LANCASTER UNIVERSITY

Geostatistical Design and Analysis for Estimating Local Variations in Malaria Disease Burden.

by

Michael Give Chipeta

A thesis submitted in partial fulfilment for the degree of Doctor of Philosophy

in the

Faculty of Health and Medicine
Lancaster Medical School

October, 2016

Declaration of Authorship

I, **Michael Give Chipeta**, declare that this thesis titled, "*Geostatistical Design and Analysis for Estimating Local Variations in Malaria Disease Burden*" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signature:________________________ **Date:**______________________________

“Everything is related to everything else, but near things are more related than distant things.”

Waldo Rudolph Tobler

LANCASTER UNIVERSITY

Abstract

Faculty of Health and Medicine
Lancaster Medical School

Doctor of Philosophy

by Michael Give Chipeta

Geostatistical design and analysis methods are increasingly used in disease mapping, particularly in resource-limited settings where uniformly precise mapping may be unrealistically costly and the priority is often to identify critical areas where interventions can have the most health impact. In this thesis, which is based on four papers, we address the problem of geostatistical *sampling design*. In the first paper, we consider the problem of sampling design for efficient spatial prediction taking account of uncertain covariance structure, in the context of non-adaptive designs. We propose two classes of designs, namely: simple inhibitory and inhibitory plus close pairs. We evaluate the performance of these designs using an average prediction variance criterion and show how the findings are applied to the design of a rolling Malaria Indicator Survey (rMIS) in an ongoing large-scale, five-year malaria transmission reduction project in Malawi. In the second paper, we address the problem of efficient spatial prediction in the context of adaptive geostatistical designs (AGD). We propose two classes of designs based on singleton and batch sampling. We show how our findings inform an AGD of rMIS, in the perimeter of Majete Wildlife Reserve (MWR) in Chikwawa, southern Malawi. The third paper is a commentary on a paper by Ferreira and Gamerman (2015), which addressed the effect of preferential sampling of the locations at which to measure a spatial process. In the fourth paper, we present the first epidemiological field application of AGD sampling in a malaria prevalence survey. We give an in-depth description of the project, the study area and practical implementation of our adaptive sampling strategy. We present prevalence maps for children 6–59 months in MWR perimeter, showing high malaria transmission areas, often called "*hotspots*", that could be targeted with interventions.

Acknowledgements

My deepest and sincere gratitude is hereby expressed to the Lord, whom I believe gave me strength, courage and ability to complete this task. I thank God for the grace and love he abundantly and continuously bestows on me. I am sincerely grateful to my supervisors, Prof. Peter J. Diggle, Dr. Dianne J. Terlouw and Prof. Kamija S. Phiri. I am certain that without your day to day support and encouragement I would not have succeeded in this endeavour. Your supervision, support, mentorship and friendship is greatly appreciated. It has been a great honour to have been your student. I would also like to thank Prof. L. N. Kazembe for making statistics interesting and accessible to a floundering undergraduate student. To Dr. E. Giorgi, many thanks for being there for me every time I called upon you.

Thanks are owed to the Economic and Social Research Council for your generosity, through the award of a studentship (grant number [ES/J500094/1]) enabling me to embark upon this research. I would also like to thank the Majete malaria project (funded by Dioraphte Foundation, The Netherlands) for allowing me to be part of the team and providing the data that served as the motivating example for this research. My sincere thanks also go to all the members of CHICAS and MALEPI groups for your professional support and day to day friendship which made me feel at home throughout my Ph.D. studies.

My deepest gratitude goes to my parents (Japhet and Vita), I couldn't have done this without you, thanks for your never-ending support, love and belief. You are the reason that I have been able to accomplish so much. I am grateful, in a very special way, to Susan P. McDowall for being there for me since my undergraduate studies up to now, providing support in every way possible. Thanks must also go to Andrew and Joe for making my stay in the U.K. so much enjoyable by introducing me to so many people, places and involving me in so many social and cultural activities.

Last but not least, I would like to thank all my friends and colleagues, too numerous to mention, for answering all my questions whenever I called upon you. All the discussions, chats, distractions and the fun we had, have kept me sane through the years.

Contents

List of Figures

List of Tables

List of Papers

Paper 1. *Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure.*
Chipeta M. G., Terlouw D. J., Phiri K. S. and Diggle P. J.
Published in *Environmetrics Journal.*, doi: 10.1002/env.2425.
Contribution: lead author, design and implementation of the simulation studies, statistical analysis and writing of the paper.

Paper 2. *Adaptive geostatistical design and analysis for prevalence surveys.*
Chipeta M. G., Terlouw D. J., Phiri K. S. and Diggle P. J.
Published in *Journal of Spatial Statistics, Vol. 15, (2016), 70–84.*
Contribution: lead author, design and implementation of the simulation studies, statistical analysis and writing of the paper.

Paper 3. *Invited Discussion of "Optimal Design in Geostatistics under Preferential Sampling", by Ferreira and Gamerman.*
Chipeta M. G. and Diggle P. J.
Published in *Journal of Bayesian Analysis, Vol. 10, No. 3, (2015), 737–739.*
Contribution: lead author and writing of the paper.

Paper 4. *Adaptive geostatistical sampling enables efficient identification of malaria hotspots in rural Chikwawa, Malawi.*
Kabaghe A. N., **Chipeta M. G.**, McCann R. S., Phiri K. S., van Vugt M., Takken W., Diggle P. J. and Terlouw D. J.
Under review in *PLoS ONE* Journal.
Contribution: development of efficient sampling design algorithms for parameter estimation and spatial prediction, statistical analysis and writing of the paper.

Abbreviations

ACTs	Artemisinin-based Combination Therapies.
AGD	Adaptive Geostatistical Design.
APV	Average Prediction Variance.
ASTER	Advanced Space-borne Thermal Emission and Reflection Radiometer.
BAS	Balanced Acceptance Sampling.
CBO	Community-Based Organisation.
COMREC	College of Medicine Research Ethics Committee.
CRAN	Comprehensive R Archive Network.
CRS	Controlled Random Search.
DEM	Digital Elevation Model.
DHO	District Health Office.
DHS	Demographic Health Survey.
EA	Estimation-Adjusted.
EK	Empirical Kriging.
ESRC	Economic and Social Research Council.
GPS	Global Positioning System.
GRTS	Generalised Random Tessellation Stratified.

HPD	Highest Posterior Density.
HR	Hierarchical Randomisation.
ICP	Inhibitory *plus* Close Pairs.
IMSE	Integrated Mean Square Error.
IPTp	Intermittent Preventive Therapy in pregnancy.
IRS	Indoor Residual Spray.
ITN	Insecticide-Treated bed Nets.
LSTM-REC	Liverpool School of Tropical Medicine Research Ethics Committee.
MBG	Model-Based Geostatistics.
MDA	Mass Drug Administration.
MERIT	Meningitis Environmental Risk Information Technologies.
MICS	Multiple Indicator Cluster Survey.
MIS	Malaria Indicator Surveys.
MLE	Maximum Likelihood Estimation.
MMP	Majete Malaria Project.
MMSE	Minimum Mean Square Error.
MoH	Ministry of Health.
MPV	Maximum Prediction Variance.
MSE	Mean Square Errors.
MSPE	Mean Squared Prediction Error.
MUAC	Mid Upper Arm Circumference.
MWR	Majete Wildlife Reserve.
NAGD	Non-Adaptive Geostatistical Design.

NDVI	Normalised Difference Vegetation Index.
NMCP	National Malaria Control Programme.
NTD	Neglected Tropical Diseases.
NWDTC	North West Doctoral Training Centre.
ODK	Open Data Kit.
PV	Prediction Variance.
RDT	Rapid Diagnostics Test.
REML	Restricted Maximum Likelihood.
rMIS	rolling Malaria Indicator Survey.
SES	Socio-Economic Status.
SI	Simple Inhibitory.
SSA	sub-Saharan Africa.
SSA	Spatial Simulated Annealing.
STH	Soil-Transmitted Helminths.
THP	The Hunger Project.
USGS	United States Geological Survey.
WHO	World Health Organisation.

