445 health centres (18,669 records), and 1194 dispensaries (40,191 records). The three modelling frameworks are presented in section 5.4 and summarised in Figure 5.5. In total, the three modelling frameworks comprised four individual prediction exercises to predict MC directly (Model 1), TC (Models 2 and 3), MP (Model 2) and SMC (Model 3). Each framework was implemented using data from all facilities combined, and also separately for the three facility classes. This meant that a total of 12 prediction exercises was carried out, implementing each of the three modelling framework for the four facility categories. Each of the 12 prediction exercises followed a similar procedure, as follows.

(1) Firstly, the space-time sample variogram surface $\hat{\gamma}_{st}(\mathbf{h}_s, h_t)$ was estimated as described in section 4.4.2. Variograms were modelled up to spatial lags of 100 km and temporal lags of 24 months. Since the objective was to interpolate (fill in gaps), rather than to extrapolate (predict into the future), time was considered isotropic (i.e. temporal lag was defined only by the number of months, and not by direction in time).

(2) The product-sum space-time variogram model (4.52) presented in section 4.4.3 (De Cesare et al. 2001, 2002) was then fitted to the sample variogram surface $\hat{\gamma}_{st}(\mathbf{h}_s, h_t)$ using the procedure outlined in section 7.3.2 of the previous chapter. In brief, sample space-and time-marginal variograms were estimated from $\hat{\gamma}_{st}(\mathbf{h}_s, h_t)$ as $\hat{\gamma}_{st}(\mathbf{h}_s, 0)$ and $\hat{\gamma}_{st}(0, h_t)$ by setting $h_t = 0$ and $\mathbf{h}_s = 0$, respectively, and variogram models were fitted to these marginal variograms by eye. The space-time sill $C_{st}(0,0)$ was also estimated from $\hat{\gamma}_{st}(\mathbf{h}_s, h_t)$ and these parameters were used to define the product-sum variogram model $\tilde{\gamma}_{st}(\mathbf{h}_s, h_t)$.

(3) The space-time variogram model was used as input into an STOK prediction carried out using the space-time GSLIB *kt3d* routine (Deutsch and Journel, 1998) modified to allow prediction of space-time points using product-sum space-time covariance structures (De Cesare et al., 2002). Again, this procedure is explained in full in section 7.3.2.

8.2.2 Comparison of modelling frameworks

Each modelling framework was implemented as a cross-validation (see section 4.3.3.6) whereby the output was a set of n predicted values of MC at the data locations, $\{z_{MC}^*((\mathbf{u}, t)_\alpha), \alpha = 1,2,\ldots, n\}$, that could be compared to the MC data themselves, $\{z_{MC}((\mathbf{u}, t)_\alpha), \alpha=1,2,\ldots, n\}$, at the same locations in order to assess the predictive accuracy of each model. For Model 1, cross-validation was applied as described to create the cross-validation set $z^*_{MC}((\mathbf{u}, t)_\alpha)$ to compare to $z_{MC}((\mathbf{u}, t)_\alpha)$. For Model 2, cross-validation sets were required for both MP and TC to define the cross-validation set $z^*_{MC}((\mathbf{u}, t)_\alpha) = z^*_{MP}((\mathbf{u}, t)_\alpha) \times z^*_{TC}((\mathbf{u}, t)_\alpha)$. For Model 3, the cross-validation procedure could not be based entirely on predictions made at data locations since the MMTC variable, by definition, required predictions at unsampled facility-months (5.1). As such, MMTC was calculated using both the available TC data $z_{TC}((\mathbf{u}, t)_\alpha)$ and predictions of TC at unsampled facility-months $z^*_{TC}((\mathbf{u}, t)_\beta)$, and a cross-validation set for MC was predicted using the resulting $z^*_{MMTC}(\mathbf{u}_k)$ values in the forward and back-transform between MC and SMC such that $z^*_{MC}((\mathbf{u},t)_\alpha) = z^*_{SMC}((\mathbf{u}, t)_\alpha) \times z^*_{MMTC}(\mathbf{u}_k)$, where $\mathbf{u}_k$ has the same spatial location as $(\mathbf{u}, t)_\alpha$ and $(\mathbf{u}, t)_\beta$.

For each cross-validation set defined above for the three modelling frameworks, the three summary statistics presented previously (ρ, MAE, and ME) were calculated to compare prediction performance.

8.3 Methodology 2: Developing an uncertainty model

8.3.1 Background: Stochastic simulation to estimate space-time regional uncertainty

In addition to providing a modelling framework to predict MC at unsampled facility-months, a

further aim of this study was to provide a measure of the uncertainty of these predictions both individually and over aggregated sets of predictions within different space-time regions. Comparison of the cross-validation statistics for the three modelling frameworks indicated that Model 3 produced the most accurate predictions of MC. This important result is presented and discussed in full later in this chapter. It is necessary to state this outcome here, however, because Model 3 was therefore chosen as the framework for which to develop an accompanying uncertainty model.

Space-time kriging procedures allow predictions to be made at a set of q unsampled space-time locations, $\{z^*((\mathbf{u}, t)_\beta), \beta = 1,2,\ldots, q\}$ over a spatiotemporal study region. In the current case, the quantity of interest is the mean or sum of values at a set of space-time locations within the study area of which some are sampled and some are unsampled (e.g. the sum or mean of MC over all facilities in a district over a year). In such cases, the relevant set of data can be combined with the relevant set of predictions and the joint mean or sum can be calculated. Whilst the contribution to this aggregated value from the original data has zero prediction uncertainty associated with it, the contribution from the predicted values has an associated prediction uncertainty that must be evaluated. Stochastic simulation is now presented in brief as a tool for estimating the uncertainty associated with the mean of a set of predictions. The procedure is equally applicable, with rudimentary adjustments, to estimating the uncertainty associated with the sum of a set of predictions.

8.3.1.1 Estimating the joint uncertainty of a set of predictions

If a set of predictions is made at q unsampled space-time locations, $\{z^*((\mathbf{u}, t)_\beta), \beta = 1,2,\ldots,q\}$, over the study region, the value of interest may be the mean $\mu[z^*((\mathbf{u}, t)_\beta)]$ of the q predicted values over the entire region $\{\beta = 1,2,\ldots, q\}$, or of a subset of v points within a sub-region, $\{\beta = 1,2,\ldots, v\}$. In addition to calculating these predicted regional means, it is necessary to provide estimates of the associated uncertainty. Although kriging systems provide 'optimum' local predictions by minimising the variance of the error of each prediction, a set of kriging predictions appears 'smoother' than the original data due to a missing error component. Conceptually, the RF $Z(\mathbf{u}, t)$ can be decomposed into the predictor $Z^*((\mathbf{u}, t)_\beta)$, as provided by kriging, and the corresponding unknown prediction error $R((\mathbf{u}, t)_\beta)$: $Z((\mathbf{u}, t)_\beta) = Z^*((\mathbf{u}, t)_\beta) + R((\mathbf{u}, t)_\beta)$. Estimates of the uncertainty associated with predictions of regional or global means

must take into account the variance introduced by this unknown error component in order to restore the full variance of the RF model.

One approach to the above problem is to simulate, for each of the β = 1,2,...,q prediction locations, l = 1,2,..., L realisations $\varepsilon^{(l)}((\mathbf{u}, t)_\beta)$ of the error component with zero mean and the correct variance and covariance which can then be added to the original prediction, $z^*((\mathbf{u}, t)_\beta)$, to give a conditional simulated prediction, $z^{(l)}((\mathbf{u}, t)_\beta)$ (Deutsch and Journel, 1998, p. 127):

$$z^{(l)}((\mathbf{u}, t)_\beta) = z^*((\mathbf{u}, t)_\beta) + \varepsilon^{(l)}((\mathbf{u}, t)_\beta) \tag{8.1}$$

If $z^{(l)}((\mathbf{u}, t)_\beta)$ is to have the same variance as the true value $z((\mathbf{u}, t)_\beta)$ then this approach requires that the error component is orthogonal to the predictor and has at least the same covariance, if not spatial distribution, as the actual error. A procedure to generate realisations of the error component under these conditions was proposed originally by Journel and Huijbregts (1978, p. 495) for a spatial-only setting and is presented here adapted to a space-time setting. l = 1,2,...,L non-conditional realisations $z_{nc}^{(l)}((\mathbf{u}, t)_v)$ that share the same covariance as the RF $Z(\mathbf{u}, t)$ are simulated at all data and prediction locations v=1,2,...,n+q. The original kriging exercise performed on the data is then repeated using the simulated values at the n data locations $\{z_{nc}^{(l)}((\mathbf{u}, t)_\alpha)$, α = 1,2,...,$n\}$, rather than the data, to obtain simulated predictions at the q unsampled locations $\{z^{*(l)}((\mathbf{u}, t)_\beta)$, β = 1,2,...,$q\}$ to compare to the simulated values at these locations $\{z_{nc}^{(l)}((\mathbf{u}, t)_\beta)$, β = 1,2,...,$q\}$. Simulated errors $\varepsilon^{(l)}((\mathbf{u}, t)_\beta)$ are then defined for each prediction location as the difference between simulated values and simulated predictions, $\varepsilon^{(l)}((\mathbf{u}, t)_\beta) = z^{*(l)}((\mathbf{u}, t)_\beta) - z_{nc}^{(l)}((\mathbf{u}, t)_\beta)$, and these can be added to the original predictions $z^*((\mathbf{u}, t)_\beta)$ to give conditional simulated predictions, $z^{(l)}((\mathbf{u}, t)_\beta)$:

$$z^{(l)}((\mathbf{u}, t)_\beta) = z^*((\mathbf{u}, t)_\beta) + [z^{*(l)}((\mathbf{u}, t)_\beta) - z_{nc}^{(l)}((\mathbf{u}, t)_\beta)] \tag{8.2}$$

The distribution of the set of L realisations $\{z^{(1)}((\mathbf{u}, t)_\beta), z^{(2)}((\mathbf{u}, t)_\beta), \ldots, z^{(L)}((\mathbf{u}, t)_\beta)\}$ at each prediction location represents the uncertainty of that prediction which can be summarised by the standard deviation of the L realisations, $\sigma_{sim}[z^*((\mathbf{u}, t)_\beta)] = \sigma[z^{(l)}((\mathbf{u}, t)_\beta)]$, l = 1,2,...,L. Where the value of interest is the mean, $\mu[z^*((\mathbf{u}, t)_\beta)]$, of a set of β = 1,2,...,q predicted values within a region, simulated realisations of the mean, $\mu[z^{(l)}((\mathbf{u}, t)_\beta)]$, can also be defined:

$$\mu[z^{(l)}((\mathbf{u},t)_\beta)] = \frac{1}{q}\sum_{\beta=1}^{q} z^{(l)}((\mathbf{u},t)_\beta) \tag{8.3}$$

and the distribution of the set of these L realisations $\{\mu[z^{(1)}((\mathbf{u},t)_\beta)], \mu[z^{(2)}((\mathbf{u},t)_\beta)], \ldots, \mu[z^{(L)}((\mathbf{u},t)_\beta)]\}$ represents the uncertainty of the predicted mean, $\mu[z^*((\mathbf{u},t)_\beta)]$. Again, this uncertainty can be summarised by the standard deviation of the L realisations, $\sigma_{\text{sim}}[\mu[z^*((\mathbf{u},t)_\beta)]] = \sigma[\mu[z^{(l)}((\mathbf{u},t)_\beta)]]$, $l = 1,2,\ldots, L$.

8.3.1.2 Sequential Gaussian simulation

The remaining issue is the choice of simulation algorithm to generate the L non-conditional simulated realisations of the RF $Z(\mathbf{u}, t)$. Sequential Gaussian simulation (sGs) is one such algorithm that creates realisations under the assumption of a multiGaussian RF model and is presented in Goovaerts (1997, pp. 380-393). The space-time equivalent, ST-sGs can be described in brief as follows. The set of n z-data $\{z((\mathbf{u}, t)_\alpha), \alpha = 1,2,\ldots, n\}$ are first transformed into a corresponding set of y-data, $y((\mathbf{u}, t)_\alpha) = \phi(z((\mathbf{u}, t)_\alpha))$, with a standard Gaussian cdf where ϕ is the normal-score transform (Goovaerts 1997, p. 266). Under the multiGaussian assumption, the ccdf at each prediction location is Gaussian and, therefore, fully characterised by its mean and variance. The sGs algorithm proceeds by visiting sequentially all v data and prediction locations, $v=1,2,\ldots,n+q$, and determining the mean and variance of each ccdf as the predicted value, $y^*_{\text{SK}}((\mathbf{u}, t)_v)$, and prediction variance, $\sigma^2_{\text{SK}}((\mathbf{u},t)_v)$, respectively, of a space-time simple kriging (STSK) prediction carried out for that location with the normal score space-time variogram model, $\tilde{\gamma}_Y(\mathbf{h}_s, h_t)$, fitted to the sample space-time variogram of the y-data $y((\mathbf{u}, t)_\alpha)$. A simulated value, $y^{(l)}((\mathbf{u}, t)_v)$, is then drawn from the ccdf for the location in question. In the non-conditional case, each subsequent prediction is conditioned only on values simulated at previously visited locations, and not on the data $y((\mathbf{u}, t)_\alpha)$. Once values have been simulated for all q locations, set $y^{(l)}((\mathbf{u}, t)_v)$ is back-transformed into the desired z-data space using the inverse normal-score transform $z^{(l)}((\mathbf{u}, t)_v) = \phi^{-1}(y_c^{(l)}((\mathbf{u}, t)_v))$.

8.3.2 Implementation of ST-sGs to estimate prediction uncertainty

The theoretical approach set out in the previous section was implemented to obtain estimates of the uncertainty associated with predictions of MC made using Model 3. This required two new developments: the conversion of techniques for spatial-only simulation to the space-time setting, and the integration of this space-time version in an uncertainty model that incorporated the different sources of prediction uncertainty within the framework of Model 3. The first requirement was met by modifying an existing algorithm for spatial-only sGs, the *sgsim* GSLIB routine (Deutsch and Journel, 1998), for a space-time setting. This entailed the replacement of sub-routines that incorporated spatial variograms and calculated covariances between spatially-separated locations with space-time equivalents. Provision was made for use of the product-sum covariance model by incorporating the modified *cova3* sub-routine presented by De Cesare et al. (2002) in the *sgsim* algorithm.

The uncertainty model was designed to replicate the prediction uncertainty inherent in Model 3. To briefly restate the structure of Model 3 (see Figure 5.5), TC was predicted first at unsampled locations and these predictions were combined with existing data to derive the mean TC per month, MMTC, at each facility. MC data from each facility were then divided by the relevant MMTC value to create standardised SMC values. SMC was then predicted at all unsampled locations and back-transformed to predictions of MC using the relevant MMTC value. The prediction procedures for both TC and SMC introduce uncertainty into the final predictions of MC, and so the uncertainty model had to incorporate the effects of both.

To construct the uncertainty model, the two STOK procedures in Model 3 that predicted TC and SMC were replaced with ST-sGs procedures. Because the objective was to develop and then evaluate the uncertainty model, it was necessary to implement it in a cross-validation mode such that each uncertainty estimate could be compared to a known prediction error derived from the cross-validation carried out in Chapter 8. This was carried out by first using ST-sGs to simulate l conditional realisations of TC at the β = 1,2,..., q unsampled locations, $z^{(l)}{}_{TC}((\mathbf{u}, t)_\beta)$. The constituent steps of this procedure are detailed in full in Figure 8.1. These realisations were then combined with the α = 1,2,..., n TC data $z_{TC}((\mathbf{u}, t)_\alpha)$ to create l realisations of MMTC. At each facility location $\mathbf{u}_k$, the d = 1,2,...,D data and s = 1,2,..., S simulated values were combined for each realisation to define the simulated MMTC value $z^{(l)}{}_{MMTC}(\mathbf{u}_k)$:

$$z^{(l)}_{\mathrm{MMTC}}(\mathbf{u}_k) = \frac{1}{D+P}\left[\sum_{d=1}^{D} z_{\mathrm{TC}}((\mathbf{u}_k,t)_d) + \sum_{p=1}^{P} z^{(l)}_{\mathrm{TC}}((\mathbf{u}_k,t)_p)\right] \tag{8.4}$$

The MC data $z_{\mathrm{MC}}((\mathbf{u}, t)_\alpha)$ were then divided by the simulated MMTC value for the facility in question $z^{(l)}_{\mathrm{MMTC}}(\mathbf{u}_k)$ to define l realisations of SMC at the data locations, $z^{(l)}_{\mathrm{SMC}}((\mathbf{u}, t)_\alpha)$:

$$z^{(l)}_{\mathrm{SMC}}((\mathbf{u},t)_\alpha) = \frac{z_{\mathrm{MC}}((\mathbf{u},t)_\alpha)}{z^{(l)}_{\mathrm{MMTC}}(\mathbf{u}_k)} \tag{8.5}$$

The simulated values $z^{(l)}_{\mathrm{SMC}}((\mathbf{u}, t)_\alpha)$ of SMC at the n data locations were then used as input into the second ST-sGs procedure that corresponded to the STOK prediction of SMC. Because cross-validation simulations were required, this ST-sGs procedure was used to obtain realisations of SMC at the n data locations. These output simulations of SMC are denoted with a (ll) superscript, $z^{(ll)}_{\mathrm{SMC}}((\mathbf{u}, t)_\alpha)$, to distinguish them from the input simulations, $z^{(l)}_{\mathrm{SMC}}((\mathbf{u}, t)_\alpha)$. This procedure is detailed in full in Figure 8.2. These output simulations of SMC were then back-transformed using the appropriate simulated MMTC value as defined in (8.4) to obtain the final simulated MC values, $z^{(l)}_{\mathrm{MC}}((\mathbf{u}, t)_\alpha)$:

$$z^{(l)}_{\mathrm{MC}}((\mathbf{u},t)_\alpha) = z^{(ll)}_{\mathrm{SMC}}((\mathbf{u},t)_\alpha) \times z^{(l)}_{\mathrm{MMTC}}(\mathbf{u}_k) \tag{8.6}$$

The above procedure resulted in a set of $l = 1,2,\ldots, L$ conditional realisations of MC, $z^{(l)}_{\mathrm{MC}}((\mathbf{u}, t)_\alpha)$, at the $\alpha = 1,2,\ldots, n$ data locations. The ST-sGs algorithm required substantial computation and the number of realisations was therefore limited to $L = 100$. The distribution of these L simulated sets provided a model of the uncertainty associated with each prediction. This model could then be compared to the known prediction errors determined in the cross-validation for Model 3, allowing assessment of the accuracy of the uncertainty model itself.

8.3.3 Testing the accuracy of the uncertainty model

The L simulated sets were tested as a model for (i) local uncertainty, that is, of predictions of MC at individual facility-months, and (ii) regional uncertainty, that is, of predictions of the regional mean MC per facility-month over aggregated sets of cross-validation predictions within

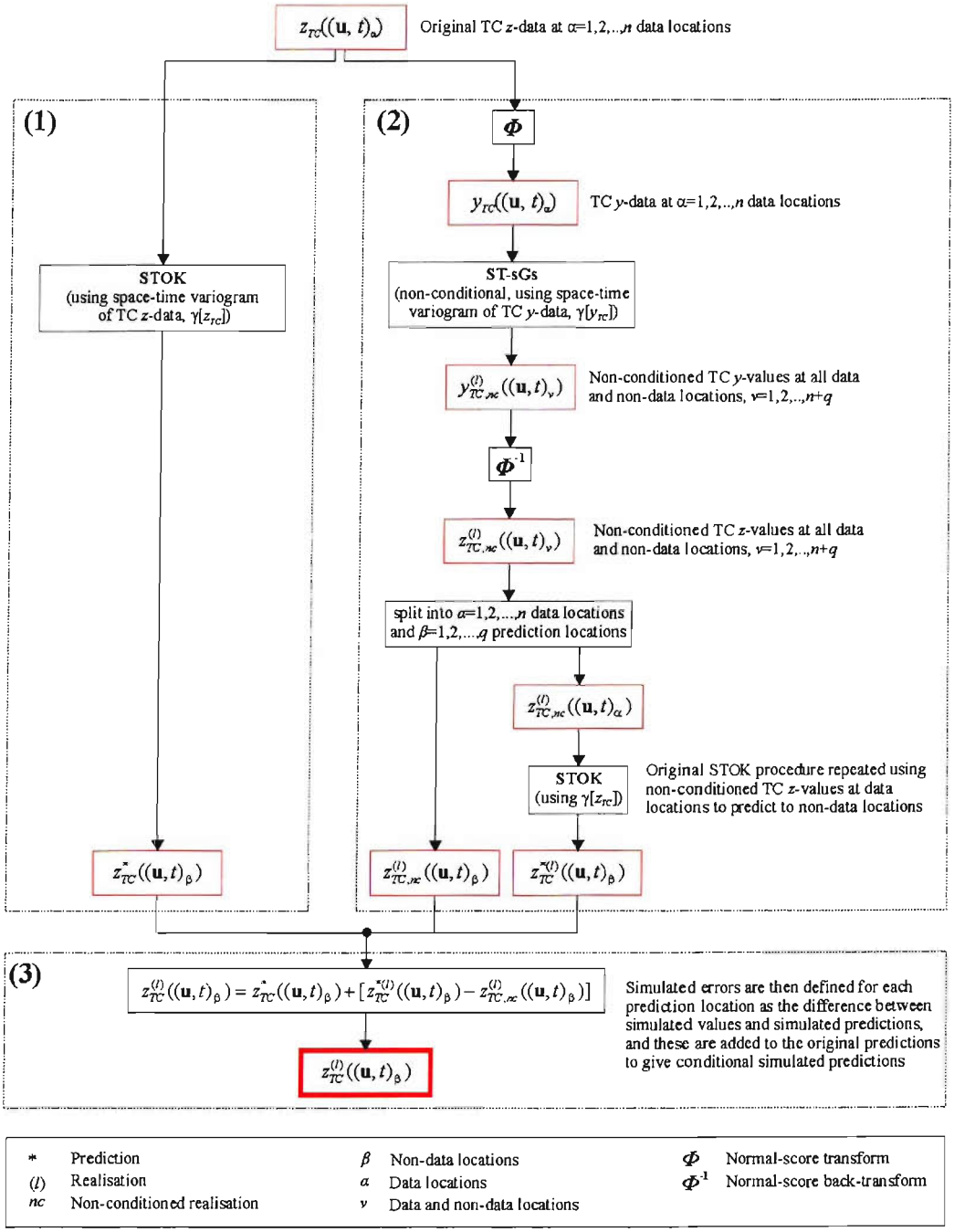


Figure 8.1 Schematic diagram showing steps involved in producing conditionally simulated realisations of TC at non-data locations. Non-conditioned fields were generated at all points using ST-sGs and these values at data locations were used for predictions at non-data locations using STOK. These simulated predictions were then compared to the original simulated values at each non-data location to generate simulated prediction errors (box 2). The original STOK prediction is shown (box 1), and the simulated errors were added to these predictions to give conditional simulated realisations (box 3). The notation used in this figure is described further in the text.