Symbols

Corr	Correlation.
Cov	Covariance.
$\boldsymbol{\mathcal{D}}$	Spatial region of interest.
$\boldsymbol{\partial}$	Partial derivative.
$\boldsymbol{E}$	Expectation of.
$\mathbb{R}^d$	d-Dimensional space.
$\boldsymbol{\mathcal{S}}$	Stochastic process.
$\boldsymbol{S}$	Realisation of $\mathcal{S}$.
$\boldsymbol{\mathcal{T}}$	Function of.
tr	Trace of.
Var	Variance.
$\boldsymbol{\mathcal{X}}$	Sampling design.

$\Vert \cdot \Vert$	Euclidean distance.
$\lvert\cdot\rvert$	Absolute value.
$\mid$	Conditional on.
$\sum$	Summation of.

$\sim$	Distributed as.
$[\cdot]$	Distribution of.
$\int_{\mathcal{D}}$	Integration over $\mathcal{D}$.
$\approx$	Approximately.
∞	Infinity.
$\propto$	Proportional to.
$\in$	Element of.
$\subset$	Subset of.
$\cup$	Set union.
$\emptyset$	Empty set.
$\setminus$	Set minus.
$\geq$	Greater than or equal to.
$\leq$	Less than or equal to.
$\boldsymbol{\beta}$	Regression coefficients (*Beta*).
$\boldsymbol{\Gamma}$	Capital letter *Gamma*.
$\boldsymbol{\gamma}$	Small letter *gamma*.
$\boldsymbol{\Delta}$	Distance between any two locations (capital letter *Delta*).
$\boldsymbol{\delta}$	Distance between any two locations (small letter *delta*).
$\boldsymbol{\epsilon}$	Some small number (*epsilon*).
$\boldsymbol{\zeta}$	Radius within which to place a close pair point of a primary location (*zeta*).
$\boldsymbol{\theta}$	Parameters (*theta*).
$\hat{\boldsymbol{\theta}}$	Parameter estimates (*theta hat*).

$\boldsymbol{\kappa}$	Smoothing parameter (*kappa*).
$\boldsymbol{\mu}$	Mean (*mu*).
$\boldsymbol{\nu}^2$	Re-parameterisation for τ^2/σ^2 (*nu*).
$\boldsymbol{\pi}$	Small letter *pi*.
$\boldsymbol{\rho}$	Correlation (*rho*).
$\boldsymbol{\sigma}^2$	Signal variance (*sigma squared*).
$\boldsymbol{\tau}^2$	Nugget variance (*tau squared*).
$\boldsymbol{\Phi}$	Cumulative distribution function (capital letter *Phi*).
$\boldsymbol{\phi}$	Scale parameter (small letter *phi*).
$\boldsymbol{\Omega}$	Regression parameter space (*Omega*).

To Japhet, Vita and Susan

Chapter 1

Introduction.

1.1 Motivation.

The study of geostatistical designs is an important topic in spatial statistics. In this thesis, we address the problem of geostatistical *sampling design* $\mathcal{X} = \{x_i,\ i = 1, \ldots, n\}$, i.e., a set of locations x_i from which data are collected to allow prediction of the unobserved spatial phenomenon of interest $\mathcal{S}$. Data $\{(x_i, y_i)\}$ are realised values of random variables Y_i associated with locations $x_i \in \mathcal{D} \subset \mathbb{R}^2$, where $\mathcal{D}$ is a geographical region of interest. Typically, each Y_i can be regarded as a noisy version of $S(x_i)$.

To motivate the sampling design problem in practice, we give two examples. The first is a malaria prevalence mapping from a cross-sectional household survey in Majete Wildlife Reserve (MWR) perimeter, in Chikwawa district, southern Malawi. The on-going Majete project is taking place in three administrative

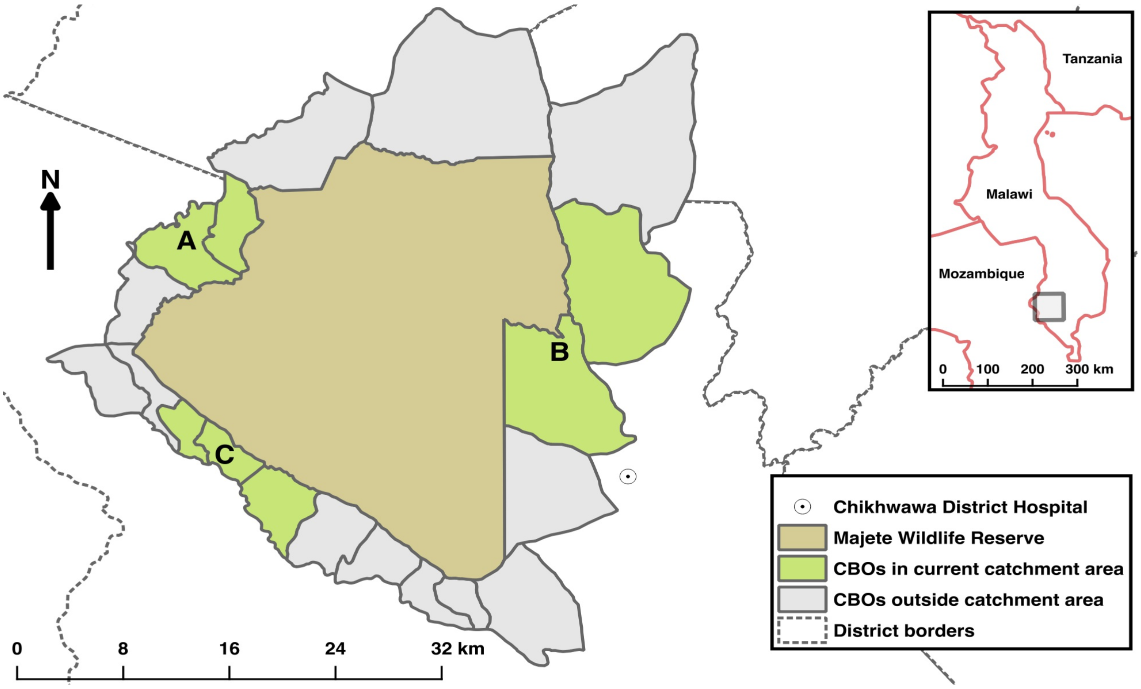


Figure 1.1: Map showing Majete Wildlife Reserve (brown) and borders of the 19 community-based organisations (CBOs) comprising the Majete perimeter. Three focal areas (green), labelled as A, B, and C, mark the communities selected for entomology/malaria indicator surveys and the trial. The rest of the CBOs (grey) are outside the project's catchment area.

units, referred to as *focal areas*, namely A, B and C, see Figure 1.1. This is a resource-limited setting where there are limited registries for disease data. Figure 1.2 shows a zoomed in map with locations for all enumerated households in focal area A. The geostatistical design problem here is to choose a finite number, n, of households to sample in an affordable and efficient manner so as to give the best possible prevalence predictions at unobserved households. The chosen *design* should enable accurate area-wide prevalence mapping so that programme implementers can identify sub-areas where targeted health interventions would have the most impact. See Section 1.8 for further details on the Majete malaria project.

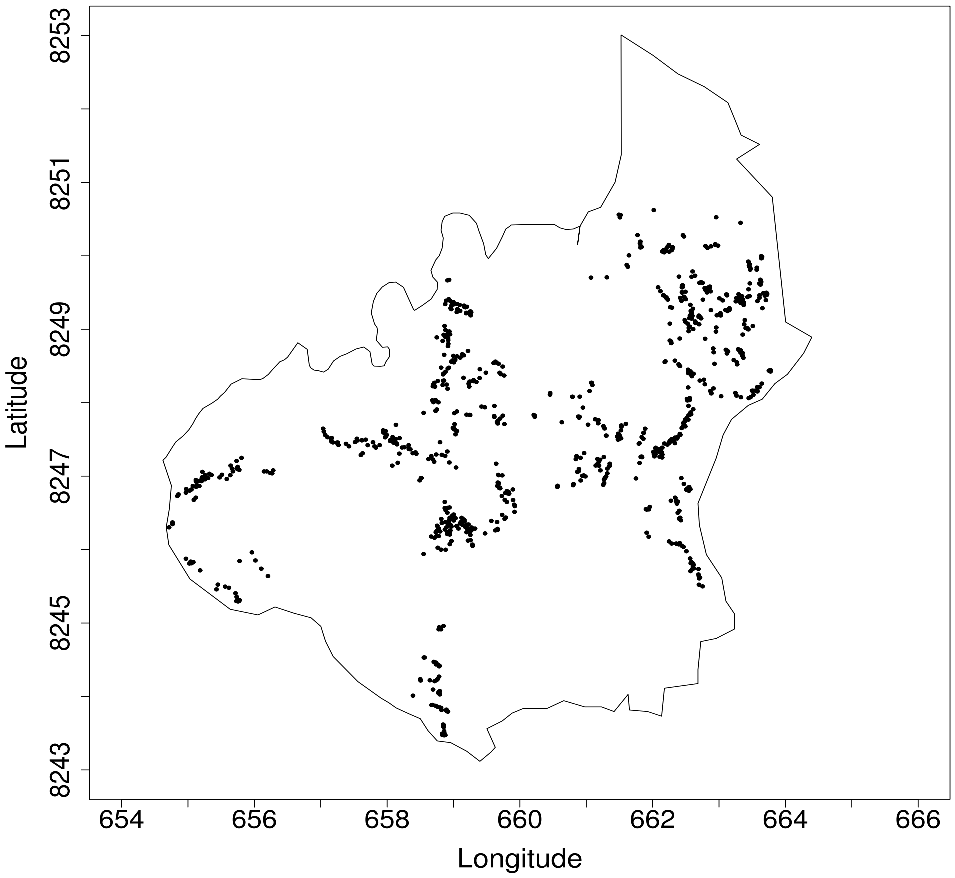


Figure 1.2: The black dots are household locations within Majete Wildlife Reserve perimeter - focal area A.

Our second example is a data-set containing lead pollution measurements taken from samples of two moss species (*Hypnum cupressiforme* and *Scleropodium purum*) collected in 2000 in Galicia, north-western Spain. These data and methods of collection have been reported elsewhere, see, for example, Fernández, Rey and Carballeira (2000), Fernández, Real et al. (2005) and Aboal et al. (2006). Briefly, samples were collected from two species of moss so as to map different metal concentrations in the whole of Galicia. Lead concentrations were measured in $\mu g/g$ dry weight. Samples were taken on an almost regular lattice, as shown in Figure 1.3, with measurement locations recorded using a global positioning system

(GPS). The data (x_i, y_i, z_i), $i = 1, \ldots, n$ are represented by location x, y and a corresponding measured value z_i. Some of the locations on the map appear to lie outside Galicia and others in the sea to the north. However, the shown boundary is both approximate and imperfectly registered, and is included only to add context to the map. In this example, policy makers would be interested in how precisely they could estimate the highest levels of pollution in the whole of the study region. Additionally, they would be interested in establishing patterns of distribution of the elements and identifying contaminated areas, including sources of contamination. The design problem here, unlike in the first example, would be to choose sample locations anywhere in the study region, not just at a pre-specified finite set of locations.

Suppose further, in both examples, that we have previously collected and analysed some data for exploratory or other purposes. How do we use this information to identify and collect additional data towards the analysis objective(s) over time? Several questions of scientific interest arise from the above scenarios. One could be interested in determining where to place the $x \in \mathcal{D} \subset \mathbb{R}^2$. One could also be interested in knowing how many design points need to be sampled to understand the heterogeneity of phenomenon of interest in the entire study region. The methods we have developed are generic in nature and widely applicable as demonstrated in the above examples. In this thesis, we focused on geostatistical sampling designs applied to the epidemiology of malaria transmission control and monitoring in a given spatial area of interest.

In what follows, we give a review of each of the following themes in relation

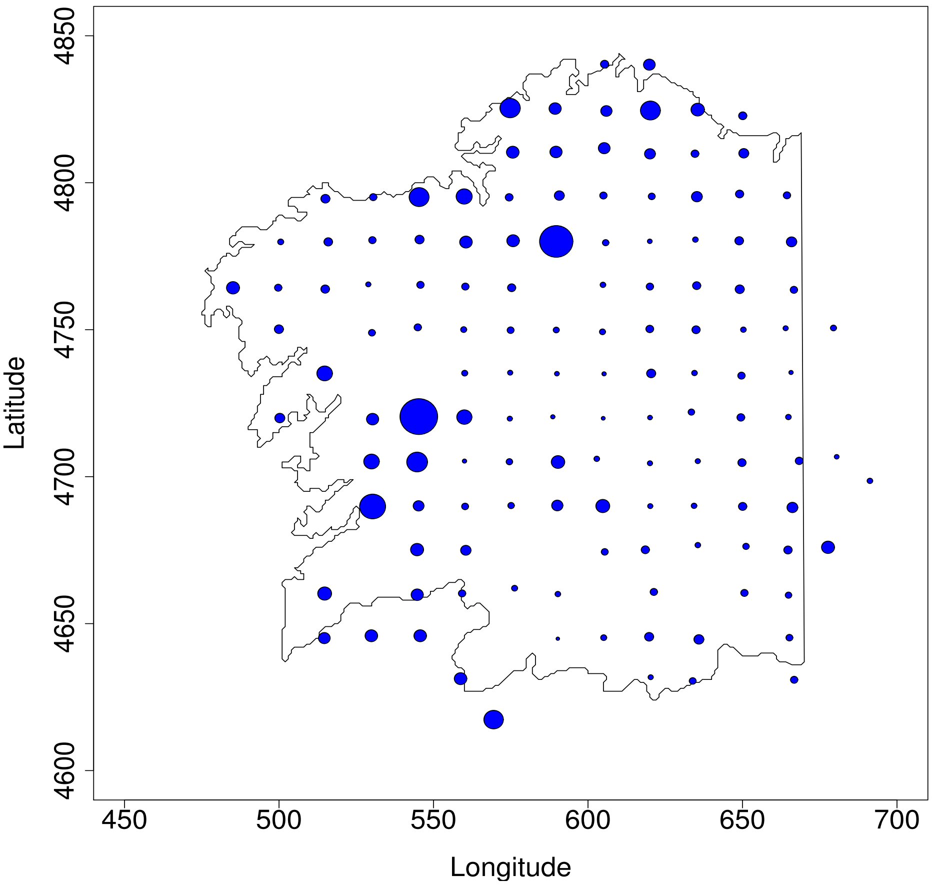


Figure 1.3: Map of lead concentrations in Galicia. Each circle centred at x_i, y_i has a radius proportional to pollution measurement z_i.

to the above-raised questions: the spatial model assumed for geostatistical data, maximum likelihood (ML) parameter estimation, spatial prediction methods, preferential sampling and the standard geostatistical model for prevalence data. We also give a brief introduction to malaria epidemiology and mapping, followed by a description of the Majete malaria project (MMP).