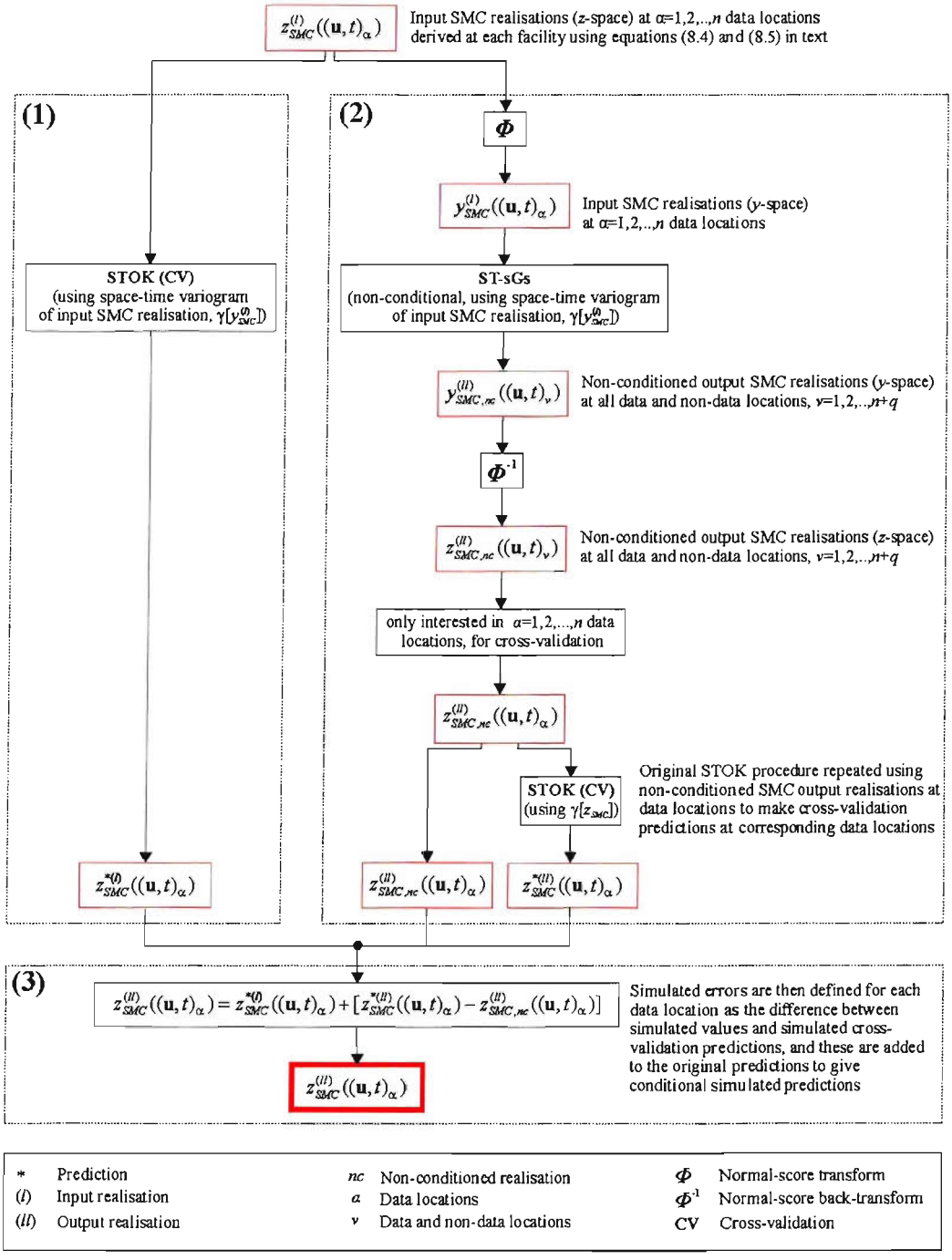


Figure 8.2 Schematic diagram showing steps involved in producing conditionally simulated cross-validation realisations of SMC at the data locations. Non-conditioned fields were generated at all points using ST-sGs and these values at the data locations were used for cross-validation predictions at the same data locations using STOK to obtain simulated cross-validation errors (box 2). The original STOK cross-validation prediction is shown (box 1), and the simulated errors were added to these predictions to give conditional simulated cross-validation realisations (box 3).

space-time regions. Each local uncertainty model was summarised by the simulated error standard deviation, $\sigma_{sim}[\varepsilon((\mathbf{u}, t)_\alpha)]$:

$$\sigma_{sim}[\varepsilon((\mathbf{u}, t)_\alpha)] = \sigma[\varepsilon^{(l)}((\mathbf{u}, t)_\alpha)], \; l = 1,2,\ldots, L \tag{8.7}$$

where simulated errors $\varepsilon^{(l)}((\mathbf{u}, t)_\alpha)$ were defined as the difference between each conditional realisation and the corresponding original prediction, $\varepsilon^{(l)}((\mathbf{u}, t)_\alpha) = z_{MC}^{(l)}((\mathbf{u}, t)_\alpha) - z_{MC}^*((\mathbf{u}, t)_\alpha)$. It was then necessary to compare the simulated error standard deviations, $\sigma_{sim}[\varepsilon((\mathbf{u}, t)_\alpha)]$, to estimates of the corresponding actual error standard deviation, $\hat{\sigma}[\varepsilon((\mathbf{u}, t)_\alpha)]$, where actual error was defined as the difference between each cross-validation prediction and the true data value, $\varepsilon((\mathbf{u}, t)_\alpha) = z^*_{MC}((\mathbf{u}, t)_\alpha) - z_{MC}((\mathbf{u}, t)_\alpha)$. The set of n errors $\varepsilon((\mathbf{u}, t)_\alpha)$, $\alpha = 1,2,\ldots, n$, was partitioned into $b = 1,2,\ldots, B$ subsets or 'bins' according to the magnitude of their corresponding simulated error standard deviations, $\sigma_{sim}[\varepsilon((\mathbf{u}, t)_\alpha)]$. Each bin spanned $1/B^{th}$ of the range of values of $\sigma_{sim}[\varepsilon((\mathbf{u}, t)_\alpha)]$ and the B was chosen as 40. Each bin therefore contained a set $\varepsilon(j)$ of $j = 1,2,\ldots,J$ error values, each with a corresponding simulated error standard deviation value, $\sigma_{sim}[\varepsilon(j)]$. For each bin, the median of the J simulated error standard deviation values was compared to the estimated actual error standard deviation, $\hat{\sigma}_b[\varepsilon(j)]$. This pair of values was obtained for each of the B bins and plotted on a scatter plot to allow visual comparison.

A large number of regionally-aggregated sets of $\alpha = 1,2,\ldots,m$ prediction locations were defined using moving space-time windows with spatial radii of between 12.5 km and 100 km and temporal radii of between 3 and 24 months. The size of aggregated sets varied from $m = 2$ to $m = 1000$ individual predictions. For each set, the true regional MC mean, $\mu[z_{MC}((\mathbf{u},t)_\alpha)]$, and predicted mean, $\mu[z^*_{MC}((\mathbf{u},t)_\alpha)]$ were calculated from the data and cross-validation predictions, respectively, and the model of prediction uncertainty was defined by the distribution of the corresponding means of the $l = 1,2,\ldots,L$ simulated realisations of the m predictions, $\mu[z^{(l)}_{MC}((\mathbf{u},t)_\alpha)]$. Each regional model of uncertainty was summarised by the simulated mean error standard deviation, $\sigma_{sim}[\mu[\varepsilon((\mathbf{u},t)_\alpha)]]$:

$$\sigma_{sim}[\mu[\varepsilon((\mathbf{u},t)_\alpha)]] = \sigma[\mu[\varepsilon^{(l)}((\mathbf{u},t)_\alpha)]], \; l = 1,2,\ldots,L \tag{8.8}$$

where each simulated mean error, $\mu[\varepsilon^{(l)}((\mathbf{u},t)_\alpha)]$, was defined as the difference between the simulated mean, $\mu[z^{(l)}_{MC}((\mathbf{u},t)_\alpha)]$, and the corresponding predicted mean, $\mu[z^*_{MC}((\mathbf{u},t)_\alpha)]$:

$$\mu[\varepsilon^{(l)}((\mathbf{u},t)_\alpha)] = \mu[z_{\mathrm{MC}}^{(l)}((\mathbf{u},t)_\alpha)] - \mu[z_{\mathrm{MC}}^{*}((\mathbf{u},t)_\alpha)] \tag{8.9}$$

The large set of regional simulated error standard deviations for different aggregated sets were compared to estimates of the actual error standard deviation using the same 'binning' approach described above for the local case, resulting in a corresponding scatter plot for the regional case.

8.4 Results

8.4.1 Variography

Figure 8.3 shows the sample space-time variogram surface, the sample space- and time-marginal variograms and corresponding fitted models, and the resulting product-sum space-time variogram model for each variable that underwent STOK (TC, MC, MP, and SMC) and for each facility category (hospitals, health centres, dispensaries, and all three combined). Table 8.1 lists all the corresponding variogram model parameters. In all cases, the temporal variograms differ substantially in structure to the corresponding spatial variograms, with temporal variograms generally having smaller relative nugget effects and smaller sills. Most temporal variograms were modelled with a periodic component to account for a pseudo-periodic structure. The spatial variogram for combined facilities was modelled as a pure nugget effect for both the non-standardised variables (MC and TC), indicating a complete absence of spatial autocorrelation in these variables. In both cases spatial variograms revealed substantially more structure when hospitals, health centres, and dispensaries were considered separately. Comparison between MC and TC indicated that MC was the more spatially structured variable, with MC spatial variograms generally having lower relative nugget effects and larger range values than those for TC.

Variograms for the two standardised variables, MP and SMC, had a number of noticeably different characteristics from those for the non-standardised variables. Firstly, the spatial variograms for combined facilities both indicated spatial autocorrelation and were modelled with structured components in contrast to the pure nugget effect models used in the equivalent variograms for MC and TC. Secondly, the spatial variograms for all facility categories indicated a greater degree of spatial structure than the non-standardised variables, with generally smaller

Facility type		MC variogram			TC variogram			MP variogram			SMC variogram		
		Model type	a	c	Model type	a	c	Model type	a	c	Model type	a	c
All types	Space-marginal variogram	nugget	-	115000	nugget	-	1140000	nugget	-	0.0075	nugget	-	0.014
								exponential	30	0.0016	exponential	5	0.016
								spherical	85	0.0084	exponential	30	0.007
											spherical	80	0.021
	Time-marginal variogram	nugget	-	24000	nugget	-	140000	nugget	-	0.003	nugget	-	0.014
		spherical	40	15000	spherical	40	150000	exponential	3.3	0.0035	exponential	4	0.028
		exponential	10	25000	exponential	10	40000	hole	6	0.0003	hole	6	0.002
		hole	6	6000	hole	6	50000	spherical	20	0.002	exponential	25	0.012
	Space-time sill	-	-	125000	-	-	1200000	-	-	0.0178	-	-	0.06
Hospitals	Space-marginal variogram	nugget	-	50000	nugget	-	0	nugget	-	0.0055	nugget	-	0.012
		exponential	30	350000	spherical	20	2200000	exponential	30	0.0016	exponential	20	0.036
								spherical	85	0.0084	exponential	80	0.016
	Time-marginal variogram	nugget	-	50000	nugget	-	800000	nugget	-	0.0025	nugget	-	0.012
		spherical	40	160000	exponential	5	400000	exponential	4	0.004	exponential	5	0.038
		exponential	6	180000	spherical	45	1200000	hole	6	0.0003	hole	6	0.001
		hole	6	20000	hole	6	400000	spherical	20	0.002	spherical	25	0.006
	Space-time sill	-	-	420000	-	-	2300000	-	-	0.0165	-	-	0.066
Health Centres	Space-marginal variogram	nugget	-	16000	nugget	-	80000	nugget	-	0.0065	nugget	-	0.01
		exponential	20	30000	exponential	3	210000	exponential	30	0.0017	exponential	8	0.02
		exponential	70	49000	exponential	40	100000	spherical	85	0.008	spherical	70	0.038
	Time-marginal variogram	nugget	-	16000	nugget	-	80000	nugget	-	0.003	nugget	-	0.01
		exponential	6	39000	exponential	4	6000	exponential	4	0.004	exponential	4	0.04
					exponential	20	95000	hole	6	0.0003	spherical	25	0.006
					hole	5.5	17000	spherical	20	0.002			
	Space-time sill	-	-	100000	-	-	400000	-	-	0.0165	-	-	0.067
Dispensaries	Space-marginal variogram	nugget	-	22000	nugget	-	175000	nugget	-	0.004	nugget	-	0.007
		exponential	90	29000	spherical	75	90000	exponential	8	0.0025	exponential	9	0.019
								exponential	60	0.005	spherical	80	0.024
								spherical	85	0.0069			
	Time-marginal variogram	nugget	-	7500	nugget	-	50000	nugget	-	0.003	nugget	-	0.007
		exponential	5	25000	exponential	5	40000	exponential	3.5	0.0033	exponential	4	0.026
		exponential	25	4000	hole	6	10000	hole	6	0.0003	hole	6	0.0025
					exponential	20	50000	spherical	20	0.002	exponential	25	0.011
	Space-time sill	-	-	55000	-	-	280000	-	-	0.0188	-	-	0.055

Table 8.1 Space-time variogram model parameters for MC,TC,MP, and SMC for each facility type. Values of the range parameter (denoted a) are given in kilometres for spatial model components and months for temporal model components. c denotes the sill parameter of each model component.

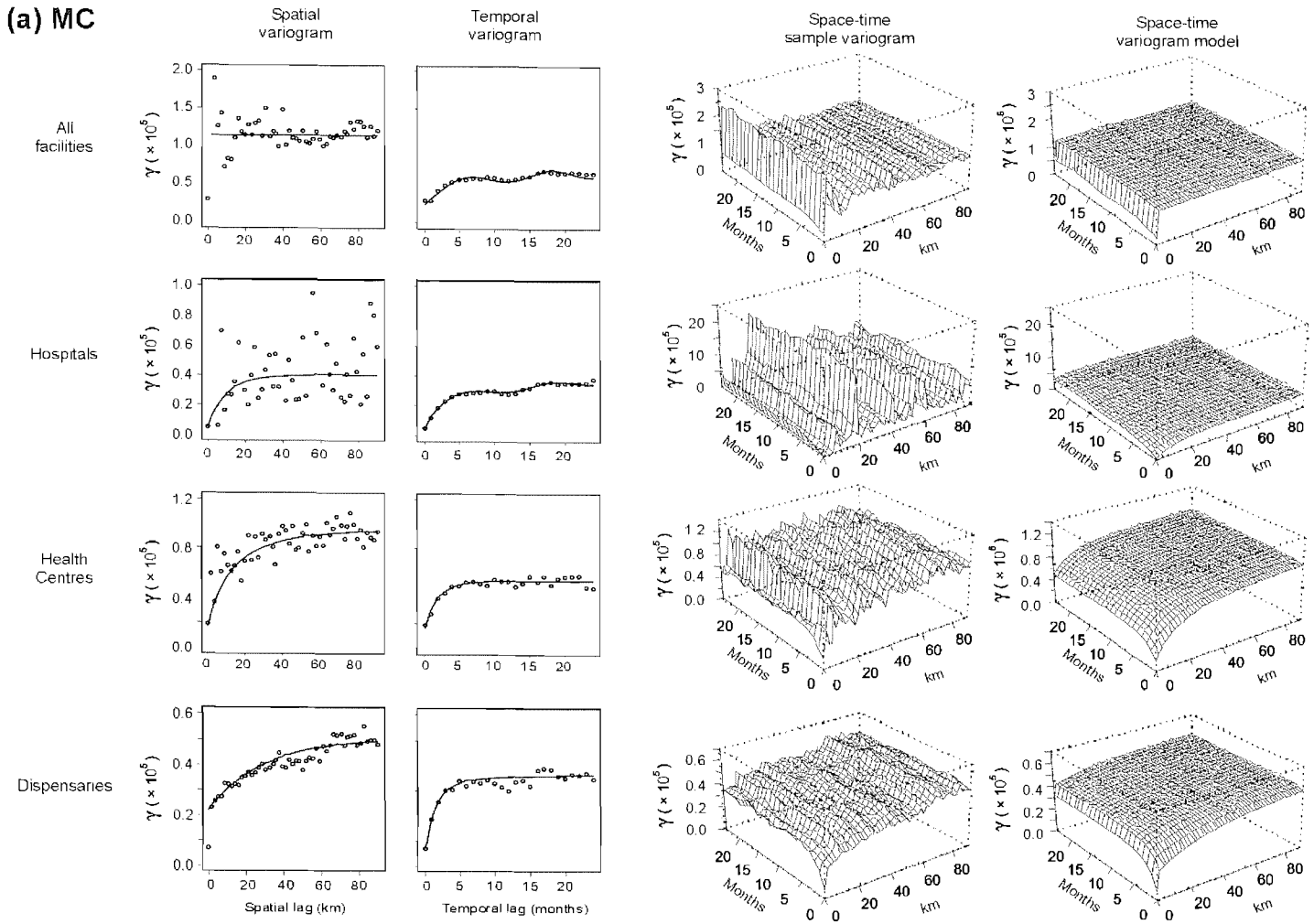


Figure 8.3 Variography of the four predicted variables MC, TC, MP, and SMC. (a) Shown on this page are the plots for MC. Column 1 (far left) shows the sample space-marginal variogram (circles) and model (line). Column 2 shows the sample time-marginal variogram (circles) and model (line). Column 3 shows the sample space-time variogram surface, and column 4 (far right) shows the space-time product-sum variogram model. The rows correspond to all facilities combined (top) and to hospitals, health centres, and dispensaries considered separately.

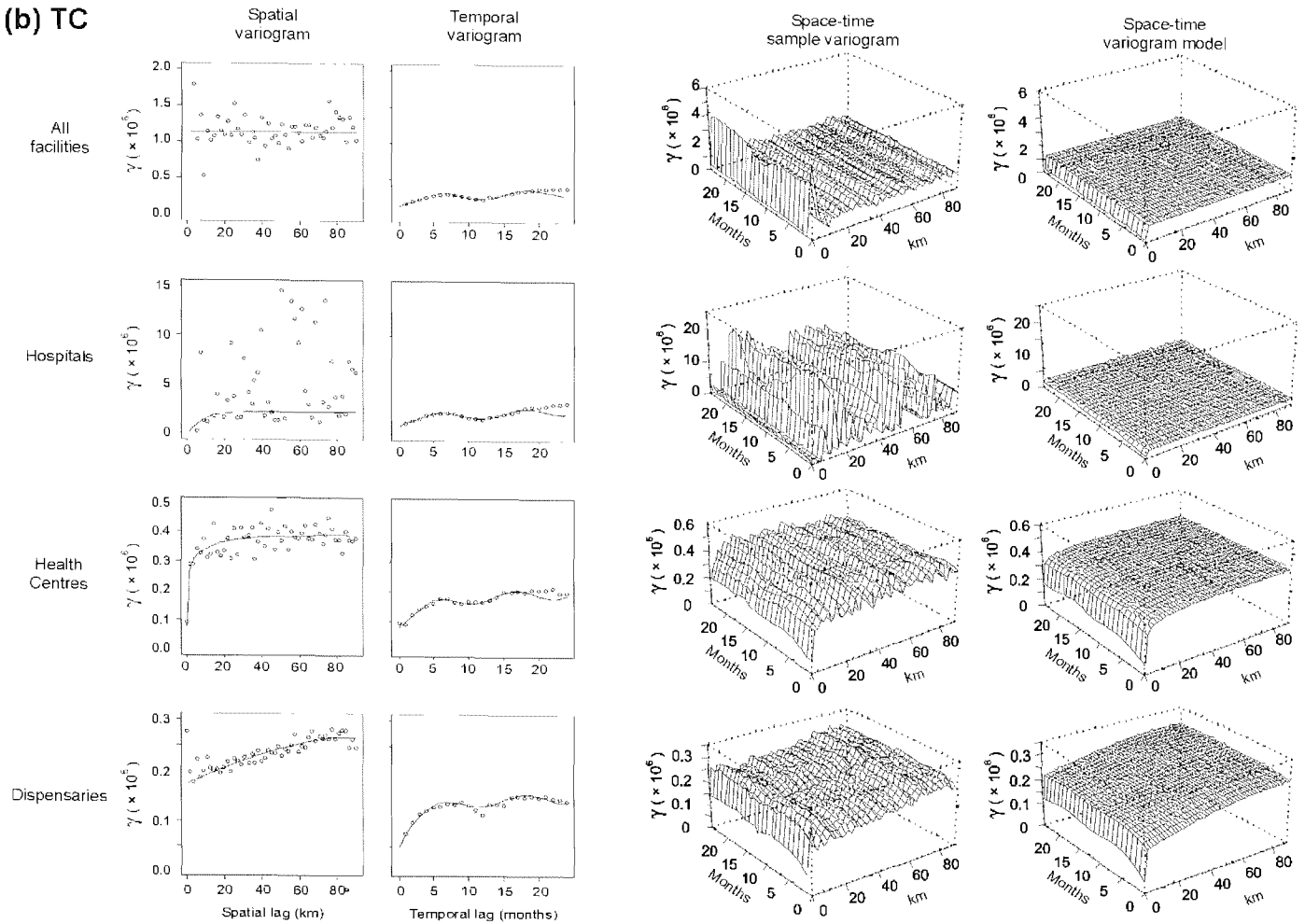


Figure 8.3 (b) Variography for TC

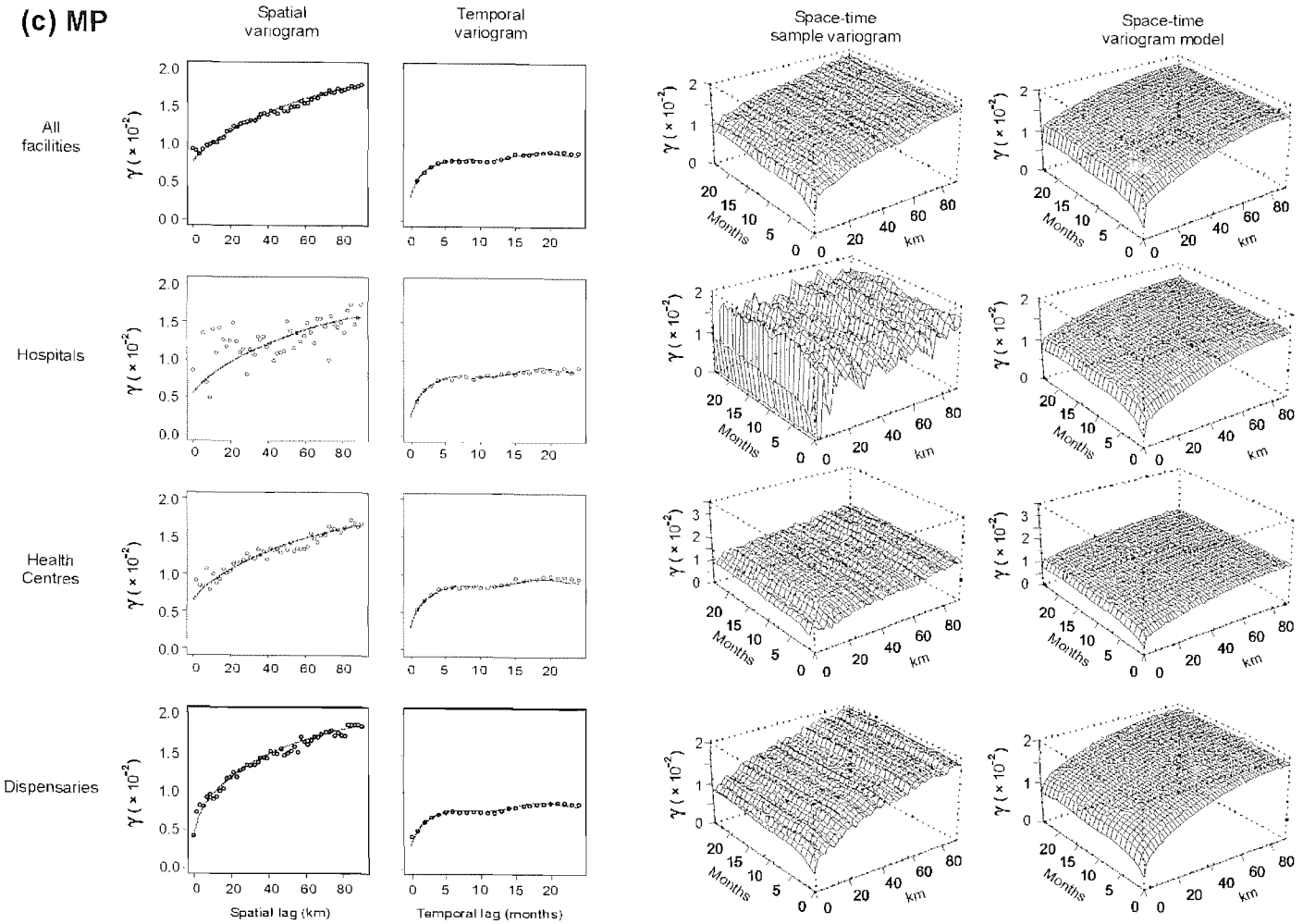


Figure 8.3 (c) Variography for MP

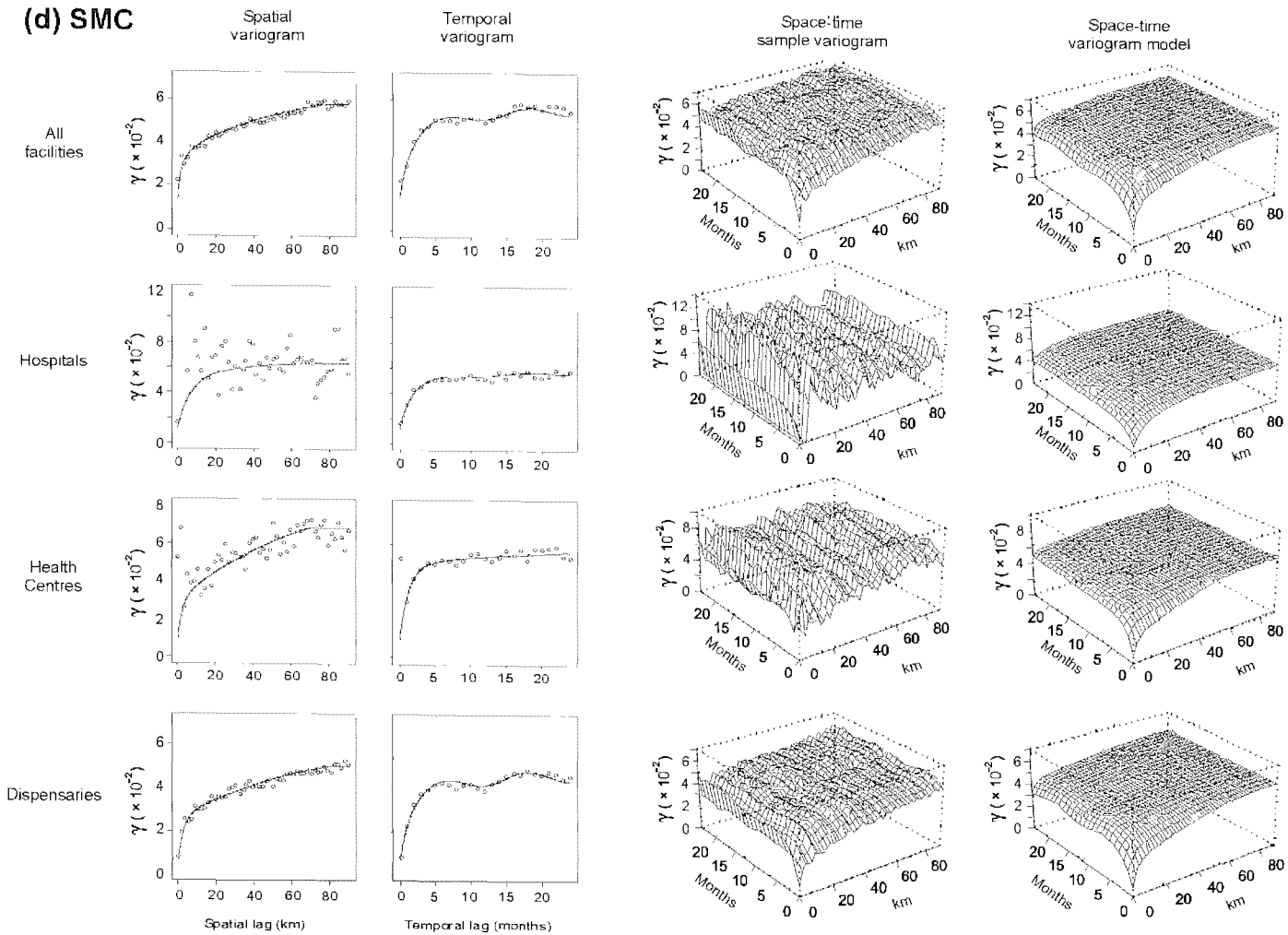


Figure 8.3 (d) Variography for SMC

relative nugget effects and structured components with larger range values. Furthermore, the sill values of spatial variograms were more similar to the sill values of the temporal variograms for MP and SMC than was the case for MC and TC. Comparison between MP and SMC variograms suggests that SMC displays marginally more structure, both spatially and temporally than does MP.

8.4.2 Comparison of modelling frameworks

The results of the cross-validation for each of the three modelling frameworks are shown in Table 8.2. Comparison between those predictions made using the combined data set for all facility types and those made with the facilities separated revealed that the latter approach resulted in substantially smaller overall bias (smaller ME), substantially smaller mean inaccuracy (smaller MAE), or both, for all facility types and for all three models. Focusing on this latter approach, comparison of the three models revealed that Model 3 produced predictions of MC which had the smallest mean inaccuracy (smallest MAE) for all three facility classes. Model 2 predictions had a smaller MAE than Model 1 for health centres and dispensaries but a larger MAE for hospitals. Results for overall bias (ME) were more mixed. The least biased prediction (smallest ME) was provided by

Table 8.2 Comparison of summary statistics for cross-validation predictions of malaria cases using three different modelling frameworks. Predictions were made separately for hospitals, health centres and dispensaries, and with all facilities combined. The statistics shown are the correlation coefficient, ρ, the mean error (ME) and mean absolute error (MAE), as described in the text. Model 3 (highlighted in bold text) was chosen as the best overall predictor of malaria cases.

Facility type	Model	Facilities separated?	ρ	ME	MAE
Hospitals	Model 1	Yes	0.859	4.439	193.188
	Model 2	Yes	0.848	6.244	205.730
	Model 3	Yes	**0.856**	**2.822**	**192.423**
	Model 1	No	0.512	-376.730	425.829
	Model 2	No	0.853	-6.945	196.637
	Model 3	No	0.850	-24.276	196.736
Health centres	Model 1	Yes	0.779	0.416	92.067
	Model 2	Yes	0.783	-2.179	90.240
	Model 3	Yes	**0.789**	**-1.050**	**89.042**
	Model 1	No	0.526	-15.783	150.227
	Model 2	No	0.781	-3.614	90.416
	Model 3	No	0.793	-1.787	89.745
Dispensaries	Model 1	Yes	0.764	0.530	69.527
	Model 2	Yes	0.776	-0.397	67.156
	Model 3	Yes	**0.774**	**-0.638**	**66.903**
	Model 1	No	0.527	58.239	136.291
	Model 2	No	0.762	1.897	69.790
	Model 3	No	0.777	0.414	67.321

Model 1 for health centres, Model 2 for dispensaries, and Model 3 for hospitals. The largest values of ρ (largest linear associations between predicted and actual values) were provided by Model 1 for hospitals, Model 2 for dispensaries and Model 3 for health centres, although differences in values of ρ between the three models were not substantial. Given these results it was decided that Model 3 was the best overall choice of predictor for MC because it resulted in the smallest mean inaccuracy for all three facility classes and, although its predictions were not the least biased for health centres and dispensaries, the bias in these cases was nevertheless very small.

8.4.3 Evaluation of the uncertainty model

The results of the procedure to test the accuracy of the simulated uncertainty model are shown in Figure 8.4 for both local predictions of MC at individual facility-months and regional predictions of mean MC for sets of between 2 and 1000 facility-months aggregated over space-time neighbourhoods. In the local case (Figure 8.4 (a)), simulated error standard deviations replicated closely actual values with no overall tendency for

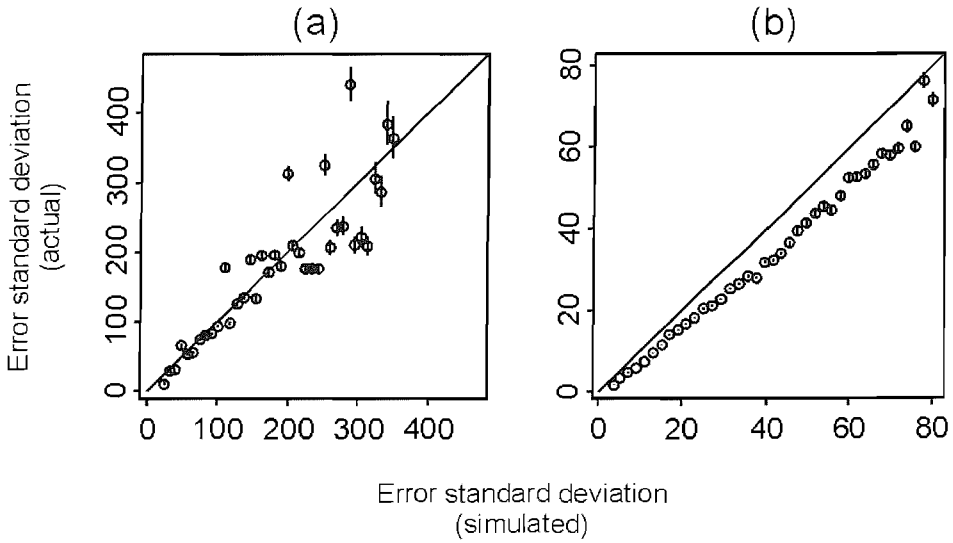


Figure 8.4 Comparison of simulated and actual standard deviations of prediction errors for MC. Simulated standard deviations were derived for individual and aggregated prediction of MC via space-time sequential-Gaussian-simulation and corresponding actual errors were obtained from a cross-validation exercise. Prediction errors were divided into bins according to their simulated standard deviation, and the actual standard deviation of the set of errors in each bin was calculated (circles) along with the 95% confidence interval (vertical bars). Results are shown for (a) predictions of MC at individual facility-months and (b) predictions of mean MC within sets of between 2 and 1000 facility-months created by aggregating points within progressively larger space-time neighbourhoods.

over or under-estimation. Points plotted for smaller standard deviations were progressively less scattered around the 1:1 line, which is indicative of the larger number of values in these bins. In the regional case (Figure 8.4 (b)), there was, again, a strong linear association between simulated and actual error standard deviations in each bin, although there was a tendency for simulated values to be slightly overestimated. This overestimation was more pronounced for larger standard deviations, although was less in relative terms. A simulated error standard deviation of 19.2 cases, for example, corresponded to an actual error standard deviation of 15.2 cases, representing an over-estimation of 4.0 cases or 26%, whilst a simulated error standard deviation of 71.7 cases corresponded to an actual value of 59.6 cases, representing an over-estimation of 12.1 cases or 17%.

8.5 Discussion

This study has presented three alternative modelling frameworks in which space-time geostatistical prediction algorithms can be used to predict MC values at missing facility-months within the Kenyan HMIS. Whilst Model 1 used these data in their raw form, Models 2 and 3 used accompanying data, TC, on total outpatient numbers to construct a denominator and predictions were made on the resulting standardised variables, MP and SMC, respectively. The rationale was that the spatial structure of the standardised variables may be greater than that of the raw count data thus yielding more accurate predictions from the geostatistical algorithms. Since the presence or absence of TC data matched that of MC, however, predictions of MP and SMC required back-transformation by corresponding predictions of the relevant denominator (TC and MMTC, respectively) at unsampled facility-months and, as such, the accuracy of the ultimate predictions of MC was dependent on the prediction accuracies of both the standardised variables and the denominator. Predictions made with all three modelling frameworks were found to be more accurate when hospitals, health centres and dispensaries were considered separately. Under this approach, Model 2 did not offer a substantial increase in predictive accuracy over Model 1, indicating that the large uncertainty associated with modelling TC negated any benefit of modelling a standardised variable. The modelling framework for Model 3, however, did result in modest increases in prediction accuracy over Model 1. The temporal variograms for MC

and SMC had almost identical structure which is to be expected since the denominator, MMTC, is constant through time at each spatial location. The benefit of standardising MC by MMTC to obtain SMC can be explained by the spatial variograms for SMC which had smaller relative nugget effects and sill values that were much nearer to the corresponding temporal sills than was the case for the MC variograms, indicating a relative reduction in the overall variance of the variable across space, of which a greater proportion was autocorrelated. These factors meant SMC could be predicted directly with greater accuracy than could MC. Although the back-transform by MMTC involved further uncertainty, the net effect was that MC was predicted with slightly greater accuracy under this framework than using raw MC data directly in Model 1. The greater spatial structure displayed by SMC emphasises the potential benefit of incorporating proxy measures of facility size and utilisation in models to predict MC. However, these results have shown that, when the only such measures available are themselves incomplete and subject to substantial uncertainty, their inclusion in a predictive model can offer only modest increases in prediction accuracy. The success of Model 3 over Model 2 can be attributed to the way that TC predictions were averaged, along with the existing TC data, over the 84-month period for each facility before being used as a denominator. The resulting MMTC values were, therefore, likely to be have a smaller error variance compared to the individual predictions of TC used as denominators in Model 2.

The fact that the use of standardised variables in Model 2 and Model 3 resulted in only modest increases in prediction accuracy over Model 1 is due partly to the effect of separating data by facility type. When data were predicted together, Model 2 and Model 3 produced dramatically more accurate predictions than Model 1. Much of this benefit of using standardised variables was negated, however, when data were separated by facility type because this separation effectively provided an alternative way of standardising the raw MC data. This effect is clear when comparing the spatial MC variogram for the three facility types combined, which indicates no spatial structure, and those for the facility types individually, in which spatial structure is clearly present. A logical explanation is that there is a degree of consistency in non-spatial factors such as catchment size and facility utilisation within each facility type, and by considering each type individually, this source of non-spatial variation is reduced. Although in this study the increase in prediction accuracy gained using Model 3 is modest compared to the

non-standardised approach in Model 1, in other situations in which the type of facility is not known and data cannot be separated by facility type, this increase will be much more pronounced.

The uncertainty model presented in this chapter provides a framework for estimating the uncertainty associated with predictions of MC made using Model 3. The results presented above indicate that this model provides accurate estimates of individual (local) prediction uncertainty. For aggregated (regional) predictions, the model marginally over-estimates uncertainty. Given that the over-estimation of prediction uncertainty is preferable to under-estimation, and that the difference between actual and modelled uncertainty is small, this model can be considered a useful means of estimating both local and regional prediction uncertainty.

8.6 Chapter summary

Three modelling frameworks were proposed in Chapter 5 that consisted of four different prediction exercises to predict MC, TC, MP, and SMC. In Chapter 7, different kriging methodologies were developed and compared and STOK was identified as the most appropriate technique for obtaining these predictions. In this chapter the three modelling frameworks were implemented using STOK. Variograms of the four variables revealed that the standardised MP and SMC variables, created by incorporating TC data as a denominator, displayed substantially more spatial structure than raw MC data. This suggests that TC provides useful information for the prediction of MC. When the different modelling frameworks were used to obtain predictions of MC, however, Model 2 did not result in more accurate predictions of MC than the null case, Model 1. Model 3 did result in more accurate predictions, although the improvement was modest, and so this modelling framework was chosen to predict MC values across Kenya in the final implementation in Chapter 9. Of equal importance to providing accurate predictions is the provision of accompanying estimates of the associated prediction uncertainty. In this chapter, an uncertainty model has been developed to produce such estimates for predictions of MC made with Model 3. This model incorporated a stochastic simulation approach using sequential-Gaussian-simulation that was adapted for a space-time situation. A framework was developed using this approach that simulated the uncertainty

associated with the prediction procedure of Model 3. In the next chapter, the modelling approach that has been developed and tested in the last two chapters is implemented to predict unsampled values of MC across Kenya.

Chapter 9

Model Implementation to Predict Malaria Treatment Burdens

Chapter 9

9. Model Implementation to Predict Malaria Treatment Burdens

9.1 Introduction

Having presented a series of modelling frameworks for the prediction of missing MC values in Chapter 5, and developed and evaluated different aspects of these frameworks in Chapters 7 and 8, the final framework (Model 3 implemented with STOK) is now implemented in this chapter to make predictions of MC at all GoK facilities across Kenya for which monthly values are missing during the 84-month study period January 1996 – December 2002. The model development and testing in Chapters 7 and 8 was carried out on an early version of the integrated HMIS-NHSD data set containing data from 1765 georeferenced GoK facilities. This version of the data set was described in detail in Chapter 3. The model implementation described in this chapter was carried out using an updated version that incorporates 2165 facilities. A summary of this updated data set is presented, along with a new assessment of the extent of under-reporting and missing records. In Chapter 8, a model-based approach was presented for estimating the uncertainty associated with predictions of MC made using Model 3. In this chapter, an empirical approach is used to validate the final model predictions and this is presented in full.

9.2 An updated version of the HMIS-NHSD data set

An updated version of the NHSD and associated HMIS data was compiled by the

KEMRI-University of Oxford-Wellcome Trust Collaborative Programme team and made available for this project in October 2005. This version of the NHSD represented the most comprehensive inventory available of health facilities in Kenya and incorporated 2165 GoK facilities, of which it has been possible to obtain georeferencing coordinates for 92%, consisting of 129 hospitals, 474 health centres and 1399 dispensaries (Table 9.1). A total of 163 facilities were included in this study that could not be georeferenced.

As before, the georeferencing and facility information in the NHSD was linked to the corresponding records in the HMIS to define spatially and temporally referenced MC and TC data. A total of 63,642 records were available in this updated data set, which represents only a modest increase from the 63,543 that constituted the earlier version, despite the fact that a further 400 facilities had been included. This apparent disparity is explained by the fact that these extra facilities were not included in the central HMIS database at the Ministry of Health headquarters in Nairobi. As such, there was no mechanism by which routine data from these facilities could be included in the national database. This substantial information gap serves to highlight the importance of obtaining a comprehensive inventory of facilities before attempting to quantify treatment

	Hospitals	Health Centres	Dispensaries	All
Number of facilities in upgraded MoH list				
Total	129	482	1,554	2,165
Georeferenced	129 (100.0%)	474 (98.3%)	1,399 (90.0%)	2,002 (92.5%)
Facility reporting rate (% of months reported)				
100%	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
>75% < 100%	19 (14.7%)	74 (15.4%)	154 (9.9%)	247 (11.4%)
> 50% ≤ 75%	31 (24.0%)	164 (34.0%)	322 (20.7%)	517 (23.9%)
> 25% ≤ 50%	45 (34.9%)	132 (27.4%)	299 (19.2%)	476 (22.0%)
> 0% ≤ 25%	24 (18.6%)	59 (12.2%)	296 (19.0%)	379 (17.5%)
0%	10 (7.8%)	53 (11.0%)	483 (31.1%)	546 (25.2%)
Overall reporting				
Records expected	10,836	40,488	130,536	181,860
Records present	4,680 (43.19%)	18,719 (46.23%)	40,243 (30.83%)	63,642 (35.00%)

Table 9.1 Summary of government health facilities in Kenya and their reporting behaviour during the 84-month study period January 1996 to December 2002. Facilities are shown disaggregated by type, georeferencing status and reporting rate. The expected and actual number of monthly records are also given for each facility type.

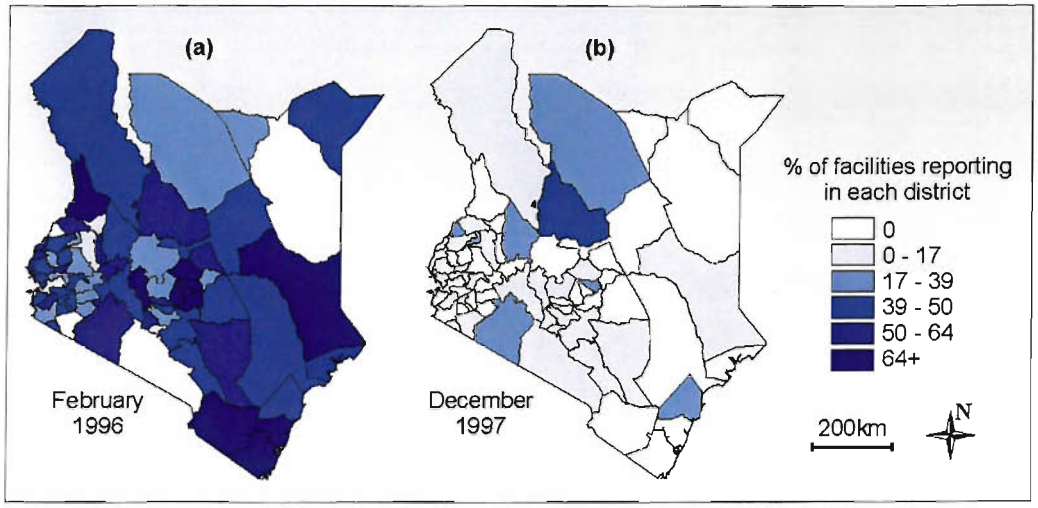


Figure 9.1 percentage of government health facilities in each Kenyan district submitting a monthly outpatient morbidity report to the HMIS. The two months shown are (a) the most complete (February 1996) and (b) the least complete (December 1997) during the 84-month study period January 1996 – December 2002.

burdens as, evidently, the first requirement is to know the number of facilities (and hence records) that are missing. In light of the substantial changes in the number of facilities, the extent of under-reporting was reassessed on this updated data set. There was considerable variation spatially and temporally (Figure 9.1) and between facility types (Table 9.1). No facilities reported in all 84 months whilst 546 facilities (25%) did not report in any month. A complete 84-month data set for each of the 2,165 facilities would consist of 181,860 facility-months. There were 63,642 records representing an overall reporting rate of 35%. The overall reporting rate varied both within and between years, with a minimum of 6% in December 1997 and a maximum of 44% in February 1996. The reporting rate displayed a seasonal pattern, with generally more facilities reporting during the first three quarters of each year (36%) than the last quarter (31%).

The 63,642 monthly records in this updated data set included a total of 18.67 million cases of presumed malaria, with a mean of 293.4 cases per facility-month. These totals (means) were 3.36 million (716.9) for hospitals, 6.05 million (323.4) for health centres and 9.26 million (230.2) for dispensaries.

9.3 Methodology 1: Implementation of Model 3 to predict MC

In this chapter, Model 3 was implemented to use the MC and TC data from the 63,642 facility-months in the HMIS data set to predict MC values at the 118,218 facility-months where records were missing. The modelling framework for Model 3 was described in Chapter 5 and its implementation was described in Chapter 8. The procedure used here was identical to that in the preceding chapter, except that the cross-validation procedure was replaced with a prediction procedure. To avoid repetition, the methodological details of this implementation are not restated in full in this chapter. In brief, data from each facility type were separated and predicted independently. For each facility type, the two prediction procedures to predict TC and then SMC were carried out using STOK. Space-time variograms of TC and SMC were estimated (4.45) and modelled (4.49) using the product-sum model (De Cesare et al., 2001; 2002). As would be expected given the small change in data sets, these variograms were almost identical to those presented in Chapter 8 (Figure 8.3) based on the test data set, and are not presented again in this chapter. For the 163 facilities that did not have georeferencing information, geostatistical techniques could not be applied to obtain predictions of missing values. The district in which each of these facilities was located was known, however, and each missing facility-month was predicted by attributing the mean MC value for that month from facilities of the corresponding type in the same district.