1.2 Geostatistical model.

Geostatistical models provide quantitative descriptions of phenomena distributed in space (Isaaks and Srivastava, 1989; Chilès and Delfiner, 2012). In geostatistics, spatial data are considered to contain deterministic and stochastic components (Müller, 2007). Data are measurements of the form $(x_i,\, y_i)$ where $i = 1, \ldots, n$, x_i is the spatial location and y_i is a response associated with $x_i \in \mathbb{R}^2$. Each y_i is a realisation of a random variable Y_i whose distribution is dependent on the value at the location x_i of an underlying spatially continuous stochastic process $S(x)$ which is not directly observable. The simplest example of a geostatistical model is the linear Gaussian model. In its most basic form, the model can be written as:

$$Y_i = S(x_i) + Z_i, \quad i = 1, \ldots, n \tag{1.1}$$

where the Z_i are mutually independent $\mathrm{N}(0, \tau^2)$ random variables and $S(x)$ is a stationary Gaussian process, with mean μ, variance $\sigma^2 = \mathrm{Var}\{S(x)\}$ and correlation function $\rho(u) = \mathrm{Corr}\{S(x), S(x')\}$, where $u = \|x - x'\|$ and $\| \cdot \|$ denotes Euclidean distance. The model is easily extended to include spatially referenced covariates, $d(x)$ say, in which case

$$Y_i = d(x_i)'\beta + S(x_i) + Z_i, \quad i = 1, ..., n. \tag{1.2}$$

where β's come from a finite dimensional parameter space $\Omega \subset \mathbb{R}^p$. This allows for the inclusion of a polynomial trend surface or, more generally, spatially referenced covariates. Writing $\mu(x) = X\beta$, a Gaussian model with a linear specification for the trend $\mu(x)$ can be expressed as a multivariate Gaussian:

$$Y \sim N(X\beta, \sigma^2 R(\phi) + \tau^2 I) \tag{1.3}$$

where X is an $n \times p$ matrix of covariates, β is the corresponding vector of regression parameters, and R depends on a scalar or vector-valued parameter ϕ, an $n \times n$ matrix with entries $\rho(u_{ij})$ where $u_{ij} = ||x_i - x_j||$. An equivalent specification to Equation (1.1) is that the Y_i are mutually independent conditional on $\{S(x) : x \in \mathbb{R}^2\}$, with

$$Y_i | \{S(x) : x \in \mathcal{D}\} \sim \mathrm{N}(S(x), \tau^2), \quad i = 1, \ldots, n. \tag{1.4}$$

This form extends more naturally to non-Gaussian models.

Other alternatives exist for non-Gaussian processes including process convolution models for a moving average with a non-normal latent process $S(x)$ (Higdon, 1998; Higdon, 2002) or allowing $S(x)$ to have non-stationary covariance structure (Plagemann, Kersting and Burgard, 2008). An example of a non-stationary Gaussian process model is:

$$S(x) = S(x-u) + Z(x) : x = 0, 1, \ldots \tag{1.5}$$

for which

$$\gamma(x, u) = \mathrm{Cov}\{S(x), S(x-u)\} = \sigma^2 ||x-u|| \tag{1.6}$$

A process of this nature is called an *intrinsic random function* (Diggle and Ribeiro, 2007; Paciorek and Schervish, 2006). Paciorek (2003) defined a class of closed-form non-stationary correlation functions, of which a special case is a non-stationary form of the Matérn correlation, as follows.

$$\rho(u,\phi,\kappa) = \frac{\sigma^2 \mid \sum_x \mid^{\frac{1}{4}} \mid \sum_{x'} \mid^{\frac{1}{4}}}{2^{\kappa-1}\Gamma(\kappa)} \left| \frac{\sum_x + \sum_{x'}}{2} \right|^{-\frac{1}{2}} (u/\phi)^{\kappa} \mathcal{K}_{\kappa}(u/\phi) \tag{1.7}$$

where $\sum_x$ and $\sum_{x'}$ are the covariance matrices of the Gaussian kernel at locations x and x'. In Equation (1.7), ϕ is the range parameter, with dimensions of distance, that determines the rate at which the correlation decays to 0, κ is the shape parameter unique to this family of correlation, known as the *order* of the Matérn model that determines the differentiability of the process $S(x)$, i.e. κ controls the smoothness of the spatial process. Larger values of κ correspond to smoother processes. The process is m times mean square differentiable if and only if $\kappa > m$ (Stein, 1999). The function $\mathcal{K}_{\kappa}(\cdot)$ is the modified Bessel function of second order κ. Both ϕ and κ must be greater than zero.

For all the simulations and computations in this thesis, we assume that the process $S(x)$ is a zero-mean, stationary and isotropic Gaussian process, i.e. with invariant distribution under translation and rotation. We work with the Matérn parametric family of correlation functions (Matérn, 1986) that is flexible yet simple. The Matérn correlation function for a stationary Gaussian process is given by the following expression:

$$\rho(u, \phi, \kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}\mathcal{K}_{\kappa}(u/\phi) \tag{1.8}$$

In Equation (1.8), the parameters ϕ and κ are as defined in Equation (1.7). The Matérn correlation function equates to $\rho(u) = \exp(-u/\phi)$ for $\kappa = 1/2$, the exponential family of correlation, and to $\rho(u) = \exp\{-(u/\phi)^2\}$ as $\kappa \to \infty$, the limiting case referred to as the Gaussian correlation function (Handcock and Stein, 1993; Diggle and Ribeiro, 2007).

1.2.1 Parameter estimation.

For most geostatistical applications, the ultimate goal is that of prediction rather than estimation of model parameters. However, reliable parameter estimates in correlated random fields are very important for prediction of the underlying signal process. Thus, in the initial phase of a geostatistical analysis, it is common to investigate the structure of the model, for a number of important key features, including the need for data transformation, the presence of anisotropy, the type of correlation function to use, and so on (Christensen, 2004).

Sometimes the estimation of model parameters themselves becomes the primary interest of analysis (Mardia and Marshall, 1984); in addition to estimation of regression coefficients. A linear Gaussian model Equation (1.1) typically has three covariance parameters that can be estimated, namely: *nugget variance*, τ^2; *scale* sometimes referred to as the *range*, ϕ; the *total sill*, $\tau^2 + \sigma^2 = \text{Var}\{Y(x)\}$. In case of a Matérn correlation function, the smoothness parameter, κ is a fourth parameter.

However, in practice, estimating κ is generally difficult, and the approach often taken is to choose from a finite set of values, for example, $\kappa = 0.5$, 1.5 or 2.5.

Maximum likelihood estimation (MLE) is asymptotically efficient for parameter estimation; it involves inference that is based on an explicit stochastic error model for the data. Mardia and Marshall (1984) were the first to use MLE in classical geostatistics, whose applications have increased tremendously over the years, mainly with the assistance of more powerful computers.

There are several advantages in using likelihood-based estimation of both covariance and regression parameters. The covariance parameters are directly estimated without having to calculate an experimental semi-variogram and then fit a model to it. The method provides uncertainty measures of the semi-variogram parameters and thus, in addition to obtaining measures of the reliability of estimates, it provides interval estimates and the ability to conduct statistical tests. Christensen (2004) showed that maximum likelihood is a feasible and powerful tool for model selection as well. In the current thesis, we restrict our attention to MLE methods.

MLE provides a general approach to simultaneously estimate $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ in a model of the form $Y \sim f(y_i, \boldsymbol{\theta}, \boldsymbol{\beta})$, where $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ represent covariance and regression parameters respectively (Kitanidis, 1987). For the multivariate Gaussian model Equation (1.3), the maximum likelihood estimates of the complete set of parameters $\boldsymbol{\theta} = (\beta, \tau^2, \phi, \sigma^2, \kappa)$ are the values that maximise the log-likelihood function:

$$\begin{aligned} L(\beta,\tau^2,\phi,\sigma^2,\kappa) = -0.5\{n\ \log(2\pi) + \log\{|(\sigma^2 R(\phi)+\tau^2 I)|\} \\ + (y - X\beta)^T(\sigma^2 R(\phi)+\tau^2 I)^{-1}(y - X\beta)\} \end{aligned} \tag{1.9}$$

Parameterising $\nu^2 = \tau^2/\sigma^2$ and $V = R(\phi) + \nu^2 I$, the log-likelihood is maximised at the following:

$$\hat{\beta}(V) = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{1.10}$$

and

$$\hat{\sigma^2}(V) = \frac{1}{n}\{y - X\hat{\beta}(V)\}^T V^{-1}\{y - X\hat{\beta}(V)\} \tag{1.11}$$

By substituting these into Equation (1.9), the log-likelihood can be written as a function of parameters ν^2 and ϕ only,

$$L(\nu^2,\phi) = -0.5\{n \log(2\pi) + n \log\ \hat{\sigma^2}(V) + \log\ |V| + n\}, \tag{1.12}$$

where R is an $n \times n$ matrix with elements $r_{ij} = \rho(||x_i - x_j||)$ and I the identity matrix. It is this function (Equation (1.12)) that is then numerically maximised. For the Matérn model, we maximise $L(\nu^2,\phi)$ separately for each individual value of κ.