The above modelling procedure resulted in predictions of MC at all facilities and months where data were missing. In combination with the original data, this set represented a complete picture of the treatment burden for presumed malaria at all facilities for all months. The level of information that is of most use to decision-makers ranges spatially from the district to national levels, and temporally from monthly to annual averages or sums. Having constructed a complete set, individual MC values could be aggregated to provide treatment burdens at any spatial or temporal level from the individual facility through to the district, provincial and national levels for the seven year period, and for any month or year in the set.

9.4 Methodology 2: Model validation

A validation exercise was undertaken to assess the performance of the model in terms of the accuracy of predictions of MC. The magnitude of error was estimated at the level of individual predictions (i.e. prediction of MC at a single facility and month) and at different levels of spatial and temporal aggregation (e.g. predictions of the sum of MC for all facilities in a district or province in a given month or year).

9.4.1 Estimating parameters of the global error distribution

A validation set $\{z_{MC}((\mathbf{u},t)_i);\ i = 1,2,\ldots,n_v\}$, was selected randomly from the full set of MC data $\{z_{MC}((\mathbf{u},t)_\alpha);\ \alpha = 1,2,\ldots,n\}$. The size of this validation set, n_v, was chosen as 6349, equivalent to 10% of the full data set, and was selected using a stratified random sample that ensured the proportions of data from hospitals, health centres and dispensaries matched those of the full data set. The validation data were removed from the full data set and the modelling procedure was repeated in its entirety using the remaining 90% of data to produce a set of predictions $z^*_{\mathrm{MC}}((\mathbf{u},t)_i)$ to compare to the validation set. The set of prediction errors, $\varepsilon_v((\mathbf{u},t)_i)$, was then defined as $\varepsilon_v((\mathbf{u},t)_i) = z^*_{\mathrm{MC}}((\mathbf{u},t)_i) - z_{\mathrm{MC}}((\mathbf{u},t)_i)$, with the v subscript used to denote the validation set. The set $\varepsilon_v((\mathbf{u},t)_i)$ was treated as a sample of $\varepsilon_u((\mathbf{u},t)_\beta)$, the full set of (unknown) errors for predictions of missing data at the β=1,2,...,q unsampled facility-months, where the u subscript is used to denote the full set of unknown errors. The mean, μ_u, and standard deviation, σ_u, of $\varepsilon_u((\mathbf{u},t)_\beta)$ were then estimated using the sample mean $\overline{X}_v$ (9.1) and standard deviation s_v (9.2) calculated from $\varepsilon_v((\mathbf{u},t)_i)$:

$$\hat{\mu}_u = \overline{X}_v = \frac{1}{n_v}\sum_{i=1}^{n_v}\varepsilon_v((\mathbf{u},t)_i) \tag{9.1}$$

$$\hat{\sigma}_u = s_v = \sqrt{\frac{1}{n_v - 1}\sum_{i=1}^{n_v}\left(\varepsilon_v((\mathbf{u},t)_i) - \overline{X}_v\right)^2} \tag{9.2}$$

9.4.2 Assessing the effect of aggregation on the variance of prediction errors

Equations (9.1) and (9.2) provide a way of estimating the mean error in predictions of MC at individual facility-months and the variability around this mean. In addition to this individual-level error, however, it was necessary to obtain estimates of the error associated with the sum of MC obtained from sets of predictions aggregated over different space-time units such as months, years, districts, provinces and so on.

Consider a set of j predictions aggregated together within a space-time unit a, $\{z^*_{MC}((\mathbf{u},t)_j); j=1,2,\ldots, n_a\}$, which is a subset of the full set of predictions, $z^*_{MC}((\mathbf{u},t)_\beta)$. The corresponding subset of prediction errors are denoted as $\varepsilon_a((\mathbf{u},t)_j)$. The task was to estimate, for each such subset, the mean of the n_a errors, $\mu_a = \mu[\varepsilon_a((\mathbf{u},t)_j)]$, and standard deviation of this mean, $\sigma[\mu_a]$. If prediction errors are assumed to be independent and identically distributed (IID) then these values can be estimated from the estimated parameters of the global error distribution as shown in equations (9.3) and (9.4), respectively:

$$\hat{\mu}_a = \hat{\mu}_u \tag{9.3}$$

$$\hat{\sigma}[\hat{\mu}_a] = \frac{\hat{\sigma}_u}{\sqrt{n_a}} \tag{9.4}$$

The assumption of IID is rarely strictly valid when dealing with spatial and/or temporal data due to the presence of spatial and/or temporal autocorrelation. Before equations (9.3) and (9.4) could be used to estimate μ_a and $\sigma[\mu_a]$ for each space-time unit it was necessary to assess the validity of this assumption for the unknown error set $\varepsilon_u((\mathbf{u},t)_\beta)$ using the sample error set $\varepsilon_v((\mathbf{u},t)_i)$. A sample space-time variogram $\hat{\gamma}_v(\mathbf{h}_s, h_t)$ was calculated for $\varepsilon_v((\mathbf{u},t)_i)$ (Figure 9.3). This provided a graphical illustration of the presence or absence of spatial and temporal autocorrelation in the validation error set. A second approach was to estimate directly the relationship between the size of each subset n_a and the standard deviation of its mean error $\sigma[\mu_a]$ using the sample error set $\varepsilon_v((\mathbf{u},t)_i)$, and compare this empirical relationship with the theoretical relationship

presented in (9.4). This was done using the following steps:

(1) A total of k=1,2,...,m different aggregated subsets, $\varepsilon_{a,k}((\mathbf{u},t)_j)$,were created from the sample error set $\varepsilon_v((\mathbf{u},t)_i)$. Each subset consisted of all elements of $\varepsilon_v((\mathbf{u},t)_i)$ that fell within a given space-time unit. All permutations of space-time units were considered leading to, for example, 56 province-years (8 provinces × 7 years), 84 national-months (1 spatial unit × 84 months) and so on, and only those space-time units that contained more than one sample error were included. In this way, a total of m = 5709 such aggregated subsets were defined. The size of these subsets sets ranged from n_a = 2 to n_a = 1533.

(2) The mean error of each aggregated set was calculated:

$$\mu_a = \frac{1}{n_a}\sum_{j=1}^{n_a}\varepsilon_a((\mathbf{u},t)_j) \tag{9.5}$$

(3) The list of m mean errors, $\mu_{a,1},\mu_{a,2},...,\mu_{a,m}$, was then plotted against the corresponding list of set sizes $n_{a,1},n_{a,2},...,n_{a,m}$ (Figure 9.4). This plot provided an illustration of the central tendency and variation of the means of aggregated sets of errors of different sizes.

(4) The list of m mean errors was sub-divided into a series of b = 1,2,..., B 'bins' according to the size of each set, such that each bin contained k=1,2,...,m_b mean errors $\mu_{a,k}$ calculated from sets of similar size. The standard deviation, σ_b, of the m_b mean errors within each bin was then calculated:

$$\sigma_b = \sqrt{\frac{1}{m_b - 1}\sum_{k=1}^{m_b}\left(\mu_{a,k} - \left(\frac{1}{m_b}\sum_{k}^{m_b}\mu_{a,k}\right)\right)^2} \tag{9.6}$$

(5) The value of σ_b for each bin was then plotted against the corresponding mean set size in each bin. The resulting plot (Figure 9.5) provided an illustration of the effect of aggregation over successively larger space-time units on the standard deviation of the mean prediction error of those units. The theoretical relationship (9.4) was also plotted for comparison.

9.4.3 Estimating the prediction error for individual space-time units

Estimates were required of the error associated with predictions of the sum of MC in each space-time unit of interest across Kenya. The units of interest were defined spatially at the district, provincial, and national level, and temporally at the monthly and annual level. This meant there were six different types of space-time unit of interest: district-months, district-years, province-months, province-years, national-months, and national-years. A total of 7371 such units were defined, and these are detailed in Table 9.2. The results of the procedure described in section 9.4.2 suggested that the use of equation (9.4) as a model for the change in the standard deviation of mean error with aggregation was reasonable, given that error did not display spatial or temporal autocorrelation (Figure 9.3), and that the empirically-observed relationship was very close to that described by this equation (Figure 9.5). In this section, the procedure by which this model was used to estimate the error associated with predictions of the sum of MC in each individual space-time unit is described. This was done in the following steps:

Spatial units (n)	Temporal units (n)	
	Month (84)	Year (7)
District (72)	6048	504
Province (8)	672	56
National (1)	84	7

Table 9.2 The number of space-time units of each type, as defined by three spatial and two temporal levels of aggregation. Figures in parentheses are the numbers of each type of unit.

(1) Rather than use the global estimates of the (unaggregated) prediction error mean $\hat{\mu}_u$ and standard deviation $\hat{\sigma}_u$ as defined in (9.1) and (9.2), it was decided that it was preferable to estimate these parameters locally to better capture regional variations, accepting the reduction in the certainty of these estimates caused by the smaller sample sizes. As such, the local prediction error mean $\hat{\mu}_L$ and standard deviation $\hat{\sigma}_L$ were calculated for each district and province from the relevant subset of the $\varepsilon_v((\mathbf{u},t)_i)$, where the subscript $_L$ is used to denote a local estimate. Where the validation set contained <30 samples for a given district, the provincial estimates were used instead.

(2) The total number of missing data (and hence predictions) in each k=1,2,...,m space-time unit was determined (this value is denoted as n_{STU} where the subscript $_{\text{STU}}$ indicates that the statistic relates to a space-time unit).

(3) The expected error Σ_{STU} of the sum of predictions in each space time-unit was estimated as the product of the local mean error μ_L and the number of predictions in the unit n_{STU}:

$$\hat{\Sigma}_{\text{STU}} = \mu_L n_{\text{STU}} \tag{9.7}$$

(4) The standard deviation of this sum $\sigma[\Sigma_{\text{STU}}]$ was then estimated using the local error standard deviation $\hat{\sigma}_{\text{L}}$ and the number of predictions in the unit n_{STU}, based on the same theoretical relationship established in section 9.4.2:

$$\hat{\sigma}[\Sigma_{\text{STU}}] = \hat{\sigma}_L \sqrt{n_{\text{STU}}} \tag{9.8}$$

This process resulted in estimates of the error associated with predictions of MC for all districts, provinces and nationally for each month and year. The estimated standard deviation of each predicted sum provides a quantification of the associated uncertainty. If a Gaussian model is adopted for the error distribution of each sum, then the estimated standard deviation can be used to calculate indicators of this uncertainty such as a 95% confidence interval.

9.4.4 Summarising the prediction error for each aggregation level

The process described above provides estimates of the mean and standard deviation of the prediction error for each of the 7371 space-time units of interest in this study. It was necessary to summarise these estimates to provide a single measure of the accuracy of predictions for each aggregation level (e.g. what is the expected range of errors for predictions of MC at the level of district-months, province-years etc). This was done in the following steps:

(1) The estimated sum of prediction errors in each space-time unit, $\hat{\Sigma}_{STU}$, and accompanying standard deviation $\hat{\sigma}[\Sigma_{STU}]$ were expressed as a percentage of the predicted total MC for that unit. Both predicted MC (missing data) and known MC (data) were included in the denominator and, as such, the estimated percentage errors accounted for the proportion of missing data. This is important since, for example, a prediction error that is large relative to the sum of predictions can still be small relative to the predicted total, if few data are missing.

(2) The distribution of the percentage error of the sum for each space-time unit was assumed Gaussian and fully defined in each case by the estimated mean percentage error and standard deviation discussed above. Each aggregation level therefore contained a set of distributions modelling the uncertainty in the m predicted MC totals at that level.

(3) A useful summary was the 95% confidence interval that defined the range of percentage errors that can be expected in 95% of cases at each aggregation level. These confidence intervals were estimated empirically using a Monte-Carlo simulation exercise. Each simulated realisation proceeded in two steps. Firstly, a single space-time unit was chosen at random from the full set that made up each aggregation level. Secondly, a random draw was made from a normal distribution defined by the estimated percentage mean error and standard deviation of that unit.

(4) 100,000 realisations were simulated and the Q_{025} and Q_{975} quantiles of the resulting distribution were used to define the lower and upper bounds of the 95% confidence interval of the percentage error of the predicted sum for each aggregation level.

This procedure resulted in, for example, estimates of the range (expressed as a 95% confidence interval) of percentage errors that could be expected for predictions of total MC for all facilities in a district over a month, all facilities in a province over a year and so on.

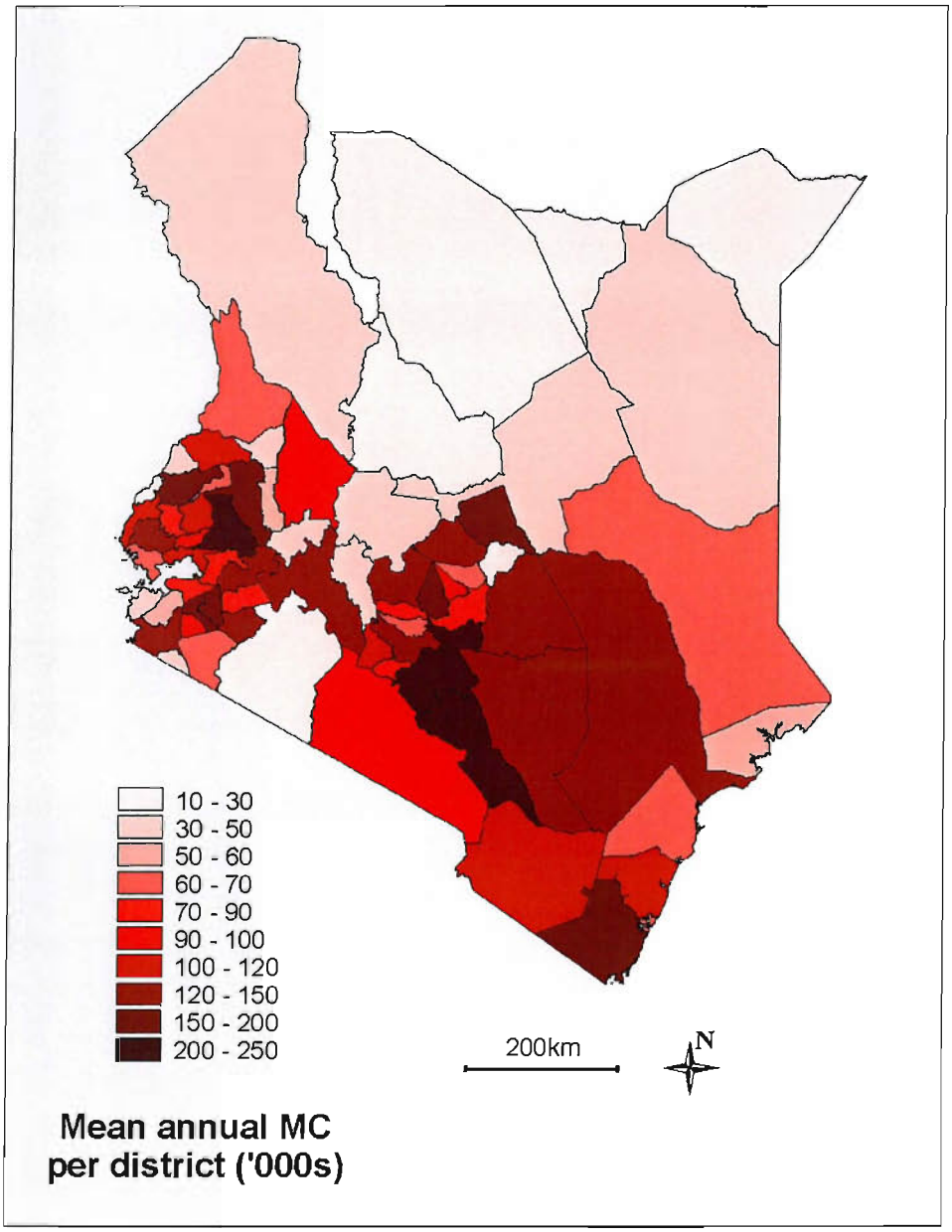


Figure 9.2 Number of outpatients treated for malaria (MC) at government facilities: Predicted mean annual totals for each district for the period 1996-2002. Values represent the combined sum of existing and predicted values.

	Data	Predictions	Combined Total
Dispensaries	1,323,271	2,625,968	3,949,239
Health Centres	864,945	872,214	1,737,159
Hospitals	479,331	628,992	1,108,324
All	2,667,547	4,127,175	6,794,722

Table 9.3 Predicted mean annual counts of outpatients treated for malaria at all Kenyan government hospitals, health centres and dispensaries for the period 1996-2002. Totals are given for data, predictions, and for the combined total.

9.5 Results

9.5.1 Prediction of treatment burdens

A total of 181,860 predictions of MC were made for the 2165 GoK health facilities over the 84- month study period. The sum of these predictions was 28.89 million cases which, when added to the 18.67 million cases reported in the existing records led to a predicted total of 47.56 million cases nationwide for the seven-year period. The mean annual total was 6.79 million cases with a mean of 261.5 cases per facility-month. The corresponding values for each facility type were 1.11 million for hospitals, 1.74 million for health centres and 3.95 million for dispensaries with means of 716.0, 300.3, 211.8 cases per facility-month, respectively. A summary of these results is presented in Table 9.3 and a complete breakdown is shown in Table 9.4.

Mean annual totals for each district displayed a pattern of spatial heterogeneity and this is illustrated in the map in Figure 9.2. Some features of this map are worthy of discussion. Firstly, these district totals have not been standardised by any measure of district population. The decision not to present any standardised version of this map reflects that the overall motivation of the project was to predict counts of malaria diagnoses in outpatients and not to use these predictions to make inferences about the distribution of malaria in the population. Without standardisation, the count of diagnoses in each district is influenced as much by the population size as by the presence and diagnosis of malaria. Nevertheless, the spatial pattern displayed in the map corresponds in broad terms with the known distribution of malaria across the country, with large predicted values (darker red districts) in areas of high prevalence around the western lake shore regions and along the Indian Ocean coastline, and smaller predicted values

(paler red districts) in areas of low prevalence in the elevated western highlands region and in the arid north-east. The set of districts with large predicted values in the south-eastern quadrant away from the coast, however, are in areas known to have a relatively low burden of malaria. Although they have relatively high population densities, this is unlikely to account entirely for the apparent anomaly. A more likely explanation is that rates of misdiagnosis of malaria (i.e. false-positive diagnosis) in these low-malaria districts are particularly high. Although beyond the scope of this thesis, a detailed comparison of the predictions of outpatient malaria counts developed in this project with the latest spatial models of malaria prevalence may highlight regions where the two appear mismatched. Such an activity may highlight those regions where diagnosis of malaria is common despite the known or assumed absence of the disease in the population.

9.5.2 Model validation

9.5.2.1 Error parameter estimation and variography

The sample mean prediction error (9.1) for predictions of MC at individual facility-months was -1.28 cases, as estimated from the validation set of 6349 known prediction errors, $\varepsilon_v((\mathbf{u},t)_i)$. The standard deviation (9.2) was 236.62 cases. The space-marginal and time-marginal variogram of the errors are shown in Figure 9.3. The spatial variogram showed no evidence of spatial autocorrelation up to lags of 90 km. Semivariances over some of the shortest lags displayed large values, although this was

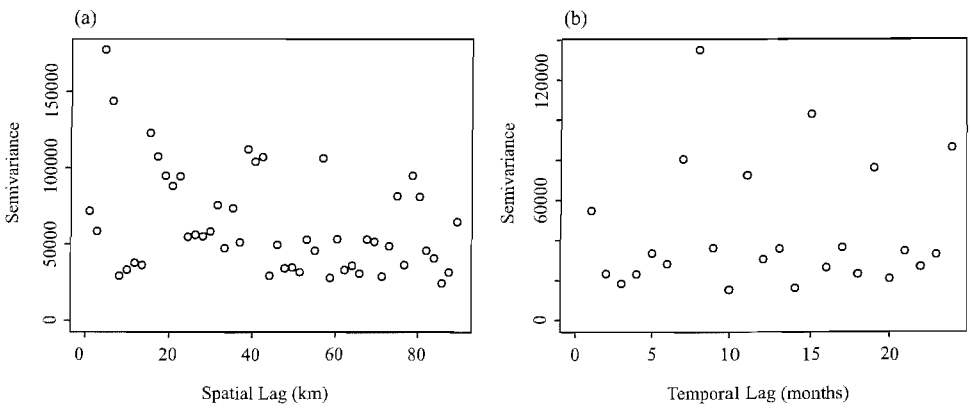


Figure 9.3 Spatial (a) and temporal (b) variograms of the error in predictions of MC, as estimated using a validation set.