1.2.2 Spatial prediction.

In some applications, the covariance function is assumed to be completely known, in which case the ultimate emphasis of design and subsequent analysis becomes prediction only. In practice, $\boldsymbol{\theta}$ has to be estimated from observations and subsequently used for making predictions at unsampled locations. In either case, the "minimum mean square error" criterion that measures, in a probabilistic sense, the error between a quantity of interest and its predictor is a reasonable criterion for predictive performance. To see why, let X, Y be random variables, then we can express $y = E[y|x] + \epsilon$, where ϵ is a random variable satisfying the following conditions. First, $E[\epsilon|x] = 0$ and secondly, $E[h(x)\epsilon] = 0$, where $h(\cdot)$ is any function of x. If say, $m(x)$ is any function of x, then the conditional expectation is the best prediction, where "best" means minimum mean squared error (MMSE), which is given by

$$E[y|x] = \operatorname*{argmin}_{m(x)} E[(y - m(x))^2] \tag{1.13}$$

Minimisation of the MSE underlies numerous methods in the statistical sciences (Guo et al., 2011). Simple kriging, the construction of a surface $\hat{S}(x)$ from the observed data $\mathbf{S}(x) = (S(x_1), \ldots, S(x_n))'$, a core geostatistical method, is equivalent to MMSE under a linear Gaussian model (Equation (1.1)). In the kriging method, estimates of all model parameters are plugged into the prediction equation as if they were the true parameter values, in a process referred to as "*plug-in prediction*" (Diggle and Ribeiro, 2007). Inferences can be made, depending on the context, for a single point, say x_0; prediction about the value of $S(\cdot)$ over an area of interest or subsets thereof; the maximum or minimum value of $S(x)$ or prediction of the probability that $S(x)$ is above or below a particular threshold $\boldsymbol{c}$. The MMSE prediction of $S(x)$ at a location x_0 as a function of data $y = (y_1, \dots, y_n)$ which minimises the quantity $E[\{\hat{S}(x) - S(x)\}^2]$ is:

$$\hat{S}(x) = \mu + \sum_{i=1}^{n} w_i(x)(y_i - \mu) \tag{1.14}$$

where $w_i(x)$ are functions of the covariance parameters σ^2, ϕ and τ^2. The μ in Equation (1.14) is referred to as the constant stationary function or the global mean and $\sum_{i=1}^{n} w_i(x)(y_i - \mu)$ is the spatially correlated stochastic part of variation. The aim is to find the predictor $\hat{T} = \hat{S}(x)$ that minimises the MSE of prediction (Schabenberger and Gotaway, 2005):

$$MSE(\hat{T}) = E[(T - \hat{T})^2] \tag{1.15}$$

The MMSE predictor of any random variable T from data Y is $\hat{T} = E[T|Y]$. We prove this in the following equations.

Write

$$E[(T - \hat{T})^2] = E_Y[E_T[(T - \hat{T})^2|Y]], \tag{1.16}$$

where the subscripts on the two expectation operators indicate that the expectations are with respect to Y and T, respectively. Write the inner expectation in Equation (1.16) as

$$E_T[(T-\hat{T})^2|Y] = \mathrm{Var}_T\{(T-\hat{T})|Y\} + \{E_T[(T-\hat{T})|Y]\}^2$$

Conditional on Y, any function of Y is a constant, so $\mathrm{Var}_T\{(T-\hat{T})|Y\} = \mathrm{Var}_T(T|Y)$ and $E_T[T-\hat{T}|Y] = E_T[T|Y] - \hat{T}$. Hence,

$$E_T[(T-\hat{T})^2|Y] = \mathrm{Var}_T(T|Y) + \{E_T(T|Y) - \hat{T}\}^2. \tag{1.17}$$

Taking the expectation of the expression on the right-hand side of Equation (1.17) with respect to Y gives

$$E[(T-\hat{T})^2] = E_Y[\mathrm{Var}_T(T|Y)] + E_Y\{[E_T(T|Y) - \hat{T}]^2\} \tag{1.18}$$

The first term on the right-hand side of Equation (1.18) does not depend on the choice of $\hat{T}$, whilst the second is non-negative, and equal to zero if and only if $\hat{T} = E[T|Y]$. This completes the proof.

The minimum mean square error predictor for $T = S(x)$ is

$$\hat{T} = \mu + r'V^{-1}(Y - \mu\mathbf{1}) \tag{1.19}$$

whose prediction variance is

$$\mathrm{Var}(T|Y) = \sigma^2(1 - r'V^{-1}r) \tag{1.20}$$

where r is an $n \times 1$ vector, $V = R + \nu^2 I$ with $\nu^2 = \tau^2/\sigma^2$, I the identity matrix and R is the $n \times n$ matrix with elements $r_{ij} = \rho(||x_i - x_j||)$. The value of the prediction variance (Equation (1.20)) at the observed value of Y estimates the achieved MSE of $\hat{T}$ (Equation (1.15)) and when the conditional variance $\mathrm{Var}(T|Y)$ does not depend on Y, MSE is equal to prediction variance. This important characteristic makes it attractive to use as a design criterion.

1.3 Preferential sampling.

Given the stochastic process $\mathcal{S}$, the design $\mathcal{X}$ and the measurements Y, a standard geostatistical analysis assumes that sampling is *non-preferential* if the joint distribution $[\mathcal{S}, \mathcal{X}, Y]$ factorises as $[S, X, Y] = [\mathcal{S}][\mathcal{X}][Y|\mathcal{S}(\mathcal{X})]$, where $[\cdot]$ refers to "distribution of". This is in line with existing knowledge in geostatistics, where models for the data treat the sampling locations x_i either as fixed by design or otherwise stochastically independent of the process $S(x)$ (Diggle, Menezes and Su, 2010). The measurements are analysed and inferences made conditional on the design $\mathcal{X}$, where $\mathcal{X}$ is stochastically independent of $\mathcal{S}$. On the other hand, *preferential* sampling allows stochastic dependence between the measurements Y and locations x_i where measurements are made i.e. the design, which depends on the unobserved quantity $\mathcal{S}$ which we are trying to predict (Diggle, Menezes and Su, 2010). The preferential sampling distribution is given as $[\mathcal{S}, \mathcal{X}, Y] = [\mathcal{S}][\mathcal{X}|\mathcal{S}][Y|\mathcal{S}, \mathcal{X}]$.

The inferences we make about a response surface are affected by the choice of sampling sites (Müller, 2007; Gelfand, Banerjee and Finley, 2012; Shaddick and Zidek, 2014); therefore a topical and important question in geostatistical designs is whether sampling is done preferentially or not. If sampling sites are preferentially chosen to capture larger (or smaller) than average values of a response, e.g., air pollution in a city or biomass in large tract of a forest, then subsequent estimation and prediction of the exposure surface using standard geostatistical methods may be misleading due to the selective sampling (Diggle and Ribeiro, 2007; Diggle, Menezes and Su, 2010; Gelfand, Banerjee and Finley, 2012). Practical needs or deliberate actions often lead to preferential sampling. For example, in an air quality monitoring network, Guttorp and Sampson (2010) state that air pollution monitoring sites may be intentionally located for a number of reasons, including to measure pollution levels: (i) outside of urban areas; (ii) in residential areas; and (iii) near pollution sources.