		Georeferenced Data			Georeferenced Predictions			Non-georeferenced Data			Non-georeferenced Predictions			All Data			All Predictions			Combined Totals		
		sum	mean	n	sum	mean	n	sum	mean	n	sum	mean	n	sum	mean	n	sum	mean	n	sum	mean	n
Hospitals	1996	596,574	847.4	704	615,525	729.3	844	-	-	0	-	-	0	596,574	847.4	704	615,525	729.3	844	1,212,099	783.0	1,548
	1997	530,883	874.6	607	736,349	782.5	941	-	-	0	-	-	0	530,883	874.6	607	736,349	782.5	941	1,267,232	818.6	1,548
	1998	502,925	785.8	640	720,885	793.9	908	-	-	0	-	-	0	502,925	785.8	640	720,885	793.9	908	1,223,810	790.6	1,548
	1999	435,363	647.9	672	703,740	803.4	876	-	-	0	-	-	0	435,363	647.9	672	703,740	803.4	876	1,139,103	735.9	1,548
	2000	500,090	648.6	771	522,448	672.4	777	-	-	0	-	-	0	500,090	648.6	771	522,448	672.4	777	1,022,538	660.6	1,548
	2001	350,706	555.8	631	593,860	647.6	917	-	-	0	-	-	0	350,706	555.8	631	593,860	647.6	917	944,566	610.2	1,548
	2002	438,778	669.9	655	510,139	571.3	893	-	-	0	-	-	0	438,778	669.9	655	510,139	571.3	893	948,917	613.0	1,548
	Total	**3,355,319**	-	**4,680**	**4,402,947**	-	**6,156**	-	-	**0**	-	-	**0**	**3,355,319**	-	**4,680**	**4,402,947**	-	**6,156**	**7,758,266**	-	**10,836**
	Mean	**479,331**	**716.9**	**669**	**628,992**	**715.2**	**879**	-	-	**0**	-	-	**0**	**479,331**	**716.9**	**669**	**628,992**	**715.2**	**879**	**1,108,324**	**716.0**	**1,548**
Health Centres	1996	1,021,279	340.3	3,001	753,023	280.2	2,687	6,302	572.9	11	32,130	378.0	85	1,027,581	341.2	3,012	785,153	283.2	2,772	1,812,734	313.4	5,784
	1997	926,960	360.4	2,572	970,458	311.4	3,116	5,453	605.9	9	42,204	485.1	87	932,413	361.3	2,581	1,012,662	316.2	3,203	1,945,075	336.3	5,784
	1998	807,898	328.1	2,462	985,550	305.5	3,226	5,906	656.2	9	29,920	343.9	87	813,804	329.3	2,471	1,015,470	306.5	3,313	1,829,274	316.3	5,784
	1999	949,075	346.8	2,737	903,500	306.2	2,951	1,047	523.5	2	37,673	400.8	94	950,122	346.9	2,739	941,174	309.1	3,045	1,891,296	327.0	5,784
	2000	886,496	312.4	2,838	755,802	265.2	2,850	1,160	386.7	3	26,965	289.9	93	887,656	312.4	2,841	782,767	266.0	2,943	1,670,423	288.8	5,784
	2001	693,158	277.3	2,500	796,711	249.9	3,188	316	316.0	1	23,025	242.4	95	693,474	277.3	2,501	819,737	249.7	3,283	1,513,211	261.6	5,784
	2002	748,770	291.4	2,570	723,450	232.0	3,118	796	199.0	4	25,088	272.7	92	749,566	291.2	2,574	748,538	233.2	3,210	1,498,104	259.0	5,784
	Total	**6,033,636**	-	**18,680**	**5,888,494**	-	**21,136**	**20,980**	-	**39**	**217,005**	-	**633**	**6,054,616**	-	**18,719**	**6,105,499**	-	**21,769**	**12,160,115**	-	**40,488**
	Mean	**861,948**	**323.0**	**2,669**	**841,213**	**278.6**	**3,019**	**2,997**	**537.9**	**6**	**31,001**	**342.8**	**90**	**864,945**	**323.4**	**2,674**	**872,214**	**280.5**	**3,110**	**1,737,159**	**300.3**	**5,784**
Dispensaries	1996	1,439,378	226.2	6,363	2,181,126	209.2	10,425	15,545	1413.2	11	456,680	247.0	1,849	1,454,923	228.3	6,374	2,637,806	214.9	12,274	4,092,729	219.5	18,648
	1997	1,244,143	251.2	4,953	2,725,141	230.3	11,835	24,358	1873.7	13	518,171	280.5	1,847	1,268,501	255.4	4,966	3,243,312	237.0	13,682	4,511,813	241.9	18,648
	1998	1,252,366	248.7	5,036	2,810,088	239.1	11,752	14,754	1341.3	11	516,469	279.3	1,849	1,267,120	251.1	5,047	3,326,556	244.6	13,601	4,593,676	246.3	18,648
	1999	1,431,059	249.7	5,731	2,295,425	207.6	11,057	16,385	1489.5	11	436,944	236.3	1,849	1,447,444	252.1	5,742	2,732,369	211.7	12,906	4,179,813	224.1	18,648
	2000	1,491,673	227.5	6,558	1,865,179	182.3	10,230	13,218	574.7	23	389,483	212.0	1,837	1,504,891	228.7	6,581	2,254,662	186.8	12,067	3,759,553	201.6	18,648
	2001	1,131,124	200.6	5,639	1,890,144	169.5	11,149	2,720	209.2	13	329,375	178.3	1,847	1,133,844	200.6	5,652	2,219,518	170.8	12,996	3,353,362	179.8	18,648
	2002	1,179,463	201.4	5,855	1,633,313	149.4	10,933	6,710	258.1	26	334,243	182.2	1,834	1,186,173	201.7	5,881	1,967,555	154.1	12,767	3,153,728	169.1	18,648
	Total	**9,169,206**	-	**40,135**	**15,400,414**	-	**77,381**	**93,690**	-	**108**	**2,981,364**	-	**12,912**	**9,262,896**	-	**40,243**	**18,381,779**	-	**90,293**	**27,644,675**	-	**130,536**
	Mean	**1,309,887**	**228.5**	**5,734**	**2,200,059**	**199.0**	**11,054**	**13,384**	**867.5**	**15**	**425,909**	**230.9**	**1,845**	**1,323,271**	**230.2**	**5,749**	**2,625,968**	**203.6**	**12,899**	**3,949,239**	**211.8**	**18,648**
All	1996	3,057,231	303.7	10,068	3,549,674	254.3	13,956	21,847	993.0	22	488,810	252.7	1,934	3,079,078	305.2	10,090	4,038,483	254.2	15,890	7,117,561	274.0	25,980
	1997	2,701,986	332.3	8,132	4,431,947	278.9	15,892	29,811	1355.0	22	560,375	289.7	1,934	2,731,797	335.0	8,154	4,992,322	280.1	17,826	7,724,119	297.3	25,980
	1998	2,563,189	315.0	8,138	4,516,523	284.3	15,886	20,660	1033.0	20	546,389	282.2	1,936	2,583,849	316.7	8,158	5,062,912	284.1	17,822	7,646,761	294.3	25,980
	1999	2,815,497	308.0	9,140	3,902,666	262.2	14,884	17,432	1340.9	13	474,618	244.3	1,943	2,832,929	309.5	9,153	4,377,283	260.1	16,827	7,210,212	277.5	25,980
	2000	2,878,259	283.1	10,167	3,143,429	226.8	13,857	14,378	553.0	26	416,447	215.8	1,930	2,892,637	283.8	10,193	3,559,877	225.5	15,787	6,452,514	248.4	25,980
	2001	2,174,988	248.0	8,770	3,280,715	215.1	15,254	3,036	216.9	14	352,400	181.5	1,942	2,178,024	248.0	8,784	3,633,115	211.3	17,196	5,811,139	223.7	25,980
	2002	2,367,011	260.7	9,080	2,866,902	191.8	14,944	7,506	250.2	30	359,331	186.6	1,926	2,374,517	260.6	9,110	3,226,232	191.2	16,870	5,600,749	215.6	25,980
	Total	**18,558,161**		**63,495**	**25,691,856**		**104,673**	**114,670**		**147**	**3,198,369**		**13,545**	**18,672,831**		**63,642**	**28,890,225**		**118,218**	**47,563,056**		**181,860**
	Mean	**2,651,166**	**292.3**	**9,071**	**3,670,265**	**245.4**	**14,953**	**16,381**	**780.1**	**21**	**456,910**	**236.1**	**1,935**	**2,667,547**	**293.4**	**9,092**	**4,127,175**	**244.3**	**16,888**	**6,794,722**	**261.5**	**25,980**

Table 9.4 Predicted mean annual counts of outpatients treated for malaria at all Kenyan government hospitals, health centres and dispensaries for the period 1996-2002.

likely to be attributable to sampling variation caused by the scarcity of pairs of sample points at these short spatial separations. The temporal variogram also showed no evidence of temporal autocorrelation.

9.5.2.2 Assessing the effect of aggregation on the variance of prediction errors

Figure 9.4 shows a plot of the mean prediction error, μ_a, of each of the subsets created by aggregating the sample error set $\varepsilon_v((\mathbf{u},t)_i)$ over facilities, districts, provinces and nationally and by month and year. Each value of μ_a is plotted against the size of the aggregated set in question, n_a. Mean errors are centred approximately on zero at all aggregation sizes, but the variation around this central value displays a marked reduction as n_a increases. This plot provides a qualitative illustration of the effect of aggregating predictions (over space and time) on the mean error of those aggregated sets, and the variation that can be expected around that mean error. Figure 9.5 shows the results of the next stage of analysis which provided a more quantitative description of this effect by estimating the standard deviation of the mean errors of sets, $\sigma[\mu_a]$, within a series of bins representing different subsets of differing size. The empirically estimated values lie very close to the line that marks the theoretical relationship shown in equation (9.4), suggesting that this equation provides a useful model of the dependence of $\sigma[\mu_a]$ on n_a in the current setting, despite possible deviations from assumptions of IID.

9.5.2.3 Prediction error at each aggregation level

Comparison of data with predictions for the 6349 randomly selected MC data in the validation set yielded mean prediction errors for hospitals, health centres, and dispensaries of 58.2, -8.8, and -4.7 cases per facility-month, respectively. The true and predicted sums of the entire national test set were 1,899,234 and 1,891,136, respectively, representing an overall prediction error of -0.4% for the validation set.

The predictive accuracy of the model increased as predictions of MC totals were made over larger aggregated space-time units. Table 9.5 shows the expected range (95% confidence interval) of percentage errors for predictions of total MC (i.e. combined total of data and predictions) at different levels of spatial and temporal aggregation. These

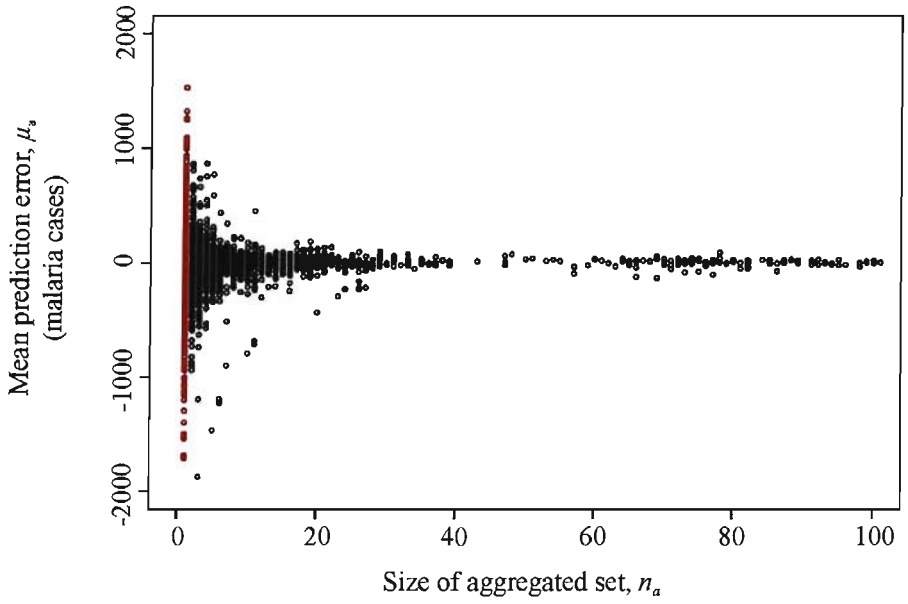


Figure 9.4 Mean prediction error, μa, against size, na, for 5709 subsets taken from the sample error set. Points marked in red are individual (unaggregated) prediction errors. A small number of sets exceeded the x-axis range of this plot (maximum n_a=1533) and these have been omitted to allow clearer display of values with smaller n_a. All omitted points showed no visible departure from $\mu_a = 0$.

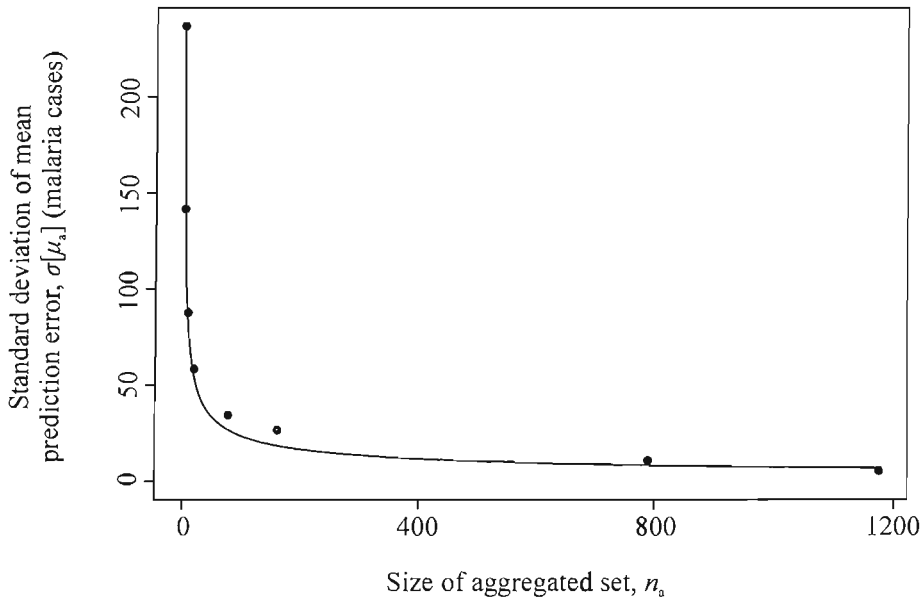


Figure 9.5 Empirical relationship between the size of subsets of the test data set and the standard deviation of their mean prediction errors. Subsets of different sizes, n_a, were created from the test set by aggregating across space (by district, province, and nationally) and through time (by month and year) and the mean prediction error, μ_a, of each subset was calculated. These subsets were then placed in bins according to their size n_a and the standard deviation of the mean errors in each bin, $\sigma[\mu_a]$, was calculated. The x-axis position of each point represents the mean subset size in that bin. The theoretical relationship $\sigma[\mu_a]=\sigma_a/ n_a^{1/2}$ is also shown (line).

confidence intervals are not symmetric as they incorporate not only the expected variance of the error, but any bias introduced by the expected mean error being non-zero. If the bound of the confidence interval that is farthest from zero is considered, then the results can be summarised by stating that 95% of errors for the prediction of total MC at the district-month level were expected to be within 35.3% of the true value. The equivalent error for predictions of annual totals at the provincial level was 12.2% and for annual national totals it was -1.3%.

	Range of expected errors	
	Month	Year
District	-32.72% to 35.31%	-15.71% to 21.25%
Province	-15.78% to 20.36%	-5.65% to 12.19%
National	-3.73% to 2.98%	-1.25% to 0.58%

Table 9.5 Expected percentage errors (95% confidence intervals) in predictions of total outpatients treated for malaria over different levels of spatial and temporal aggregation. Errors were calculated from a validation exercise in which 6349 monthly records (10%) were removed from the data set and predicted using the remaining 90%.

9.6 Chapter Summary

This chapter has presented the implementation of Model 3 to predict MC at all missing facility-months. This final implementation was carried out using an updated version of the integrated NHSD-HMIS database which, although containing very few additional data compared to the version presented in Chapter 3, incorporated a further 400 government health facilities, thus representing the most comprehensive inventory of facilities that currently exists for Kenya. In order to validate the predictions of MC, an empirical validation approach was developed and presented in this chapter. In this approach, 10% of the available data were selected at random, temporarily removed, and predicted using the remaining 90% of data to obtain a set of prediction errors. This set was then used to infer the expected prediction errors of the main prediction exercise at different levels of spatial and temporal aggregation, after establishing that the use of a theoretical relationship was appropriate in the current setting.

After combining data and predictions together, the predicted mean annual total number

of outpatients treated for malaria was 6.79 million cases. The validation exercise suggested that there is a 95% chance that this prediction is accurate to within 1.3% of the true value. The predictions presented in this chapter, and the associated validation exercise address the primary aims of this project. The nature of these results and their implications are discussed in detail in the next, and penultimate, chapter.

Chapter 10

Discussion

Chapter 10

10. Discussion

10.1 Introduction

The purpose of this penultimate chapter is to provide an overview of the progression of ideas and techniques that resulted from this project, and to discuss the most significant issues that have arisen. In the next section, the evolution, rationale and ultimate success of the modelling strategy that incorporated TC data in models to predict MC are discussed. This is followed by an appraisal of the techniques developed to quantify the uncertainty of MC predictions. The output of the final model implementation is then considered and these results and their associated uncertainty are examined in the context of their utility to public health decision-makers. The wider applicability of the modelling strategies developed in this project are then discussed. The chapter concludes by considering possible avenues by which this work can be taken forward in future studies.

10.2 Use of TC in predicting MC

10.2.1 Overview of evolution of modelling strategy

A conceptual consideration of the factors that determine the MC variable (Chapter 5) proposed that MC is driven by spatially-dependent factors, mostly related to environmental heterogeneity, and spatially-independent factors, mostly related to characteristics of individual health facilities and their catchment populations. Geostatistical techniques are aimed at characterising and predicting spatial (and/or temporal) variability in the variable of interest and where non-spatial variability exists it inevitably introduces greater uncertainty in predictions of that variable. In response to

this, two alternative strategies were conceived with the aim of accounting for some of the non-spatial variability present in MC.

The first strategy was to investigate the feasibility of developing catchment models that would allow the size of individual health facility catchment populations to be predicted across Kenya. Such predictions could then be used as denominator values to standardise the raw MC data. This strategy was pursued in two collaborative studies presented in Chapter 6. The first study investigated the use of Thiessen polygons, one of the most straightforward and widely-implemented approaches to developing catchment boundary models. The principal assumptions of this approach, that all care-seekers utilise their nearest facility, and that the rate of utilisation is even within each catchment, were found to be inappropriate in four sample districts. The second study implemented a Thiessen polygon model in the same four districts and also constructed a series of more refined models. A journey-time metric was developed that replaced straight-line distance as a more realistic way of assessing care-seekers' physical access to health facilities and data from a household survey was used to model the way different facility types compete to draw in patients from different distances. When the basic Thiessen polygon model was adjusted to incorporate these refinements, the resulting catchment boundaries were found to predict more accurately the patterns of facility choice made by individual homesteads. Because the data required to develop the more refined catchment models were not available for facilities across Kenya, the only feasible approach to estimating catchment population sizes was to use Thiessen polygons. An important conclusion of the two catchment modelling studies discussed above, however, was that such a model is inappropriate in the Kenyan setting and likely to produce misleading predictions. As such, directly predicted catchment populations remained unavailable for the purposes of this project. To exemplify this point, when catchment population estimates were derived using Thiessen polygon boundaries in conjunction with enumeration area-level population data, and these estimates were used as denominators to the raw MC counts, the spatial variograms of the resulting standardised variables displayed a complete absence of spatial autocorrelation.

In light of the limitations of catchment models to provide denominator values, as described above, a second strategy was developed. This was to use the TC data on total outpatient diagnoses that accompanied the disease-specific MC values. The rationale for

the use of TC was that these values were driven in broad terms by the size of each facility and its utilisation by the population and, therefore, acted as a proxy measure of catchment population size. As such, these values were likely to contain information that might be used to assist in the prediction of MC.

10.2.2 Incorporation of TC in the modelling framework

The most straightforward way of incorporating TC values as a way of standardising MC is to use them directly as a denominator, thus defining the variable called MP in this project. When the spatial variogram was estimated using all available MC data, and compared to the corresponding variogram for MP, the difference was dramatic. Whilst the former indicated zero spatial autocorrelation, the latter indicated considerable spatial continuity. These variograms suggested that the use of TC as a denominator did result in an effective standardisation of the raw MC data and accounted for much of its non-spatial variability, allowing the inherent spatial structure in MC to be revealed. However, when the raw MC data were split up according to facility type, the MC variogram also indicated considerable spatial structure. This was an important result that suggested that much of the non-spatial variation in MC is caused by differences between the three facility types and that, when these are considered separately, there is a degree of within-class consistency in factors such as catchment size and utilisation. This simple approach also negated some of the benefit of using the standardised MP variable, since spatial MC variograms for each facility type indicated only marginally less spatial continuity than those for MP.

If TC data were available at all facility-months, then the standardised MP variable could be predicted at all locations with missing MC values and back-transformed to the desired MC predictions using these TC data. Crucially, however, TC data were co-located with MC such that the respective patterns of data presence or absence corresponded exactly. Where MC data were missing, therefore, no TC data were available to perform this back-transform. This presented a potential road-block but a possible solution was to predict TC values at these locations using STOK in the same way that MC values were predicted. Examination of the variograms for TC indicated that, when data were separated by facility type, the variable displayed a degree of both spatial and temporal structure and this allowed prediction of the missing TC values which could then be used

for the back-transforms. This approach (Model 2) was compared to the simple use of raw MC values (Model 1) in Chapter 8. The respective characteristics of the MC and MP variograms of the full data set were manifest in the model outputs, with Model 2 able to predict MC with much greater accuracy than Model 1. When facilities were considered separately, however, the differences in accuracy between Model 1 and Model 2 were negligible with Model 1 predicting more accurately for some facility types. The failure of Model 2 to provide significantly more accurate predictions of MC than Model 1 can be explained by two factors. The first factor was the unexpected degree of spatial structure that was revealed in MC by simply considering each facility type independently, which acted to standardise MC values almost as effectively as dividing them by TC. The second was the need, in Model 2, to predict TC at unsampled locations which inevitably introduced further uncertainty in the ultimate back-transformed predictions of MC.

The third modelling framework, Model 3, was developed as a way of reducing the uncertainty introduced by the need to predict TC. By averaging all predictions of TC, along with the available TC data, over the 84 monthly values at each location, a single facility-specific denominator was obtained (MMTC) that was more robust to both monthly fluctuations in the true value of TC and to error in the TC predictions. When this denominator was used to standardise MC, variograms of the resulting standardised variable (SMC) indicated a greater degree of spatial continuity than both MC and MP. Predictions of MC made using this model were the most accurate of the three approaches tested and this model was therefore used in the final implementation. The increases in accuracy offered by Model 3, however, were not substantial when compared to the simple approach of Model 1. Whilst the enhanced spatial structure indicated by the SMC variogram confirms that TC data contain information of use in predicting MC, the benefit of using this standardised variable is, again, largely negated by the uncertainty introduced by having to predict TC. Whilst Model 3 did offer a partial solution to this problem compared to Model 2, it remains a substantial limitation in both approaches.

An important conclusion from this work is that the strategy of standardising MC to account for facility-specific, non-spatial, sources of variation before predicting missing MC values is the correct one, and this is illustrated by the substantially greater spatial

continuity indicated in the SMC variograms. Furthermore, TC data contain useful information that can be used in this standardisation. The absence of these data at the locations where MC predictions are required, however, presents a major limitation to their use in this context, and this limitation has been only partially overcome in this project. Model 3 can be thought of more generally as a generic approach to predicting MC that can be built upon and enhanced as external sources of data become available for predicting facility-specific denominators. Detailed data from each facility on factors such as its size, staffing levels, and services offered may assist in predicting these values, as may data on catchment population access and behaviour and such data are currently being assembled for Kenya by the KEMRI-University of Oxford-Wellcome Trust Collaborative Programme team.

10.3 Assessing prediction uncertainty

In Chapter 8, the development of an uncertainty model was described that used a space-time adaptation of sequential Gaussian simulation in a framework that simulated the uncertainty associated with predictions of MC made using Model 3. Evaluation of this uncertainty model suggested that it provided accurate measures of the uncertainty in predictions of both individual MC values and sets of values aggregated over space-time regions. This model stands as a useful accompaniment to the proposed prediction framework. The principal downside of this model, however, is its conceptual and implementational complexity. The large number of processing steps involved mean that the model must be implemented with care and numerous modelling decisions have to be made by the user. Furthermore, the use of simulation algorithms on a large space-time data set is extremely computationally demanding and the resources used to implement these algorithms in this project (a high-performance Beowulf cluster of several hundred parallel processors) are not available in most settings. The conceptual complexity does not limit the usefulness of the model *per se*, although it may reduce its appeal to potential end-users such as public health decision-makers who may be less willing to put faith in uncertainty measures when the underlying model is not easily understood. Concerns of this type were raised by practitioners in Nairobi and, whilst strictly unjustified from a theoretical standpoint, this project is concerned with a real-world issue and the perception of end-users is a factor that cannot be ignored.

In response to the factors stated above, an alternative approach to assessing prediction uncertainty was developed for the final implementation of Model 3 in Chapter 9. The uncertainty model presented in Chapter 8 extended the model-based approach used to predict MC to the assessment of prediction uncertainty based on inferred characteristics of the RF model. As such, uncertainty estimates in this model took into account the spatiotemporal configuration of data and prediction locations and the likely strength of relationships between these locations. In contrast to this model-based approach, the method proposed in Chapter 9 used an empirical approach that relied on summary statistics derived from a sample of known prediction errors to infer characteristics of the unknown errors at all prediction locations. A simple model was used to determine how the uncertainty associated with aggregated sets of predictions changed as these sets became larger. Although conceptually much more simple than the model-based simulation model, this approach was nevertheless likely to have provided reasonable measures of prediction uncertainty at different levels of spatial and temporal aggregation. Furthermore, this was achieved using an approach that was simple, straightforward and quick to implement, and that may be conceptually more transparent to potential end-users.

10.4 The modelling output in context

Predictions of variables derived from HMIS outpatient data in a low-income country are likely to have a large inherent uncertainty associated with them and this is reflected in this study in both the variograms and model outputs. At the level of individual facility-months, predictions with the accuracies presented in Table 8.1 are likely to be of only limited use to health system decision-makers (MAE was 26.8%, 27.6%, and 22.9% of the mean MC value for hospitals, health centres and dispensaries, respectively). Strategic decision-making is rarely made at this level, however, and the accuracy of aggregated predictions of MC at monthly and annual district, provincial, and national levels are of greater importance. Predictions of mean MC at these levels entail the averaging or summation of many individual predictions (along with existing data) across space-time regions. The results of the empirical validation exercise described in the previous section suggested that the accuracy of predictions of total MC over different space-time units would increase as more individual MC predictions were aggregated. As

such, predictions of total MC for a given district and month could be expected to be within 35.3% of the true value, whilst the equivalent provincial and national values were 20.4% and 3.7%, respectively. Predictions of the national annual total could be expected to be within 1.3% of the true value. Such an effect is expected as individual predictions are aggregated. The mean error is expected to remain small (since kriging is designed to produce unbiased predictions, with a zero expected mean error), whilst the variance of this error decreases, allowing more precise predictions of aggregated values.

It is understood that no equivalent exercises have been undertaken to systematically evaluate the malaria treatment burden in the government sector in Kenya and so the predictions made in this project stand as probably the most reliable source of information for related decision-making at district, provincial, and national levels. The accompanying measures of uncertainty further enhance the utility of these predictions, allowing decision-makers to identify realistic ranges of possible values. The prediction of a mean national annual total of 6.8 million cases during the study period with an expected margin of error of 1.3% represents a tangible difference to the rudimentary approach of multiplying nationally available data by a proportion of under-reporting which would result in a crude estimate of 7.6 million cases.

In the current setting, it is difficult to quantify the levels of accuracy required by public health decision-makers to allow effective evidence-based decision-making. A valid question is whether a more rudimentary approach to predicting treatment burden would be sufficient for effective public health decision making. It is argued here, however, that the difference between national predictions of 6.8 million cases and 7.6 million cases is likely to be a substantial one in the context of national-level policy and management decisions. Furthermore, as predictions are made at progressively finer levels of spatial and temporal aggregation, the relative disparities between rudimentary and sophisticated methods of estimation are likely to increase. Even if a rudimentary method resulted in similar predictions to the approach presented in this project, there are at least two further arguments for using the more sophisticated approach. Firstly, the approach presented here provides a realistic measure of the uncertainty in the final predictions, which is an essential accompaniment to any prediction, allowing decision-makers a tangible yardstick of prediction reliability. Secondly, the process of developing and testing a more sophisticated approach adds credence to the resulting predictions. This may be

important when, for example, these predictions are used in the procurement of national drug supplies through international donor agencies since any predictions must be based on reliable methods.

It is important to note that the various measures of uncertainty presented in this project refer only to prediction uncertainty. Sources of uncertainty that are inherent in the data themselves have not been considered. Factors such as the incorrect filling out of reporting forms, or subsequent errors in data transmission and entry into the HMIS database introduce uncertainty into the data which will be transferred to the predicted MC totals but not included in the estimates of prediction error. Such sources of uncertainty can only be quantified reliably by obtaining 'gold-standard' data from a sample of facilities and comparing this to the corresponding routine HMIS data.

10.4.1 What variable has been quantified?

Having presented quantifications of MC totals across Kenya, it is important to reassess what such totals actually represent. Of crucial importance is the distinction between MC and the burden of malaria in the population. Because only a small proportion of incidences of malaria result in a visit to a formal health service provider, as discussed in Chapter 5, the pattern of malaria seen at health facilities is only loosely connected to that in the population as a whole and the results of this project should not be used to evaluate the latter. Furthermore, MC totals do not even represent the true number of incidences of malaria that are seen at health facilities due to the misdiagnosis of malaria and other conditions. Because of these factors, MC must be interpreted as quantifying the number of diagnoses that have been made for malaria and, importantly, the number of malaria treatments that have been administered. Despite the disparities between MC and the true pattern of population and outpatient malaria morbidity, the variable remains critical for health-service planning because it determines the level of resources required to treat patients under this diagnosis.

10.5 Wider applicability

Whilst the primary aim of this project was the quantification of the treatment burden for malaria in government facilities in Kenya, the value of the approaches presented in this thesis can extend to other settings. It is likely that the treatment burden for other diseases can be predicted in a similar way, although the accuracy of such predictions will be influenced by the degree of spatial dependence related to the disease in question. Whilst many diseases are environmentally constrained in a similar way to malaria, others are not and geostatistical approaches may be less appropriate in these cases.

In principle, the techniques developed in this project could be applied in other countries where HMIS data are incomplete. A fundamental requirement, however, is that a comprehensive list of health facilities is available and that these facilities are georeferenced. This was made possible for this project by the work of the KEMRI-University of Oxford-Wellcome Trust Collaborative Programme team and the construction of the NHSD resource. Rather than a limitation to wider application of the approach outside Kenya, however, it is argued here that knowing where service providers are located is a must for any health planning agency and that health service GIS-frameworks such as the NHSD should be developed everywhere.

In general terms, this project represents, to the best knowledge of the author and all involved, the first attempt to tackle the problem of missing HMIS data by predicting individual missing records through the exploitation of space-time structure as opposed to the crude adjustment of aggregated totals based on the proportion of missing data. As such, the approach developed in this project stands as a useful tool that can be applied to HMIS settings to obtain more reliable information for public health decision-making.

10.6 Future work

The model development and results presented in this study raise several important questions that require attention and can form the basis of further studies.

10.6.1 Incorporation of information from covariates

A source of potentially useful data for the prediction of MC exists in the form of environmental covariates provided by remote sensing satellites and other sources. Remotely sensed covariates such as land-surface-temperature, cold-cloud-duration (Tucker and Sear, 2001) (used as an indicator of rainfall in the tropics) and the normalised difference vegetation index (Justice et al., 1985) (a function of soil moisture availability) are known to be related to malaria distribution (Hay et al., 1996, 1997, 1998a, 1998b; Thomson et al., 1996, 1997; Snow et al., 1998; Craig et al., 1999; Hay and Lennon, 1999; Hay, 2000; Omumbo et al., 2002, 2005; Rogers et al., 2002), either synchronously or with a time-lag. If such data are obtained at appropriate spatial and temporal resolutions, then such data can be assimilated into the framework for predicting MC. Several geostatistical techniques are available to assimilate these data into the predictions, including STK with an external drift (STKED) (Goovaerts, 1997; Lloyd, 2002). STKED utilises the local regression relation between the property of interest and the covariates to estimate a local trend surface, such that STK can proceed on the residuals.

10.6.2 Updating predictions using sentinel facilities

This project has focused on HMIS data for the years 1996 – 2002. For the duration of this project, more recent data from the HMIS were not available. A delay is inevitable between patients being diagnosed and treated at outpatient departments, and the data being recorded, transferred and ultimately entered into the national HMIS database. The total delay between diagnosis and acquisition of useable data is currently approximately two years. Therefore, although the approach developed in this project allows a complete HMIS record to be reconstructed for the period under study, this record will inevitably lag behind the current situation by at least two years. As such, critical information on current and future malaria treatment burdens, that would be of greatest use to public health planners, is missing.

Extension of the prediction framework to the present day will require the addition of supplemental data. Of particular interest is the addition of data from a few sentinel facilities. These are facilities where projects have been initiated to ensure data collection

and collation is both rapid and of a high quality. Such sentinel facilities exist across Kenya and an important research question is: what advantage would be gained by supplementing the HMIS data with more timely data from sentinel sites for which up-to-date information of trusted quality could be received in near real-time?

To incorporate information from these sentinels, the modelling framework could be modified to include a temporal trend component that is fitted to sentinel data that span across all available years, including up to the present day, to capture inter-annual variation. Each sentinel facility would be fitted with a temporal trend model most likely involving a two-parameter linear trend component and four-parameter double periodic component (to represent the first and second seasons of malaria - with the two amplitude and two phase parameters adjusted locally spatially).

The temporal trends estimated for each sentinel facility could then be used to interpolate the six temporal trend parameters across all facilities. Various approaches could be implemented and compared to achieve this. Firstly, a spatial trend surface could be fitted to each of the (six) parameters of the temporal trend. This spatial trend could be fitted at the same time as the per-sentinel temporal trends. The number of spatial polynomial coefficients would need to be chosen carefully such as to represent the spatial variation adequately. It would be possible to fit the spatial trend as an integral part of the temporal trend fitting procedure at each sentinel. A second approach would be to predict the temporal trend parameters for non-sentinel facilities using OK. Once the temporal trend parameters had been estimated at each facility, it would then be possible to predict MC values up to the present day using a geostatistical approach such as STK with a trend (STKT) (Journel and Rossi, 1989; Lloyd and Atkinson, 2002).

If the above procedure could be implemented successfully then an intuitive extension would be to attempt to predict treatment burdens into the future. Temporal forecasting is generally hazardous and the estimation of valid confidence intervals, while crucial to the sensible use of forecasts, can be extremely difficult. Nevertheless, the utility of forecasts for public health planning needs far exceeds that of any historical data. The potential accuracy of forecasts could be evaluated by artificially removing the last year of known data and attempting to predict these values with data from the preceding years.

Chapter 11

Conclusions

Chapter 11

11. Conclusions

Public health decision-makers require accurate and timely information on disease-specific treatment burdens within a health system to allow the monitoring and planning of resource needs. A basic requirement is reliable national and sub-national data detailing the number of treatment events for a given disease or condition occurring at health facilities each month or year. In most African settings, this requirement is addressed with an HMIS that coordinates the routine acquisition of treatment records from health facilities and the transfer, compilation and analysis of these data through district, regional and national levels.

A perfect HMIS requires all health facilities to report promptly in all months, allowing a comprehensive quantification of treatment events through time and space across the health system. The reality of HMIS in Africa and elsewhere stands in marked contrast to this ideal. Typically, many facilities never report or report only intermittently resulting in spatially and temporally incomplete national data. Following several decades of donor investment in HMIS across Africa the incomplete nature of routine national reporting has shown little improvement.

Faced with poor data coverage, national treatment burdens are often estimated using rudimentary methods to account for missing values. The aim of this project was to develop a statistical approach to provide more reliable estimates of national outpatient treatment burdens. This project has focused on the Kenyan HMIS and has used the example of presumed malaria cases seen at government outpatient facilities around the country, a variable important to health-system planners.

Probabilistic models have been developed in this project to predict MC, the monthly number of malaria cases diagnosed at each facility where HMIS records were missing. These predictions were required at locations distributed in space and time and the modelling task was therefore framed as a space-time problem and addressed within the conceptual framework provided by geostatistics. Such an approach relies on the presence of autocorrelation in the variable of interest. This study found that raw HMIS data on malaria diagnoses can display substantial spatial autocorrelation when data from different facility types are considered independently. This justifies the use of spatial prediction techniques such as OK. This study also found, however, that these data display temporal autocorrelation, and that exploiting continuity in both the spatial and temporal domain using STOK results in considerably more accurate predictions. Furthermore, the heterogeneity of spatial patterns of malaria across Kenya suggests that assumptions of second-order stationarity of the RF model used in these predictions may be sub-optimal. An approach was developed that allowed space-time variograms to be estimated locally in order to more accurately represent second-order heterogeneities and this approach was found to result in marginally more accurate predictions, although its implementation was computationally demanding.

The number of outpatients diagnosed with malaria each month at a given health facility is a complex variable driven by a wide range of factors. A simple conceptual model was proposed that divides these factors into spatially-dependent determinants, principally caused by the heterogeneity of environmental conditions, and spatially-independent determinants, principally caused by factors specific to each facility and catchment population. Accounting for these non-spatial effects by standardising the MC variable by measures of these facility-specific factors can enhance the spatial continuity of the MC variable. Attempts to derive such measures directly using catchment population models highlighted the importance of using refined models that required detailed nationwide data. Because such data were unavailable, a different approach was devised that used TC data on the total monthly number of outpatients diagnosed for all conditions at each facility as a proxy measure of facility catchment size. By developing and testing two alternative prediction frameworks, this study showed that the use of TC data as a denominator to standardise MC data can account for much of the non-spatial variation present in MC. However, because TC values were unavailable at MC prediction locations, these values themselves required prediction and this introduced substantial

uncertainty into the resulting predictions of MC, negating much of the benefit of using a standardised numerator.

Two approaches were developed for providing measures of the uncertainty associated with predictions of MC at different levels of spatial and temporal aggregation. The first was a model-based geostatistical approach that involved a space-time adaptation of sequential Gaussian simulation. Evaluation of this uncertainty model found that it provided accurate measures of local and regional uncertainty. Because the model was complex and required intricate implementation, an alternative approach was also developed that could be more widely understood and implemented. This second approach used an internal validation procedure to obtain a sample of known prediction errors, and used these to infer the expected errors associated with the real predictions.

The predictive framework presented in this project allowed the incomplete Kenyan national HMIS database on outpatient malaria to be reconstructed and national treatment burdens to be estimated. The resulting estimate of the national annual treatment burdens for presumed outpatient malaria within the government sector was 6.8 million cases, with an expected margin of error of 1.3%. This figure is substantially different to the equivalent value of 7.6 million cases derived using rudimentary methods to account for the proportion of missing records. As such, this project has used geostatistics to provide results that are of direct use to public health decision-makers in Kenya.

Whilst the underlying problem of inadequate national health reporting systems can only be fully remedied by substantial and sustained investment in the infrastructure of these systems, the findings of this study and the predictive tools developed represent an important contribution that can be used to improve the reliability of information from HMIS and to enhance their utility as an evidence-base.

References

Abdulla,S., Gemperli,A., Mukasa,O., Schellenberg,J.R.M.A., Lengeler,C., Vounatsou,P., and Smith,T., 2005. Spatial effects of the social marketing of insecticide-treated nets on malaria morbidity. Tropical Medicine and International Health, 10, 11-18.

AbouZahr,C., and Boerma,T., 2005. Health information systems: The foundations of public health. Bulletin of the World Health Organization, 83, 578-583.

Afrane,Y.A., Lawson,B.W., Githeko,A.K., and Yan,G., 2005. Effects of microclimatic changes caused by land use and land cover on duration of gonotrophic cycles of *Anopheles gambiae* in western Kenya highlands. Journal of Medical Entomology, 42, 974-980.

Al Laham,H., Khoury,R., and Bashour,H., 2001. Reasons for under-reporting of notifiable diseases by Syrian paediatricians. Eastern Mediterranean Health Journal, 7, 590-596.

Albert,D.P., Gesler,W.M., and Levergood,B., 2000. Spatial Analysis, GIS, and Remote Sensing Applications in the Health Sciences. Ann Arbor Press, Michigan.

Amin,A.A., Marsh,V., Noor,A.M., Ochola,S.A., and Snow,R.W., 2003. The use of formal and informal curative services in the management of paediatric fevers in four districts of Kenya. Tropical Medicine and International Health, 8, 1143-1152.

Amin,A.A., and Snow,R.W., 2005. Brands, costs and registration status of antimalarial drugs in the Kenyan retail sector. Malaria Journal, 4.

Araghinejad,S., and Burn,D.H., 2005. Probabilistic forecasting of hydrological events using geostatistical analysis. Hydrological Sciences Journal, 50, 837-856.

Aron,J.L., and Schwartz,I.B., 1984. Seasonality and period-doubling bifurcations in an epidemic model. Journal of Theoretical Biology, 110, 665-679.

Ashraf,H., 2005. Countries need better information to receive development aid. Bulletin of the World Health Organization, 83, 565-566.

Avgerou,C., 1993. Information systems for development planning. International Journal of Information Management, 13, 260-273.

Bailey,T., and Gatrell,A., 1995. Interactive Spatial Data Analysis. Longman, Harlow.

Barat,L., Chipipa,J., Kolczak,M., Sukwa,T., 1999. Does the availability of blood slide microscopy for malaria at health centers improve the management of persons with fever in Zambia? American Journal of Tropical Medicine and Hygiene, 60, 1024-1030.

Beales,P.F., and Gilles,H.M., 2002. Rationale and technique of malaria control. In: Warrell,D.A., and Gilles,H.M., (Eds.), Essential Malariology. Arnold, London, pp. 107-190.

Beier,J.C., Copeland,R., Oyaro,C., Masinya,A., Odago,W.O., Oduor,S., Koech,D.K., and Roberts,C.R., 1990. *Anopheles gambiae* complex egg-stage survival in dry soil from larval development sites in western Kenya. Journal of the American Mosquito Control Association, 6, 105-109.

Berke,O., 2004. Exploratory disease mapping: Kriging the spatial risk function from regional count data. International Journal of Health Geographics, 3.

Bilonick,R.A., 1985. The space-time distribution of sulfate deposition in the northeastern United States. Atmospheric Environment - Part A General Topics, 19, 1829-1845.

Binka,F.N., and Adongo,P., 1997. Acceptability and use of insecticide impregnated bednets in northern Ghana. Tropical Medicine and International Health, 2, 499-507.

Bloland,P.B., 2001. Drug Resistance in Malaria. WHO, Geneva.

Bloland,P.B., Kachur,S.P., and Williams,H.A., 2003. Trends in antimalarial drug deployment in sub-Saharan Africa. Journal of Experimental Biology, 206, 3761-3769.

Bogaert,P., and Christakos,G., 1997. Spatiotemporal analysis and processing of thermometric data over Belgium. Journal of Geophysical Research D: Atmospheres, 102.

Braa,J., Heywood,A., and Shung King,M., 1997. District level information systems: two cases from South Africa. Methods of Information in Medicine 36, 115-121.

Brinkmann,U., and Brinkmann,A., 1995. Economic aspects of the use of impregnated mosquito nets for malaria control. Bulletin of the World Health Organization, 73, 651-658.

Brooker,S., Clarke,S., Njagi,J.K., Polack,S., Mugo,B., Estambale,B., Muchiri,E., Magnussen,P., and Cox,J., 2004. Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya. Tropical Medicine and International Health, 9, 757-766.

Brown,V., Issak,M.A., Rossi,M., Barboza,P., and Paugam,A., 1998. Epidemic of malaria in north-eastern Kenya. Lancet, 352, 1356-1357.

Carrat,F., and Valleron,A.J., 1992. Epidemiologic mapping using the 'kriging' method: Application to an influenza-like illness epidemic in France. American Journal of Epidemiology, 135, 1293-1300.

Casado,L.S., Rouhani,S., Cardelino,C.A., and Ferrier,A.J., 1994. Geostatistical analysis and visualization of hourly ozone data. Atmospheric Environment, 28, 2105-2118.

CBS Kenya, 2001a. 1999 population and housing census: counting our people for development. Volume I: Population distribution by administrative areas and urban centres. Central Bureau of Statistics, Ministry of Finance and Planning, Republic of Kenya, Nairobi.

CBS Kenya, 2001b. National Economic Survey 2000. Cental Bureau of Statistics, Ministry of Planning and National Development, Republic of Kenya, Nairobi.

CBS Kenya, 2003. Economic Survey. Cental Bureau of Statistics, Ministry of Planning and National Development, Republic of Kenya, Nairobi.

CBS Kenya, ORC Macro, 2003. Kenya Demographic and Health Survey 2003. Central Bureau of Statistics, Ministry of Health, and ORC Macro, Calverton, Maryland, USA.

CDC, 2000. Assessment of infectious disease surveillance - Uganda, 2000. Morbidity and Mortality Weekly Report, 49, 687-691.

Cham,M.K., Olaleye,B., D'Alessandro,U., Aikins,M., Cham,B., Maine,N., Williams,L.A., Mills,A., and Greenwood,B.M., 1997. The impact of charging for insecticide on the Gambian National Impregnated Bednet Programme. Health Policy and Planning, 12, 240-247.

Chaulagai,C.N., Moyo,C.M., Koot,J., Moyo,H.B., Sambakunsi,T.C., Khunga,F.M., and Naphini,P.D., 2005. Design and implementation of a health management information system in Malawi: issues, innovations and results. Health Policy and Planning, 20, 375-384.

Chilès,J.P., and Delfiner,P., 1999. Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons, Toronto.

Chilundo,B., Sundby,J., and Aanestad,M., 2004. Analysing the quality of routine malaria data in Mozambique. Malaria Journal, 3.

Christakos,G., and Lai,J.J., 1997. A study of the breast cancer dynamics in North Carolina. Social Science & Medicine, 45, 1503-1517.

Christakos,G., and Vyas,V.M., 1998. A composite space/time approach to studying ozone distribution over Eastern United States. Atmospheric Environment, 32, 2845-2857.

Christakos,G., Hristopulos,D.T., and Bogaert,P., 2000. On the physical geometry concept at the basis of space/time geostatistical hydrology. Advances in Water Resources, 23, 799-810.

Cibulskis,R.E., and Hiawalyer,G., 2002. Information systems for health sector monitoring in Papua New Guinea. Bulletin of the World Health Organization, 80, 752-758.

Costa,L.S., Nassi,C.D., Pinheiro,R.S., and Almeida,R.M.V.R., 2003. Accessibility of selected hospitals and medical procedures by means of aerial and transit network-based measures. Health Services Management Research, 16, 136-140.

Craig,M., Snow,R.W., and Le Sueur,D., 1999. A climate-based distribution model of malaria transmission in Sub-Saharan Africa. Parasitology Today, 15, 105-111.

Cressie,N., 1985. Fitting variogram models by weighted least squares. Mathematical Geology, 17, 563-586.

Cressie,N., 1993. Statistics for Spatial Data. Wiley, New York.

Cressie,N., and Huang,H.C., 1999. Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions. Journal of the American Statistical Association, 448, 1330-1340.

CSO Zambia, and ORC Macro, 2003. Zambia Demographic and Health Survey 2001-2002. Central Statistical Office, Central Board of Health, and ORC Macro, Calverton, Maryland, USA.

CSO Zimbabwe, and ORC Macro, 2006. Zimbabwe Demographic and Health Survey 1999. Central Statistical Office and ORC Macro, Calverton, Maryland, USA.

D'Agostino,V., Greene,E.A., Passarella,G., and Vurro,M., 1998. Spatial and temporal study of nitrate concentration in groundwater by means of coregionalization. Environmental Geology, 36, 285-295.

De Cesare,L., Myers,D.E., and Posa,D., 2001. Estimating and modeling space-time correlation structures. Statistics & Probability Letters, 51, 9-14.

De Cesare,L., Myers,D.E., and Posa,D., 2002. FORTRAN programs for space-time modeling. Computers and Geosciences, 28, 205-212.

De Iaco,S., Myers,D.E., and Posa,D., 2002. Space-time variograms and a functional form for total air pollution measurements. Computational Statistics and Data Analysis, 41, 311-328.

de Savigny,D., Mayombana,C., Mwageni,E., Masanja,H., Minhaj,A., Mkilindi,Y., Mbuya,C., Kasale,H., and Reid,G., 2004. Care-seeking patterns for fatal malaria in Tanzania. Malaria Journal, 3.

Deressa,W., Ali,A., and Enqusellassie,F., 2003. Self-treatment of malaria in rural communities, Butajira, southern Ethiopia. Bulletin of the World Health Organisation, 81, 261-268.

Derriennic,Y., 2003. Costing of Anti- malarial Drug Treatments in Ghana. The Partners for Health Reform*plus* Project, Abt Associates Inc., Bethesda, MD, USA.

Deutsch,C.V., and Journel,A.G., 1998. GSLIB: Geostatistical Software Library and User's Guide. Second edition. Oxford University Press, New York.