An important issue, therefore, is the knowledge of any preferential sampling process in order to avoid misleading inferences. Given this knowledge, then effects of preferential sampling on parameter estimation and spatial prediction can be assessed (Shaddick and Zidek, 2014; Zidek, Shaddick and Taylor, 2014). In a recent paper, Ferreira and Gamerman (2015) address the effect of preferential sampling in geostatistics when the choice of new sampling locations is the main interest of the researcher. We come back to this topic in Chapter 4, where we give theoretical remarks on what we call *adaptive designs* (see Chapter 3), including an explanation of why this does not necessarily require consideration of preferential sampling.

1.4 Geostatistical model for prevalence surveys.

A disease prevalence survey involves visiting communities at locations x_i distributed over a region of interest where, in each community, field teams sample n_i individuals at risk and record whether each individual tests positive or negative for the disease in question, for $i = 1, \ldots, n$. Let Y_i be the number of positive outcomes out of n_i individuals tested at location x_i in a region of interest $\mathcal{D} \subset \mathbb{R}^2$, and $d(x_i) \in \mathbb{R}^p$ a vector of associated covariates. Then the standard model assumes that $Y_i \sim \text{Binomial}(n_i, p(x_i))$ where $p(x)$ is the prevalence of disease at location x. Linkage of the $p(x_i)$ at different locations is usually desirable and is essential if we wish to make inferences about $p(x)$ at unsampled locations x (Diggle and Giorgi, 2015). The model further assumes that

$$\log[p(x)/\{1 - p(x)\}] = d(x)'\beta + S(x) \tag{1.21}$$

where $S(x)$ is a stationary Gaussian process with zero mean, variance σ^2 and correlation function $\rho(u) = \text{Corr}\{(S(x), S(x')\}$, where u is the distance between x and x'. Stanton and Diggle (2013) showed that provided the binomial denominators n_i are large (i.e., $n_i \geq 100$) and the underlying prevalence is not too close to zero (i.e.,

$|p(x) - 0.5| \leq 0.4$), reliable predictions can be obtained using a computationally simpler non-hierarchical approximate trans-Gaussian model. Define the *empirical logit transform*,

$$Y_i^* = \log\{(Y_i + 0.5)/(n_i - Y_i + 0.5)\}$$

and assume that

$$Y_i^* = d(x_i)'\beta + S(x_i) + Z_i, \tag{1.22}$$

where the Z_i are mutually independent zero-mean Gaussian random variables with variance τ^2. In circumstances where n_i is less than 100, the approximate trans-Gaussian method continues to give reliable results provided the prevalence is correspondingly closer to 0.5, for example, if $n_i \approx 50$ and $|p(x) - 0.5| \leq 0.25$. Stanton and Diggle (2013) further state that for larger values of n_i, the approximate trans-Gaussian method can tolerate more extreme values of underlying prevalence. In both exact and approximate trans-Gaussian methods, predictive inferences need to be back-transformed from the logit to the prevalence scale.

1.5 Bayesian geostatistical analysis.

In the classical geostatistics inference framework in Sections 1.2.1 to 1.2.2, the covariance structure is estimated first, then the estimated covariance structure is used for prediction. Estimates of all model parameters are *plugged* into the prediction equation as if they were the true parameter values. It is common to ignore the effect of uncertainty in the covariance structure on subsequent predictions (Stein, 1999). Unlike this approach, a Bayesian inference approach to parameter estimation and prediction of spatial processes provides a general methodology for taking into account the uncertainty about parameters on subsequent predictions (Diggle, Tawn and Moyeed, 1998). The Bayesian inferential framework enables inference to use information from data via the likelihood function as well as from other sources such as previous studies, expert judgement and researchers' own subjective judgement, which is formalised by placing prior distributions on the model parameters.

Bayes' theorem, whose statement is: for any two events, say A and B, with $0 < Pr(A) < 1$ and $Pr(B) > 0$, defined as:

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}, \tag{1.23}$$

is fundamental to Bayesian inference. Given data, say $Y = y$, Bayes' theorem combines three key elements namely:

- the **prior distribution** of the parameters, $p(\theta)$, i.e. prior beliefs about the values of θ;
- the **likelihood** function of the data given the parameters, $f(y|\theta)$;
- the **posterior distribution** of the parameters given the data, $p(\theta|y)$, i.e. a combination of prior beliefs about the unknown parameter θ and information from data y.

The posterior distribution is then expressed as follows:

$$p(\theta|y) \propto f(y|\theta)p(\theta) \tag{1.24}$$

Given the geostatistical model of Equation (1.2), the residual has two components namely the correlated and uncorrelated terms. The correlated term, $S(x)$, introduces the partial sill, σ^2, and the range, ϕ, and the uncorrelated term, Z, introduces the nugget effect, τ^2. The nugget effect represents the measurement error and/or micro-scale variability.

In order to construct a Bayesian model formulation for Equation (1.2), specification of the prior distributions for β and θ is a necessary step. Where there is no prior knowledge about the β, the pragmatic solution is to adopt a non-informative, but improper prior distributions with bounds $-\infty$ and ∞ which reflects lack of prior knowledge other than that the regression coefficients can take any positive or negative value. For the spatial parameters σ^2, ϕ and τ^2, any distribution of θ can be used (Giorgi and Diggle, 2015); gamma, inverse gamma and normal are some of the distributions that have been used in literature.

In Chapter 5 we implement a geostatistical binary probit model within a Bayesian framework. The model has a hierarchical two-level structure so as to include individual- and household- level (or any other unit comprising a group of individuals, e.g. village or school) variables (Giorgi and Diggle, 2015).

1.6 Malaria epidemiology.

Globally, an estimated 3.2 billion people are at risk of malaria (World Health Organisation, 2015b). The World Health Organisation (2015b) indicates that 88 % of the cases of malaria occurred in sub-Saharan Africa (SSA) region. In 2015, it was estimated that there were 429,000 deaths from malaria globally, 92 % of which occurred in Africa and most of these were children under 5 years old (World Health Organisation, 2016). The most at-risk populations are those that live in stable transmission areas as shown in Figure 1.4. A smaller proportion of the at-risk population lives in areas where the risk of malaria is more seasonal and less predictable, because of either altitude or rainfall patterns.

In areas of stable malaria transmission, young children and pregnant women are the population groups at highest risk for malaria morbidity and mortality (World Health Organisation, 2012; Nansseu; et al., 2013). Most children experience their first malaria infections during the first two years of life (Ouédraogo et al., 2013) when they have not yet acquired adequate clinical immunity, which makes these early years particularly dangerous. Adult women in areas of stable transmission have a high level of immunity, but this is impaired especially in the first pregnancy, resulting in a higher risk of infection increases (Steketee et al., 2001). Malaria in pregnancy affects both mother and unborn child; it is associated with, among others, anaemia, pre-term delivery, high risk of maternal death and low birth weight (Huynh et al., 2011; De Beaudrap et al., 2013; Kalilani-Phiri et al., 2013). In malaria endemic regions, malaria control, prevention and treatment take up large

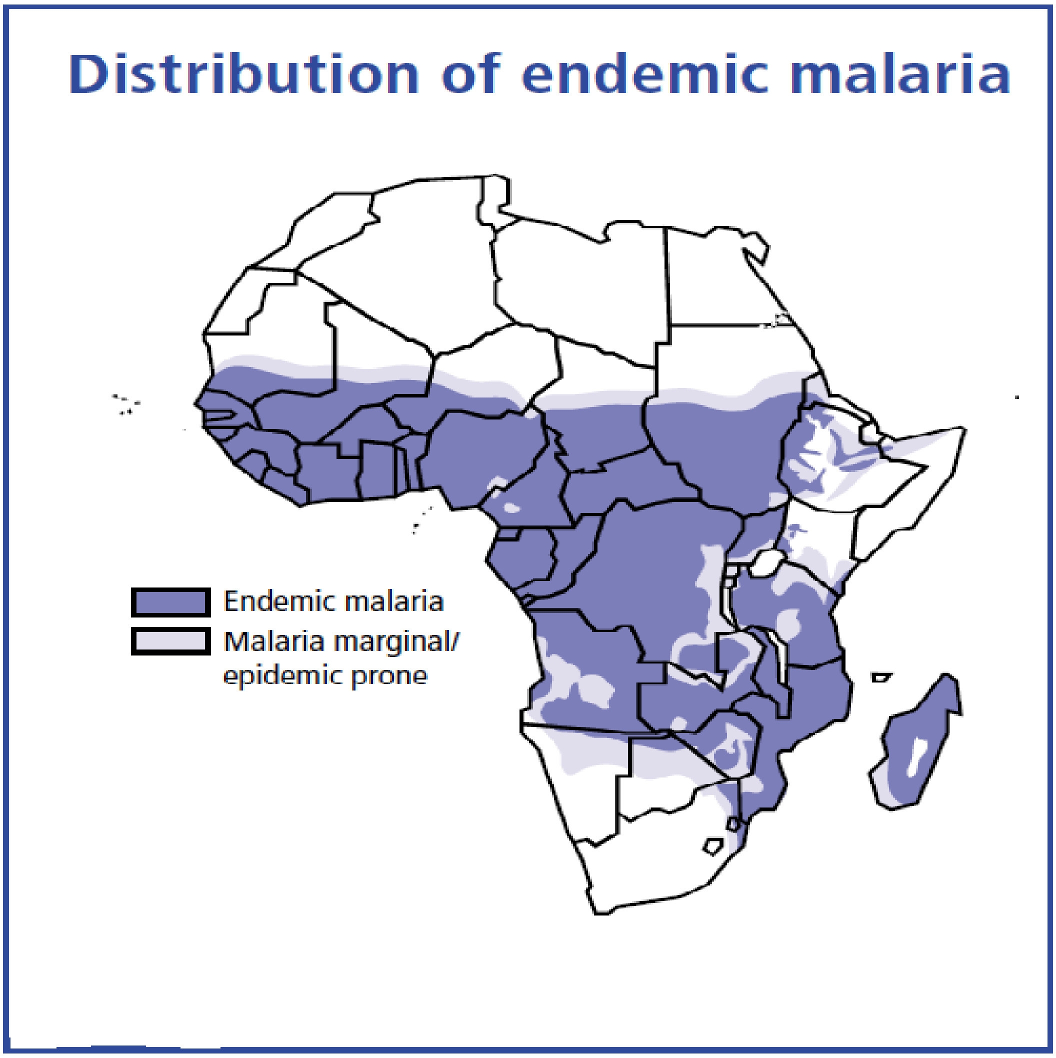

Distribution of endemic malaria
Endemic malaria
Malaria marginal/ epidemic prone

Figure 1.4: Distribution of malaria in sub-Saharan Africa. ***Source:*** World health malaria report (2013).

proportions of national health budgets at the expense of other equally important developmental and economic activities (Kleinschmidt, 2001). Therefore, in addition to the impact on health status, malaria also has economic consequences inhibiting economic development in SSA and other endemic regions.

Malaria is a preventable and treatable disease, provided the recommended interventions are properly implemented (World Health Organisation, 2013). These interventions include: (i) vector control through the use of insecticide-treated nets

(ITNs), indoor residual spraying (IRS) and, in some specific settings, larval control; (ii) chemo-prevention for the most vulnerable populations, particularly pregnant women and infants; (iii) timely confirmation of malaria diagnosis through microscopy or rapid diagnostic tests (RDTs) for every suspected case; and (iv) timely treatment with appropriate antimalarial medicines with artemisinin-based combination therapies (ACTs) (World Health Organisation, 2013; World Health Organisation, 2015b).

A better understanding of vector distribution and malaria risk through geostatistical mapping is an important tool in its control and eventual elimination. With more accurate local maps, it is possible to target interventions to areas and populations where they are needed and could have the most health impact.

1.7 Malaria disease mapping.

Over the past decade, great progress has been achieved in malaria control globally, thanks to unprecedented financial investments. As efficacious interventions such as ITNs, IRS, and effective artemisinin-based antimalarials were scaled up successfully, this triggered a renewed global commitment and push towards transmission reduction targets, and ultimately elimination (Bhatt et al., 2015; World Health Organisation, 2015a). The 2016 – 2030 WHO global technical strategy for malaria aims to reduce malaria case incidence to 10 % of 2015 levels by 2030 (World Health Organisation, 2015a).

To support control programmes and achieve such ambitious targets within resource-limited settings requires timely and frequent identification of sub-national variation and areas that lag behind in performance. At the same time, there is increasing evidence to suggest that malaria control in countries with substantial heterogeneity of malaria transmission may be more effective if additional control efforts are targeted towards so-called "*hotspots*" of transmission within districts (Woolhouse et al., 1997; Bousema, Griffin et al., 2012; Bousema, Stevenson et al., 2013). It is important that resources are targeted effectively, and this requires accurate and detailed information about which areas are worst affected, and where malaria might conceivably be eliminated altogether (van der Hoek et al., 2003). This is currently limited by the lack of user-friendly and affordable tools. Additionally, in the resource-limited settings where malaria is endemic, there are typically limited or no registries of disease burden.

Important aspects of malaria control include proper identification of vector distributions, vector survival conditions/environments, malaria distribution and risk through mapping. In particular, maps could enable targeting control measures and interventions at high-risk areas and greatly increase the cost efficiency of malaria control programmes. Additionally, maps can be used for policy decision making in development projects, especially settlement locations relative to vector habitats environment in general.

The methodology developed in the current thesis permits designing of disease prevalence studies, taking into account of spatial characteristics and statistics. It enables generation of accurate fine-scale spatial risk maps using model-based geostatistical methods, adjusted for environmental, public health and climatic covariates. These maps can subsequently inform posterior burden surveys with sampling frames that are specifically designed to monitor disease burden heterogeneity. The methods, although specifically motivated by the problem of malaria monitoring and evaluation in the Majete malaria project (MMP), are generic in nature and could, therefore, prove useful for applications to other diseases, not only in Malawi but also in other resource-limited settings in sub-Saharan Africa.

1.8 Majete malaria project.

The Majete Malaria Project (MMP) is an operational research collaboration comprising multidisciplinary researchers from College of Medicine and Malawi Liverpool Wellcome Trust (Malawi), Academic Medical Center – University of Amsterdam and Wageningen University (The Netherlands) and Lancaster University (United Kingdom), with operational contribution from Malawi Ministry of Health, Malawi National Malaria Control Programme (NMCP), African Parks-Majete and The Hunger Project (THP). The main aim of the project is to reduce the burden of malaria in the communities surrounding Majete Wildlife Reserve (MWR) in Chikwawa and Mwanza districts of southern Malawi. To achieve this outcome, the project is systematically implementing interventions in three focal areas in the "Majete perimeter" (see Figure 1.1). The interventions are as follows:

- Community awareness and health promotion campaign on malaria symptoms, treatment, community impact, complications, and prevention. These campaigns teach communities about the health system structure and function from the highest level i.e. District Health Office (DHO) to the community level and links between the community systems and the formal health system. The expected output is an increased knowledge of malaria and the health system, and an increased commitment to participate in malaria prevention and control.

- Scaling up of malaria control interventions based on the national malaria control policy. This involves:

  1. Universal coverage of insecticide-treated mosquito net ownership for pregnant women and children below 5 years of age;
  2. Household ownership of an insecticide-treated net (ITN) per 1.8 persons;
  3. Universal access to prompt diagnosis and appropriate treatment of malaria;
  4. Access to malaria prevention during pregnancy through intermittent preventive therapy in pregnancy (IPTp).

  The output will be an increased coverage and utilisation of these interventions in the focal areas.

- An assessment of the health system within the Majete perimeter to identify and address gaps in health service delivery.

- A cluster randomised trial comparing the effectiveness of the current national control policy with other combinations of malaria transmission reduction methods. These non-routine interventions involve household improvement to reduce the amount of mosquitoes entering houses and mosquito larviciding using a toxic microbe, *Bacillus thurengesis israelensis* (BTI).