Diggle,P., Moyeed,R., Rowlingson,B., and Thomson,M.C., 2002. Childhood malaria in The Gambia: a case-study in model-based geostatistics. Applied Statistics, 51, 493-506.

Dimitrakopoulos,R., and Luo,X., 1994. Spatiotemporal modeling:covariances and ordinary kriging systems. In: Dimitrakopoulos,R., (Ed.), Geostatistics for the Next Century. KluwerAcademic Publishers, Dordrecht, pp. 88-93.

Douaik,A., Van Meirvenne,M., and Toth,T., 2005. Soil salinity mapping using spatio-temporal kriging and Bayesian maximum entropy with interval soft data. Geoderma, 128, 234-248.

Evans,T., and Stansfield,S., 2003. Health information in the new millennium: A gathering storm? Bulletin of the World Health Organization, 81.

Font,F., Alonso Gonzalez,M., Nathan,R., Kimario,J., Lwilla,F., Ascaso,C., Tanner,M., Menendez,C., and Alonso,P.L., 2001. Diagnostic accuracy and case management of clinical malaria in the primary health services of a rural area in south-eastern Tanzania. Tropical Medicine and International Health, 6, 423-428.

Foster,S., 1995. Treatment of malaria outside the formal health services. Journal of Tropical Medicine and Hygiene, 98, 29-34.

Fosu,G.B., 1994. Childhood morbidity and health services utilisation: Cross-national comparisons of user-related factors from DHS data. Social Science & Medicine, 38, 1209-1220.

Garner,P., and Graves,P.M., 2005. The Benefits of Artemisinin Combination Therapy for Malaria Extend Beyond the Individual Patient. PLoS Medicine, 2, e105.

Gemperli,A., Vounatsou,P., Kleinschmidt,I., Bagayoko,M., Lengeler,C., and Smith,T., 2004. Spatial Patterns of Infant Mortality in Mali: The Effect of Malaria Endemicity. American Journal of Epidemiology, 159, 64-72.

Gemperli,A., Vounatsou,P., Sogoba,N., and Smith,T., 2006. Malaria Mapping Using Transmission Models: Application to Survey Data from Mali. American Journal of Epidemiology, 163, 289-297.

Gething,P.W., Noor,A.M., Zurovac,D., Atkinson,P.M., Hay,S.I., Nixon,M.S., and Snow,R.W., 2004. Empirical modelling of government health service use by children with fevers in Kenya. Acta Tropica, 91, 227-237.

GFATM, 2005. Guide to the Global Funds's Policies on Procurement and Supply Management. The Global Fund to Fight Aids, Tuberculosis and Malaria.

Gimnig,J.E., Kolczak,M.S., Hightower,A.W., Vulule,J.M., Schoute,E., Kamau,L., Phillips-Howard,P.A., Ter Kuile,F.O., Nahlen,B.L., and Hawley,W.A., 2003. Effect of permethrin-treated bed nets on the spatial distribution of malaria vectors in western Kenya. American Journal of Tropical Medicine and Hygiene, 68, 115-120.

Girt,J.L., 1973. Distance to general medical practice and its effect on revealed ill health in a rural environment. Canadian Geographer, 17, 154-166.

Gladwin,J., Dixon,R.A., and Wilson,T.D., 2002. Rejection of an innovation: health information management training materials in east Africa. Health Policy and Planning, 17, 354-361.

Gladwin,J., Dixon,R.A., and Wilson,T.D., 2003. Implementing a new health management information system in Uganda. Health Policy and Planning, 18, 214-224.

Goodman,C., Kachur,S.P., Abdulla,S., Mwageni,E., Nyoni,J., Schellenberg,J.A., Mills,A., and Bloland,P., 2004. Retail supply of malaria-related drugs in rural Tanzania: Risks and opportunities. Tropical Medicine and International Health, 9, 655-663.

Goovaerts,P., Jacquez,G.M., and Greiling,D., 2005. Exploring Scale-Dependent Correlations Between Cancer Mortality Rates Using Factorial Kriging and Population-Weighted Semivariograms. Geographical Analysis, 37, 152-182.

Goovaerts,P., 1997. Geostatistics for Natural Resource Evaluation. Oxford University Press, New York.

Goovaerts,P., 2005a. Analysis and detection of health disparities using geostatistics and a space-time information system. Proceedings of GIS planet 2005, May 30 - June 2 2005, Estoril, Portugal.

Goovaerts,P., 2005b. Detection of spatial clusters and outliers in cancer rates using geostatistical filters and spatial neutral models. GeoENV V - Geostatistics for Environmental Applications, 149-160.

Goovaerts,P., 2005c. Geostatistical analysis of disease data: Estimation of cancer mortality risk from empirical frequencies using Poisson kriging. International Journal of Health Geographics, 4.

Goovaerts,P., 2006. Geostatistical analysis of disease data: Visualization and propagation of spatial uncertainty in cancer mortality risk using Poisson kriging and p-field simulation. International Journal of Health Geographics, 5.

Goovaerts,P., and Jacquez,G.M., 2004. Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: The case of lung cancer in Long Island, New York. International Journal of Health Geographics, 3.

Gove,S., 1998. Integrated management of childhood illness by outpatient health workers: Technical basis and overview. Bulletin of the World Health Organization, 75, 7-24.

Greenwood,B.M., 1989. The microepidemiology of malaria and its importance to malaria control. Transactions of the Royal Society of Tropical Medicine and Hygiene, 83, 25-29.

Guyatt,H.L., Corlett,S.K., Robinson,T.P., Ochola,S.A., Snow,R.W., 2002. Malaria prevention in highland Kenya: Indoor residual house-spraying vs. insecticide-treated bednets. Tropical Medicine and International Health, 7, 298-303.

Guyatt,H.L., Noor,A.M., Ochola,S.A., and Snow,R.W., 2004. Use of intermittent presumptive treatment and insecticide treated bed nets by pregnant women in four Kenyan districts. Tropical Medicine and International Health, 9, 255-261.

Guyatt,H.L., and Snow,R.W., 2004. The management of fevers in Kenyan children and adults in an area of seasonal malaria transmission. Transactions of the Royal Society of Tropical Medicine and Hygiene, 98, 111-115.

Haas,T.C., 1990. Kriging and automated variogram modeling within a moving window. Atmospheric Environment - Part A General Topics, 24 A, 1759-1769.

Haas,T.C., 1995. Local prediction of a spatio-temporal process with an application to wet sulfate deposition. Journal of the American Statistical Association, 90, 1189-1199.

Hamel,M.J., Odhacha,A., Roberts,J.M., and Deming,M.S., 2001. Malaria control in Bungoma District, Kenya: a survey of home treatment of children with fever, bednet use and attendance at antenatal clinics. Bulletin of the World Health Organisation, 79, 1014-1023.

Haslett,J., and Raftery,A.E., 1989. Space-time modelling with long-memory dependence: assessing Ireland's wind power resource. Applied Statistics, 38, 1-50.

Hawley,W.A., Phillips-Howard,P.A., Ter Kuile,F.O., Terlouw,D.J., Vulule,J.M., Ombok,M., Nahlen,B.L., Gimnig,J.E., Kariuki,S.K., Kolczak,M.S., and Hightower,A.W., 2003. Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in western Kenya. American Journal of Tropical Medicine and Hygiene, 68, 121-127.

Hay,S.I., 2000. An overview of remote sensing and geodesy for epidemiology and public health application. Advances in Parasitology, 47, 1-35.

Hay,S.I., Tucker,C.J., Rogers,D.J., Packer,M.J., 1996. Remotely sensed surrogates of meteorological data for the study of the distribution and abundance of arthropod vectors of disease. Annals of Tropical Medicine and Parasitology, 90, 1-19.

Hay,S.I., Packer,M.J., and Rogers,D.J., 1997. The impact of remote sensing on the study and control of invertebrate intermediate hosts and vectors for disease. International Journal of Remote Sensing, 18, 2899-2930.

Hay,S.I., Snow,R.W., and Rogers,D.J., 1998a. From predicting mosquito habitat to malaria seasons using remotely sensed data: Practice, problems and perspectives. Parasitology Today, 14, 306-313.

Hay,S.I., Snow,R.W., and Rogers,D.J., 1998b. Predicting malaria seasons in Kenya using multitemporal meteorological satellite sensor data. Transactions of the Royal Society of Tropical Medicine and Hygeine, 92, 12-20.

Hay,S.I., and Lennon,J.J., 1999. Deriving meteorological variables across Africa for the study and control of vector-borne disease: a comparison of remote sensing and spatial interpolation of climate. Tropical Medicine and International Health, 4, 58-71.

Hay,S.I., Rogers,D.J., Shanks,G.D., Myers,M.F., and Snow,R.W., 2001. Malaria early warning in Kenya. Trends in Parasitology, 17, 95-99.

Hay,S.I., Cox,J., Rogers,D.J., Randolph,S.E., Stern,D.I., Shanks,G.D., Myers,M.F., and Snow,R.W., 2002. Climate change and the resurgence of malaria in the East African highlands. Nature, 415, 905-909.

Hay,S.I., Were,E.C., Renshaw,M., Noor,A.M., Ochola,S.A., Olusanmi,I., Alipui,N., and Snow,R.W., 2003. Forecasting, warning, and detection of malaria epidemics: A case study. Lancet, 361, 1705-1706.

Hay,S.I., Guerra,C.A., and Snow,R.W., 2004. Determination of country populations at malaria risk of different endemicities: report on agreement to perform work (APW) for WHO/Roll Back Malaria. TALA Research Group, Department of Zoology, University of Oxford, Oxford.

Health Metrics Network, 2005a. Statistics save lives: strengthening country health information systems, draft. Geneva: Health Metrics Network. 2005.

Health Metrics Network, 2005b. Issues in Health Information. Available online at http://www.who.int/healthmetrics/library/en/. Accessed June 23rd 2006.

Health Metrics Network, 2005c. Issues in Health Information (1) National and Subnational Health Information Systems. Available online at http://www.who.int/healthmetrics/library/en/. Accessed June 23rd 2006.

Health Metrics Network, 2005d. Issues in Health Information (5) Household and Facility Surveys. Available online at http://www.who.int/healthmetrics/library/en/. Accessed June 23rd 2006.

Health Metrics Network, 2005e. Review of health information systems (IX) Uganda. Available online at http://www.who.int/healthmetrics/library/en/. Accessed June 23rd 2006.

Horton,R., 2005. The Ellison Institute: monitoring health, challenging WHO. The Lancet, 366, 179-181.

Hoshen,M.B., and Morse,A.P., 2004. A weather-driven model of malaria transmission. Malaria Journal, 3.

Host,G., Omre,H., and Switzer,P., 2004. Spatial Interpolation Errors for Monitoring Data. Journal of the American Statistical Association, 90, 853-861.

Howard,S.C., Omumbo,J., Nevill,C., Some,E.S., Donnelly,C.A., and Snow,R.W., 2000. Evidence for a mass community effect of insecticide-treated bednets on the incidence of malaria on the Kenyan coast. Transactions of the Royal Society of Tropical Medicine and Hygiene, 94, 357-360.

IDRC, 2002. Population and Health in Developing Countries: Population, Health, and Survival at INDEPTH Sites. International Development Research Centre, Ottowa.

Ijumba,J.N., Mosha,F.W., and Lindsay,S.W., 2002. Malaria transmission risk variations derived from different agricultural practices in an irrigated area of northern Tanzania. Medical and Veterinary Entomology, 16, 28-38.

INE Moçamique, and ORC Macro, 2003. Moçamique Inquérito Demográfico e de Saúde 2003. Instituto Nacional de Estatística, Ministério da Saúde and ORC Macro, Calverton, Maryland, USA.

Ingram,D.R., Clarke,D.R., and Murdie,R.A., 1978. Distance and the decision to visit an emergency department. Social Science and Medicine, 12, 55-62.

Isaaks,E.H., and Srivastava,R.H., 1989. Applied Geostatistics. Oxford University Press, New York.

James,G., 2001. Modern Engineering Mathematics, 3rd ed. Pearson Education Limited, Harlow.

Jost,G., Heuvelink,G.B.M., and Papritz,A., 2005. Analysing the space-time distribution of soil water storage of a forest ecosystem using spatio-temporal kriging. Geoderma, 128, 258-273.

Journel,A.G., and Huijbregts,C.J., 1978. Mining Geostatistics. Academic Press, New York.

Journel,A.G., and Rossi,M.E., 1989. When do we need a trend model in kriging? Mathematical Geology, 21, 715-739.

Justice,C.O., Townshend,J.R.G., Holben,B.N., and Tucker,C.J., 1985. Analysis of the phenology of global vegetation using meteorological satellite data. International Journal of Remote Sensing, 6, 1271-1318.

Kallander,K., Nsungwa-Sabiiti,J., and Peterson,S., 2004. Symptom overlap for malaria and pneumonia--policy implications for home management strategies. Acta Tropica, 90, 211-214.

Karanja,F.K., and Mutua,F.M., 2000. Reducing the impact of environmental emergencies through early warning and preparedness - the case of El Niño-southern oscillation (ENSO): Impacts of the 1997-98 El Niño event in Kenya. United Nations, Nairobi.

Keiser,J., Utzinger,J., and Singer,B.H., 2002. The potential of intermittent irrigation for increasing rice yields, lowering water consumption, reducing methane emissions, and controlling malaria in African rice fields. Journal of the American Mosquito Control Association, 18, 329-340.

Keller,A., 1991. Management information systems in maternal and child health/family planning programs: a multi-country analysis. Studies in Family Planning, 22, 19-30.

Kindermans,J., 2002. Changing national malaria treatment protocols in Africa: What is the cost and who will pay? RBM Partnership meeting on improving access to antimalarial treatment 30 September 2 October 2002. Medicins Sans Frontieres, Geneva.

Kleinschmidt,I., Bagayoko,M., Clarke,G.P.Y., Craig,M., and Le Sueur,D., 2000. A spatial statistical approach to malaria mapping. International Journal of Epidemiology, 29, 355-361.

Kleinschmidt,I., Omumbo,J., Briët,O., van de Giesen,N., Sogoba,N., Mensah,N.K., Windmeijer,P., Moussa,M., and Teuscher,T., 2001. An empirical malaria distribution map for West Africa. Tropical Medicine and International Health, 6, 779-786.

Kloos,H., 1990. Utilisation of selected hospitals, health centres and health stations in Central, Southern and Western Ethiopia. Social Science and Medicine, 31, 101-114.

Koenraadt,C.J., Paaijmans,K.P., Githeko,A.K., Knols,B.G., and Takken,W., 2003. Egg hatching, larval movement and larval survival of the malaria vector *Anopheles gambiae* in desiccating habitats. Malaria Journal, 2.

Kohli,S., Sahlén,K., Sivertun,Å., Löfman,O., Trell,E., and Wigertz,O., 1995. Distance from the primary health centre: a GIS method to study geographical access to health care. Journal of Medical Systems, 19, 425-436.

Korenromp,E.L., 2004. Malaria incidence estimates at country level for the year 2004. Report prepared for the Roll Back Malaria monitoring and evaluation group. Roll Back Malaria, World Health Organisation, Geneva.

Krige,D.G., 1951. A statistical approach to some mine valuations and allied problems at the Witwatersrand. Masters thesis submitted to University of Witwatersrand.

Kyriakidis,P.C., and Journel,A.G., 1999. Geostatistical space-time models: a review. Mathematical Geology, 31, 651-684.

Kyriakidis,C., and Journel,G., 2001. Stochastic modeling of atmospheric pollution: a spatial time-series framework. Part II: application to monitoring monthly sulfate deposition over Europe. Atmospheric Environment, 35, 2339-2348.

Kyriakidis,P.C., Miller,N.L., and Kim,J., 2004. A spatial time series framework for simulating daily precipitation at regional scales. Journal of Hydrology, 297, 236-255.

Law,D.C.G., Serre,M.L., Christakos,G., Leone,P.A., and Miller,W.C., 2004. Spatial analysis and mapping of sexually transmitted diseases to optimise intervention and prevention strategies. Sexually Transmitted Infections, 80, 294-299.

Lawson,A.B., 2001. Statistical Methods in Spatial Epidemiology. John Wiley and Sons, Ltd., Chichester.

Lengeler,C., and Snow,R.W., 1996. From efficacy to effectiveness: Insecticide-treated bednets in Africa. Bulletin of the World Health Organization, 74, 325.

Lengeler,C., Smith,T., and Armstrong Schellenberg,J.R.M., 1997. Focus on the effect of bednets on malaria morbidity and mortality. Parasitology Today, 13, 123-124.

Lippeveld,T., Sauerborn ,R., and Bodart,C., 2000. Design and implementation of health information systems. World Health Organisation.

Littlejohns,P., Wyatt,J.C., and Garvican,L., 2003. Evaluating computerised health information systems: hard lessons still to be learnt. BMJ, 326, 860-863.

Lloyd,C.D., 2002. Increasing the accuracy of predictions of monthly precipitation in Great Britain using kriging with an external drift. In: Foody,G.M., and Atkinson,P.M., (Eds.), Uncertainty in Remote Sensing and GIS. John Wiley and Sons, Chichester, pp. 243-267.

Lloyd,C.D., and Atkinson,P.M., 2002. Non-stationary approaches for mapping terrain and assessing prediction uncertainty. Transactions in GIS, 6, 17-30.

Lophaven,S., and Carstensen,J., and Rootzen,H., 2006. Stochastic modelling of dissolved inorganic nitrogen in space and time. Ecological Modelling, 193, 467-478.

Loveridge,B.W., Henner,J.R., and Lee,F.C., 2003. Accurate clinical diagnosis of malaria in a postflood epidemic: A field study in Mozambique. Wilderness and Environmental Medicine Journal, 14, 17-19.

Macfarlane,S.B., 2005. Harmonizing health information systems with information systems in other social and economic sectors. Bulletin of the World Health Organization, 83, 590-596.

Mahapatra,P., and Chalapati Rao,P.V., 2001. Cause of death reporting systems in India: A performance analysis. National Medical Journal of India, 14, 154-162.

Makemba,A.M., Winch,P.J., Kamazima,S.R., Makame,V.R., Sengo,F., Lubega,P.B., Minjas,J.N., and Shiff,C.J., 1995. Community-based sale, distribution and insecticide impregnation of mosquito nets in Bagamoyo District, Tanzania. Health Policy and Planning, 10, 50-59.

Marsh,V.M., Mutemi,W., Some,E.S., Haaland,A., and Snow,R.W., 1996. Evaluating the community education programme of an insecticide-treated bed net trial on the Kenyan coast. Health Policy and Planning, 11, 280-291.

Marsh,K., 1998. Malaria disaster in Africa. Lancet, 352, 924-924.

Marsh,V.M., Mutemi,W.M., Willetts,A., Bayah,K., Were,S., Ross,A., and Marsh,K., 2004. Improving malaria home treatment by training drug retailers in rural Kenya. Tropical Medicine and International Health, 9, 451-460.

Matheron,G., 1971. The Theory of Regionalized Variables and Its Applications. Ecole Nationale Supérieure des Mines de Paris, Fontainebleau.

Mathers,C.D., Fat,D.M., Inoue,M., Rao,C., and Lopez,A.D., 2005. Counting the dead and what they died from: An assessment of the global status of cause of death data. Bulletin of the World Health Organization, 83, 171-177.

Mbogo,C.N.M., Baya,N.M., Ofulla,A.V.O., Githure,J.I., and Snow,R.W., 1996. The impact of permethrin-impregnated bednets on malaria vectors of the Kenyan coast. Medical and Veterinary Entomology, 10, 251-259.

McCombie,S.C., 1996. Treatment seeking for malaria: a review of recent research. Social Science and Medicine, 43, 933-945.

McCombie,S.C., 2002. Self-treatment for malaria: the evidence and methodological issues. Health Policy and Planning, 17, 333-344.

Meiring,W., Guttorp,P., and Sampson,P.D., 1998. Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. Environmental and Ecological Statistics, 5, 197-222.

Minakawa,N., Munga,S., Atieli,F., Mushinzimana,E., Zhou,G., Githeko,A.K., and Yan,G., 2005. Spatial distribution of anopheline larval habitats in Western Kenyan highlands: Effects of land cover types and topography. American Journal of Tropical Medicine and Hygiene, 73, 157-165.

MoH Kenya, 1994. Health Policy Framework. Ministry of Health, Republic of Kenya, Nairobi.

MoH Kenya, 1999. The National Health Sector Strategic Plan 1999-2004. Ministry of Health, Republic of Kenya, Nairobi.

MoH Kenya, 2000. Strengthening Health Information Systems (HIS) Towards Health Management Information System (HMIS). Ministry of Health, Republic of Kenya, Nairobi.

MoH Kenya, 2001a. Health Management Information Systems: report for the 1996 to 1999 period. Ministry of Health, Republic of Kenya, Nairobi.

MoH Kenya, 2001b. National Malaria Strategy 2001-2010. Division of Malaria Control, Ministry of Health, Republic of Kenya, Nairobi.

MoH Kenya, 2003. Kenya National Health Accounts 2001-2002. Ministry of Health, Republic of Kenya, Nairobi.

MoH Kenya, 2004. Position Paper by Subgroup on Demand, Costs, Financing, Procurement and Distribution. Drug Policy Technical Working Group, Division of Malaria Control, Ministry of Health, Kenya, Nairobi.

MoH Kenya, 2005a. The Second National Health Sector Strategic Plan of Kenya (NHSSP II 2005-10). Ministry of Health, Republic of Kenya, Nairobi.

MoH Kenya, 2005b. The Second National Health Sector Strategic Plan of Kenya (NHSSP II 2005-10): Reversing the Trends. Ministry of Health, Republic of Kenya, Nairobi.

MoH Kenya, 2005c. Transition Plan for Implementation of Artemisinin-Based Combination Therapy (ACT) Malaria Treatment Policy in Kenya. Ministry of Health, Republic of Kenya, Nairobi.

Molyneux,C.S., Mung'ala-Odera,V., Harpham,T., and Snow,R.W., 1999. Maternal responses to childhood fevers: a comparison of rural and urban residents in coastal Kenya. Tropical Medicine and International Health, 4, 836-845.

Molyneux,C.S., Murira,G., Masha,J., and Snow,R.W., 2002. Intra-household relations and treatment decision-making for childhood illness: A Kenyan case study. Journal of Biosocial Science, 34, 109-131.

Morrill,R.L., and Earickson,R., 1968. Variation in the character and use of Chicago area hospitals. Health Services Research, 3, 224-238.

Morris,S.S., Black,R.E., and Tomaskovic,L., 2003. Predicting the distribution of under-five deaths by cause in countries without adequate vital registration systems. International Journal of Epidemiology, 32, 1041-1051.

Mouchet,J., Manguin,S., Sircoulon,J., Laventure,S., Faye,O., Onapa,A., Carnevale,P., Julvez,J., and Fontenille,D., 1998. Evolution of malaria in Africa for the past 40 years: impact of climatic and human factors. Journal of the American Mosquito Control Association, 14, 121-130.

Müller,I., Smith,T., Mellor,S., Rare,L., and Genton,B., 1998. The effect of distance from home on attendance at a small rural health centre in Papua New Guinea. International Journal of Epidemiology, 27, 878-884.

Muller,O., Traore,C., Kouyate,B., Ye,Y., Frey,C., Coulibaly,B., and Becher,H., 2006. Effects of insecticide-treated bednets during early infancy in an African area of intense malaria transmission: A randomized controlled trial. Bulletin of the World Health Organization, 84, 120-126.

Munga,S., Minakawa,N., Zhou,G., Mushinzimana,E., Barrack,O.O., Githeko,A.K., and Yan,G., 2006. Association between land cover and habitat productivity of malaria vectors in western Kenyan highlands. The American Journal of Tropical Medicine and Hygiene. 74, 69-75.

Murray,C.J.L., Mathers,C.D., and Salomon,J.A., 2003. Towards evidence-based public health. In: Murray,C.J.L., Evans,D.B., (Eds.), Health Systems Performance Assessment: Debates, Methods and Empiricism. World Health Organisation, Geneva, pp. 715-726.

Murray,C.J.L., Lopez,A.D., and Wibulpolprasert,S., 2004. Monitoring global health: Time for new solutions. British Medical Journal, 329, 1096-1100.

Mutemwa,R.I., 2006. HMIS and decision-making in Zambia: re-thinking information solutions for district health management in decentralized health systems. Health Policy and Planning, 21, 40-52.

Mwenesi,H., Harpham,T., and Snow,R.W., 1995. Child malaria treatment practices among mothers in Kenya. Social Science & Medicine, 40, 1271-1277.

NBS Tanzania, and ORC Macro, 2005. Tanzania Demographic and Health Survey 2004-05. National Bureau of Statistics and ORC Macro, Dar es Salaam, Tanzania.

NCAPD/MOH/CBS Kenya, and ORC Macro, 2005. Kenya Service Provision Assessment Survey 2004. National Coordinating Agency for Population and Development, Ministry of Health, Central Bureau of Statistics, and ORC Macro, Nairobi.

Nevill,C.G., Some,E.S., Mung'ala,V.O., Mutemi,W., New,L., Marsh,K., Lengeler,C., and Snow,R.W., 1996. Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast. Tropical Medicine and International Health, 1, 139-146.

Noor,A.M., 2005. Developing spatial models of health service and utilisation to define health equity in Kenya. PhD. thesis submitted to The Open University.

Noor,A.M., Zurovac,D., Hay,S.I., Ochola,S.A., and Snow,R.W., 2003. Defining equity in physical access to clinical services using geographical information systems as part of malaria planning and monitoring in Kenya. Tropical Medicine and International Health, 8, 917-926.

Noor,A.M., Gikandi,P.W., Hay,S.I., Muga,R.O., and Snow,R.W., 2004. Creating spatially defined databases for equitable health service planning in low-income countries: the example of Kenya. Acta Tropica, 91, 239-251.

Noor,A.M., Amin,A.A., Gething,P.W., Atkinson,P.M., Hay,S.I., and Snow,R.W., 2006. Modelling distances travelled to government health services in Kenya. Tropical Medicine and International Health, 11, 188-196.

NSEO Eritrea, and ORC Macro, 2003. Eritrea Demographic and Health Survey 2002. National Statistics and Evaluation Office and ORC Macro, Calverton, Maryland, USA.

NSO Malawi, and ORC Macro, 2005. Malawi Demographic and Health Survey 2004. National Statistical Office and ORC Macro, Calverton, Maryland, USA.

Nunes,C., and Soares,A., 2005. Geostatistical space-time simulation model for air quality prediction. Environmetrics, 16, 393-404.

Olea,R.A., 1991. Geostatistical Glossary and Multilingual Dictionary. Oxford university Press, New York.

Oliver,M.A., Muir,K.R., Webster,R., Parkes,S.E., Cameron,A.H., Stevens,M.C.G., and Mann,J.R., 1992. A geostatistical approach to the analysis of pattern in rare disease. Journal of Public Health Medicine, 14, 280-289.

Oliver,M.A., Webster,R., Lajaunie,C., Muir,K.R., Parkes,S.E., Cameron,A.H., Stevens,M.C.G., and Mann,J.R., 1998. Binomial cokriging for estimating and mapping the risk of childhood cancer. IMA Journal of Mathematics Applied in Medicine and Biology, 15, 279-297.

Omumbo,J.A., Hay,S.I., Goetz,S.J., Snow,R.W., and Rogers,D.J., 2002. Updating historical maps of malaria transmission duration in east Africa using remote sensing. Photogrammetric Engineering and Remote Sensing, 68, 161-166.

Omumbo,J.A., Hay,S.I., Snow,R.W., Tatem,A.J., and Rogers,D.J., 2005. Modelling malaria risk in East Africa at high-spatial resolution. Tropical Medicine and International Health, 10, 557-566.

ONAPO Rwanda, and ORC Macro, 2001. Enquête Démographique et de Santé, Rwanda 2000. Ministère de la Santé, Office National de la Population et ORC Macro, Calverton, Maryland, USA.

Onokerhoraye,A.G., 1999. Access and utilisation of modern health care facilities in the petroleum-producing region of Nigeria: the case of Bayelsa state. Harvard School of Public Health, Boston.

Oranga,H.M., Nordberg,E., 1995. A longitudinal health interview survey in rural Kenya: potentials and limitations for local planning. East African Medical Journal, 72, 241-247.

Pardo-Igúzquiza,E., 1999. VARFIT: A fortran-77 program for fitting variogram models by weighted least squares. Computers and Geosciences, 25, 251-261.

PARIS21, 2004. Making the Case:National Strategy for the Development of Statistics. The Partnership in Statistics for Development in the 21st Century. Available online at http://www.paris21.org/documents/1406.pdf. Accessed 23rd June 2006.

Patz,J.A., Strzepek,K., Lele,S., Hedden,M., Greene,S., Noden,B., Hay,S.I., Kalkstein,L., and Beier,J.C., 1998. Predicting key malaria transmission factors, biting and entomological inoculation rates, using modelled soil moisture in Kenya. Tropical Medicine and International Health, 3, 818-827.

Patz,J.A., and Olson,S.H., 2006. Malaria risk and temperature: Influences from global climate change and local land use practices. Proceedings of the National Academy of Sciences of the United States of America, 103, 5635-5636.

Perkins,B.A., Zucker,J.R., Otieno,J., Jafari,H.S., Paxton,L., Redd,S.C., Nahlen,B.L., Schwartz,B., Oloo,A.J., Olango,C., Gove,S., Campbell,C.C., 1998. Evaluation of an algorithm for integrated management of childhood illness in an area of Kenya with high malaria transmission. Bulletin of the World Health Organization, 75, 33-42.

Perry,B., and Gesler,W., 2000. Physical access to primary health care in Andean Bolivia. Social Science and Medicine, 50, 1177-1188.

RBM, 2000. Framework for Monitoring Progress and Evaluating Outcomes and Impact. Roll Back Malaria/World Health Organisation, Geneva.

Redd,S.C., Redd,S.C., Bloland,P.B., Campbell,C.C., Kazembe,P.N., Tembenu,R., and Patrick,E., 1992. Usefulness of clinical case-definitions in guiding therapy for African children with malaria or pneumonia. The Lancet, 340, 1140-1143.

Redd,S.C., Kazembe,P.N., Luby,S.P., Nwanyanwu,O., Hightower,A.W., Ziba,C., Wirima,J.J., Chitsulo,L., Franco,C., and Olivar,M., 1996. Clinical algorithm for treatment of *Plasmodium falciparum* malaria in children. Lancet, 347, 223-227.

Reiter,P., Thomas,C.J., Atkinson,P.M., Hay,S.I., Randolph,S.E., Rogers,D.J., Shanks,G.D., Snow,R.W., and Spielman,A., 2004. Global warming and malaria: a call for accuracy. The Lancet Infectious Diseases, 4, 323-324.

Ribeiro,J.M.C., Seulu,F., Abose,T., Kidane,G., and Teklehaimanot,A., 1996. Temporal and spatial distribution of anopheline mosquitos in an Ethiopian village: Implications for malaria control strategies. Bulletin of the World Health Organization, 74, 299-305.

Rodriguez-Iturbe,I., and Mejia,J.M., 1974. Design of rainfall networks in time and space. Water Resources Research, 10, 713-728.

Rogers,D.J., Randolph,S.E., Snow,R.W., and Hay,S.I., 2002. Satellite imagery in the study and forecast of malaria. Nature, 415, 710-715.

Rouhani,S., and Myers,D.E., 1990. Problems in space-time kriging of geohydrological data. Mathematical Geology, 22, 611-623.

RPM plus, 2005. Changing Malaria Treatment Policy to Artemisinin-Based Combinations: An Implementation Guide Submitted to the U.S. Agency for International Development by the RPM Plus Program. Management Sciences for Health, Arlingotn, VA, USA.

Rudan,I., Lawn,J., Cousens,S., Rowe,A.K., Boschi-Pinto,C., Tomaskovic,L., Mendoza,W., Lanata,C.F., Roca-Feltrer,A., Carneiro,I., Schellenberg,J.A., polasek,O., Weber,M., Bryce,J., Morris,S.S., Black,R.E., and Campbell,H., 2005. Gaps in policy-relevant information on burden of disease in children: A systematic review. Lancet, 365, 2031-2040.

Ruebush,T.K., Kern,M.K., Campbell,C.C., and Oloo,A.J., 1995. Self treatment of malaria in a rural area of western Kenya. Bulletin of the World Health Organisation, 73, 229-236.

Ryan,G.W., 1998. What do sequential behavioral patterns suggest about the medical decision-making process?: Modeling home case management of acute illnesses in a rural Cameroonian village. Social Science & Medicine, 46, 209-225.

Saito,H., Goovaerts,P., 2000. Geostatistical interpolation of positively skewed and censored data in a dioxin-contaminated site. Environmental Science and Technology, 34, 4228-4235.

Sankoh,O., De Savigny,D., and Binka,F., 2004. INDEPTH Network: Generating Empirical Population and Health Data in Resource-constrained Countries in the Developing World for assessing the Millennium Development Goals. Paper presented at the Global Forum for Health Research, Forum 8, Mexico City, November 2004.

Schellenberg,J.A., Newell,J.N., Snow,R.W., Mung'ala,V., Marsh,K., Smith,P.G., and Hayes,R.J., 1998. An analysis of the geographical distribution of severe malaria in children in Kilifi District, Kenya. International Journal of Epidemiology, 27, 323-329.

Scott,C., 2005. Measuring up to the Measurement Problem: The Role of Statistics in Evidence Based Policy-making. Partnership in Statistics for Development in the 21st Century (PARIS 21).

Setel,P.W., Sankoh,O., Rao,C., Velkoff,V.A., Mathers,C., Gonghuan,Y., Hemed,Y., Jha,P., and Lopez,A.D., 2005. Sample registration of vital events with verbal autopsy: A renewed commitment to measuring and monitoring vital statistics. Bulletin of the World Health Organization, 83, 611-617.

Shanks,G.D., Biomndo,K., Hay,S.I., and Snow,R.W., 2000. Changing patterns of clinical malaria since 1965 among a tea estate population located in the Kenyan highlands. Transactions of the Royal Society of Tropical Medicine and Hygeine, 94, 253-255.

Shannon,G.W., Bashshur,R.L., and Metzner,C.A., 1969. The concept of distance as a factor in accessibility and utilisation of health care. Medical Care Review, 26, 143-161.

Shannon,G.W., Skinner,J.L., and Bashshur,R.L., 1973. Time and distance: the journey for medical care. International Journal of Health Services, 3, 237-244.

Shretta,R., Omumbo,J., Rapuoda,B., and Snow,R.W., 2000. Using evidence to change antimalarial drug policy in Kenya. Tropical Medicine and International Health, 5, 755-764.

Sibai,A.M., 2004. Mortality certification and cause-of-death reporting in developing countries. Bulletin of the World Health Organization, 82.

Silvi,J., 2003. On the estimation of mortality rates for countries of the Americas. Epidemiological bulletin, 24, 1-5.

Slack,A., Cumming,J., Maré,D., and Timmins,J., 2002. Variations in secondary care utilisation and geographic access. Motu/Health Services Research Centre, University of Wellington, New Zealand.

Smith,T., Schellenberg,J.A., and Hayes,R., 1994. Attributable fraction estimates and case definitions for malaria in endemic areas. Statistics in Medicine, 13, 2345-2358.

Snepvangers,J.J.J.C., Heuvelink,G.B.M., and Huisman,J.A., 2003. Soil water content interpolation using spatio-temporal kriging with external drift. Geoderma, 112, 253-271.

Snow,R.W., Peshu,N., Forster,D., Mwenesi,H., and Marsh,K., 1992a. The role of shops in the treatment and prevention of childhood malaria on the coast of Kenya. Transactions of the Royal Society of Tropical Medicine and Hygiene, 86, 237-239.

Snow,R.W., Winstanley,M.T., Marsh,V.M., Newton,C.R.J.C., Waruiru,C., Mwangi,I., Winstanley,P.A., Marsh,K., Snow,R.W., and Forster,D., 1992b. Childhood deaths in Africa: uses and limitations of verbal autopsies. The Lancet, 340, 351-355.

Snow,R.W., Armstrong Schellenberg,J.R.M., Peshu,N., Forster,D., Newton,C.R.J.C., Winstanley,P.A., Mwangi,I., Waruiru,C., Warn,P.A., Newbold,C., and Marsh,K., 1993. Periodicity and space-time clustering of severe childhood malaria on the coast of Kenya. Transactions of the Royal Society of Tropical Medicine and Hygeine, 87, 386-390.

Snow,R.W., Omumbo,J.A., Lowe,B., Molyneux,C.S., Obiero,J.O., Palmer,A., Weber,M.W., Pinder,M., Nahlen,B., Obonyo,C., Gupta,S., and Marsh,K., 1997. Relation between severe malaria morbidity in children and level of *Plasmodium falciparum* transmission in Africa. Lancet, 349, 1650-1654.

Snow,R.W., Gouws,E., Omumbo,J.A., Craig,M., Tanser,F.C., Le Sueur,D., and Ouma,J., 1998. Models to predict the intensity of *Plasmodium falciparum* transmission: applications to the burden of disease in Kenya. Transactions of the Royal Society of Tropical Medicine and Hygeine, 92, 601-606.

Snow,R.W., Craig,M., Deichmann,U., and Marsh,K., 1999a. Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population. Bulletin of the World Health Organisation, 77, 624-640.

Snow,R.W., McCabe,E., Mbogo,C.N.M., Molyneux,C.S., Some,E.S., Mung'ala,V.O., and Nevill,C.G., 1999b. The effect of delivery mechanisms on the uptake of bed net re-impregnation in Kilifi District, Kenya. Health Policy and Planning, 14, 18-25.

Snow,R.W., and Gilles,H.M., 2002. The epidemiology of malaria. In: Warrell,D.A., and Gilles,H.M., (Eds.), Essential Malariology. Arnold, London, pp. 85-106.

Snow,R.W., Eckert,E., and Teklehaimanot,A., 2003a. Estimating the needs for artesunate-based combination therapy for malaria case-management in Africa. Trends in Parasitology, 19, 363-369.

Snow,R.W., Craig,A., Newton,C.R.J.C., and Steketee,R.W., 2003b. The public health burden of *Plasmodium falciparum* malaria in Africa: Deriving the numbers. Working Paper No. 11, Disease Control Priorities Project. Fogarty International Center, National Institutes of Health, Bethesda, Maryland.

Snow,R.W., Noor,A.M., Gikandi,P.W., Tetteh,G., and Ochola,S.A., 2003c. Modeling the anti-malarial drug requirements for the Kenyan Government's formal health sector using imperfect data. Ministry of Health, Republic of Kenya., Nairobi.

Snow,R.W., Guerra,C.A., Noor,A.M., Myint,H.Y., and Hay,S.I., 2005. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. Nature, 434, 214-217.

Stock,R., 1983. Distance and the utilisation of health facilities in rural Nigeria. Social Science and Medicine, 17, 563-570.

Tanser,F.C., Hosegood,V., Benzler,J., and Solarsh,G., 2001. New approaches to spatially analyse primary health care usage patterns in rural South Africa. Tropical Medicine and International Health, 6, 826-838.

Tanser,F., Gijsbertsen,B., and Herbst,K., 2006. Modelling and understanding primary health care accessibility and utilisation in rural South Africa: An exploration using a geographical information system. Social Science & Medicine, 63, 691-705.

Ter Kuile,F.O., Terlouw,D.J., Phillips-Howard,P.A., Hawley,W.A., Friedman,J.F., Kolczak,M.S., Kariuki,S.K., Shi,Y.P., Kwena,A.M., Vulule,J.M., and Nahlen,B.L., 2003. Impact of permethrin-treated bed nets on malaria and all-cause morbidity in young children in an area of intense perennial malaria transmission in Western Kenya: Cross-sectional survey. American Journal of Tropical Medicine and Hygiene, 68, 100-107.

Thomson,M.C., Connor,S.J., Milligan,P., and Flasse,S.P., 1996. The ecology of malaria - As seen from Earth-observation satellites. Annals of Tropical Medicine and Parasitology, 90, 243-264.

Thomson,M.C., Connor,S.J., Milligan,P., and Flasse,S.P., 1997. Mapping malaria risk in Africa: What can satellite data contribute? Parasitology Today, 13, 313-318.

Trape,J.F., Pison,G., Preziosi,M.P., Enel,C., du Lou,A.D., Delaunay,V., Samb,B., Lagarde,E., Molez,J.F., and Simondon,F., 1998. Impact of chloroquine resistance on malaria mortality. Comptes Rendus de l'Academie des Sciences - Series III - Sciences de la Vie, 321, 689-697.

Trape,J.F., 2001. The public health impact of chloroquine resistance in Africa. American Journal of Tropical Medicine and Hygiene, 64, 12-17.

Tucker,M.R., and Sear,C.B., 2001. A comparison of Meteosat rainfall estimation techniques in Kenya. Meteorological Applications, 8, 107-117.

Twigg,L., 1990. Health based geographical information systems: their potential examined in the light of existing data sources. Social Science and Medicine, 30, 143-155.

UBOS Uganda, and ORC Macro, 2001. Uganda Demographic and Health Survey 2000-2001. Uganda Bureau of Statistics and ORC Macro, Calverton, Maryland, USA.

UN, 2000. United Nations millennium declaration. United Nations General Assembly, A/RES/55/2.

UN, 2001. Road map towards the implementation of the UN Millennium Declaration. United Nations General Assembly, A/56/326.

UN, 2005. The Millenium Development Goals Report 2005. United Nations, New York.

UN, 2006. Development Indicators. United Nations Economic and Social Council, E/CN.3/2006/14.

UNICEF, 1999. End-decade assessment - indicators for assessing progress globally, Executive Directive CF/EXD/1999-03, 23 April 1999. United Nations Children's Fund.

Üstün,T.B., Chatterji,S., Mechbal,A., and Murray,C.J.L., 2003. The World Health Surveys. In: Murray,C.J.L., and Evans,D.B., (Eds.), Health Systems Performance Assessment: Debates, Methods and Empiricism. World Health Organisation, Geneva, pp. 797-808.

Vanderlinden,K., Ordonez,R., Polo,M.J., and Giraldez,J.V., 2006. Mapping residual pyrite after a mine spill using non co-located spatiotemporal observations. Journal of Environmental Quality, 35, 21-36.

Warrell,D.A., 2002. Clinical features of malaria. In: Warrell,D.A., and Gilles,H.M., (Eds.), Essential Malariology. Arnold, London, pp. 191-205.

Warrell,D.A., Watkins,W.M., and Winstanley,P.A., 2002. Treatment and Prevention of Malaria. In: Warrell,D.A., and Gilles,H.M., (Eds.), Essential Malariology. Arnold, London, pp. 268-312.

Webster,R., Oliver,M.A., Muir,K.R., and Mann,J.R., 1994. Kriging the local risk of a rare disease from a register of diagnoses. Geographical Analysis, 26, 168-185.

Wernsdorfer,W.H., 1994. Epidemiology of drug resistance in malaria. Acta Tropica, 56, 143-156.

White,N.J., 1998. Preventing antimalarial drug resistance through combinations. Drug Resistance Updates, 1, 3-9.

White,N.J., Nosten,F., Looareesuwan,S., Watkins,W.M., Marsh,K., Snow,R.W., Kokwaro,G., Ouma,J., Hien,T.T., Molyneux,M.E., Taylor,T.E., Newbold,C.I., Ruebush II,T.K., Danis,M., Greenwood,B.M., Anderson,R.M., and Olliaro,P., 1999. Averting a malaria disaster. Lancet, 353, 1965-1967.

WHO, 1993. Guidelines for the Developmentof Health Management Information Systems. World Health Organisation Regional Office for the Western Pacific, Manila.

WHO, 1995. Estimating Drug Requirements: A Practical Manual. World Health Organisation Action Programme on Essential Drugs, Geneva.

WHO, 1997. Integrated Management of Childhood Illnesses Adaptation Guide. Part 2. C. Technical basis for adapting clinical guidelines, feeding recommendations, and local terms. World Health Organisation, Geneva.

WHO, 1988. Estimating Drug Requirements: A Practical Manual. World Health Organisation Action Programme on Essential Drugs, Geneva.

WHO, 1999. Malaria in Kenya: disease outbreak reported 14th July 1999. Communicable Disease Surveillance and Response, World Health Organisations.

WHO, 2000a. World Health Report: Health Systems: Improving Performance. World Health Organisation, Geneva.

WHO, 2000b. An integrated approach to communicable disease surveillance. Weekly Epidemiological Record 75, 1-8. 2000a.

WHO, 2001a. Antimalarial Drug Combination Therapy. Report of a WHO Technical Consultation. World Health Organization, Geneva.

WHO, 2001b. Assessment of the national communicable disease surveillance and response system, Ethiopia. World Health Organization Weekly epidemiological record. 76, 9-16.

WHO, 2003a. Access to Antimalarial Medicines: Improving the Affordability and Financing of Artemisinin-Based Combination Therapies. Malaria Control Department and Essential Drugs and Medicines Policy Department, World Health Organisation, Geneva.

WHO, 2003b. The world health report 2003: shaping the future. World Health Organisation, Geneva.

WHO, 2004a. Developing Health Management Information Systems: A Practical Guide for Developing Countries. The World Health Organisation Regional Office for the Western Pacific, Manila.

WHO, 2004b. International Statistical Classification of Diseases and Health Related Problems. World Health Organisation, Geneva.

WHO, 2005a. World Malaria Report 2005. World Health Organisation, Geneva.

WHO, 2005b. Malaria Control Today: Current WHO Recommendations. Roll Back Malaria Department, World Health Organization, Geneva.

WHO, 2006. Procurement of Artemether-Lumefantrine (Coartem®) through WHO. World Health Organisation, Geneva.

WHO/AFRO, 2003. African regional consultation on health systems performance assessment. In: Murray,C.J.L., Evans,D.B., (Eds.), Health Systems Performance Assessment: Debates, Methods and Empiricism. World Health Organisation, Geneva.

WHO/SEARO, 2003. South-East Asian Regional Consultation on Health Systems Performance Assessment. In: Murray,C.J.L., Evans,D.B., (Eds.), Health Systems Performance Assessment: Debates, Methods and Empiricism. World Health Organisation, Geneva.

World Economic Forum, 2006. Global Governance Initiative, 3rd Annual Report (2006).

Zurovac,D., Midia,B., Ochola,S.A., Barake,Z., and Snow,R.W., 2002. Evaluation of Malaria Case Management of Sick Children Presenting in Outpatient Departments in Government Health Facilities in Kenya. Division of malaria Control, Ministry of Health, Republic of Kenya, Nairobi.

Zurovac,D., Midia,B., Ochola,S.A., English,M., and Snow,R.W., 2006. Microscopy and outpatient malaria case management among older children and adults in Kenya. Tropical Medicine and International Health, 11, 432-440.

Zwarenstein,M., Krige,D., and Wolff,B., 1991. The use of a geographical information-system for hospital catchment-area research in Natal Kwazulu. South African Medical Journal, 80, 497-500.