To measure the impact of these interventions, specific malaria indicators are monitored in the community and health facilities through a rolling malaria indicator survey (rMIS) (Roca-Feltrer et al., 2012), a house improvement trial and an entomology survey, among other studies. These studies are being implemented using adaptive and/or inhibitory geostatistical design methodologies, developed and demonstrated in later chapters of this thesis.

1.9 The structure of the thesis.

In Chapter 2 (Paper 1), we develop non-adaptive inhibitory geostatistical designs in the context of malaria prevalence surveys. The methodology extends a simple inhibition point process (Diggle, 2013) to geostatistical designs for prevalence data to allow the inclusion of close pairs in an otherwise spatially-regular but randomised layout. We give an overview of non-adaptive geostatistical strategies, including classes of the designs. We define and develop our class of inhibitory geostatistical designs, assuming a stationary Matérn correlation structure. In our simulation studies, we consider two model classes, namely the linear Gaussian and Binomial geostatistical models. In both cases, the predictive target is $\mathcal{S}$. In our application, we use data from the Majete malaria project to demonstrate the implementation of the inhibitory geostatistical methodology to design a malaria prevalence study in focal area A.

In Chapter 3 (Paper 2), we develop adaptive geostatistical designs in the context of malaria prevalence surveys. We give an overview of geostatistical designs, then define a class of adaptive designs. We again make an assumption of a stationary Gaussian process with Matérn correlation structure. In our simulation studies, we compare the predictive efficiency of adaptive and non-adaptive geostatistical designs. We also analyse data from two sampling waves of rolling malaria indicator survey (rMIS) sampling in Majete. We then demonstrate the implementation of adaptive sampling in practice using the accumulating data to determine new sampling locations for each subsequent sampling wave.

Chapter 4 (Paper 3) is an invited discussion of an article entitled: "Optimal Design in Geostatistics under Preferential Sampling" by Gustavo da Silva Ferreira and Dani Gamerman (2015). The paper analyses the effect of preferential sampling in geostatistics when the choice of new sampling locations is the main interest of the researcher. In the commentary, we address two issues. The first is a set of theoretical remarks on adaptive design, including an explanation of why this does not necessarily require consideration of preferential sampling. The second issue is on practical constraints that may limit the scope for theoretically optimal designs to be used in practice, especially in low-resource settings.

In Chapter 5 (Paper 4) we describe the first field epidemiological application of adaptive geostatistical sampling design in continuous malaria prevalence surveys for a 12 month period (from April 2015 to April 2016). We conducted repeated cross-sectional surveys guided by an adaptive sampling design to monitor the prevalence of malaria parasitaemia in children aged 6–59 months and women aged 15–49 years within the Majete Malaria Project. More specifically, in this paper we analyse and present maps for malaria prevalence in children 6–59 months from Majete Wildlife Reserve perimeter in Chikwawa district, southern Malawi. We also show how the methodology can be used by programme managers and implementers to identify and map "*hotspots*" as well as intervention coverage in practice.

Chapter 6 is a concluding general discussion where we present a summary of the main contributions, the implications of our results in malaria transmission control and briefly explore possible future extensions of the developed methodologies in the previous chapters.

References

Aboal, J.R., Real, C., Fernández, J.A. and Carballeira, A. (2006) 'Mapping the results of extensive surveys: The case of atmospheric biomonitoring and terrestrial mosses', *Science of the Total Environment* 356 (1-3), pp. 256–274.

Bhatt, S., Weiss, D.J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K.E., Moyes, C.L., Henry, A., Eckhoff, P.A., Wenger, E.A., Briet, O., Penny, M.A., Smith, T.A., Bennett, A., Yukich, J., Eisele, T.P., Griffin, J.T., Fergus, C.A., Lynch, M., Lindgren, F., Cohen, J.M., Murray, C.L.J., Smith, D.L., Hay, S.I., Cibulskis, R.E. and Gething, P.W. (2015) 'The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015', *Nature* 526 (7572), pp. 207–211.

Bousema, T., Griffin, J.T., Sauerwein, R.W., Smith, D.L., Churcher, T.S., Takken, W., Ghani, A., Drakeley, C. and Gosling, R. (2012) 'Hitting Hotspots: Spatial Targeting of Malaria for Control and Elimination', *PLoS Medicine* 9 (1), e1001165.

Bousema, T., Stevenson, J., Baidjoe, A., Stresman, G., Griffin, J.T., Kleinschmidt, I., Remarque, E.J., Vulule, J., Bayoh, N., Laserson, K., Desai, M., Sauerwein, R., Drakeley, C. and Cox, J. (2013) 'The impact of hotspot-targeted interventions on malaria transmission: study protocol for a cluster-randomized controlled trial.', *Trials* 14 (1), p. 36.

Chilès, J.-P. and Delfiner, P. (2012) *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed., New Jersey: John Wiley & Sons, Inc.

Christensen, O.F. (2004) 'Monte Carlo Maximum Likelihood in Model-Based Geostatistics', *Journal of Computational and Graphical Statistics* 13 (3), pp. 702–718.

De Beaudrap, P., Turyakira, E., White, L.J., Nabasumba, C., Tumwebaze, B., Muehlenbachs, A., Guérin, P.J., Boum, Y., McGready, R. and Piola, P. (2013) 'Impact of malaria during pregnancy on pregnancy outcomes in a Ugandan prospective cohort with intensive malaria screening and prompt treatment.', *Malaria Journal* 12, p. 139.

Diggle, P.J. (2013) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns.* 3rd ed., Boca Raton: CRC Press.

Diggle, P.J. and Giorgi, E. (2015) 'Model-Based Geostatistics for Prevalence Mapping in Low-Resource Settings.', *Journal of the American Statistical Association (in press)*, pp. 1–42.

Diggle, P.J., Menezes, R. and Su, T.-L. (2010) 'Geostatistical inference under preferential sampling', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59 (2), pp. 191–232.

Diggle, P.J. and Ribeiro, J.P. (2007) *Model-based Geostatistics*, New York: Springer.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998) 'Model-based geostatistics (with discussion)', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47 (3), pp. 299–350.

Fernández, J.A., Real, C., Couto, J.A., Aboal, J.R. and Carballeira, A. (2005) 'The effect of sampling design on extensive bryomonitoring surveys of air pollution', *Science of the Total Environment* 337 (1-3), pp. 11–21.

Fernández, J.A., Rey, A. and Carballeira, A. (2000) 'An extended study of heavy metal disposition in Galicia (NW Spain) based on moss analysis', *The Science of the Total Environment* 254, pp. 31–44.

Ferreira, G.d.S. and Gamerman, D. (2015) 'Optimal Design in Geostatistics under Preferential Sampling', *Bayesian Analysis* 10 (3), pp. 711–735.

Gelfand, A.E., Banerjee, S. and Finley, A.O. (2012) 'Spatial Design for Knot Selection in Knot-Based Dimension Reduction Models', *Spatio-temporal design: Advances in efficient data acquisition*, ed. by J. Mateu and W.G. Müller, 1st ed., Chichester, UK: John Wiley & Sons, Ltd, chap. 7, pp. 142–169.

Giorgi, E. and Diggle, P.J. (2015) 'PrevMap : an R Package for Prevalence Mapping', *Journal of Statistical Software (to appear)*, pp. 1–27.

Guo, D., Wu, Y., Shamai, S. and Verdú, S. (2011) 'Estimation in Gaussian noise: Properties of the minimum mean-square error', *IEEE Transactions on Information Theory* 57 (4), pp. 2371–2385.

Guttorp, P. and Sampson, P.D. (2010) 'Discussion of Geostatistical inference under preferential sampling by Diggle, P.J., Menezes, R. and Su, T.', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59 (2), pp. 191–232.

Handcock, M.S. and Stein, M.L. (1993) 'A Bayesian Analysis of Kriging', *Technometrics* 35 (4), pp. 403–410.

Higdon, D. (1998) 'A process-convolution approach to modeling temperatures in the North Atlantic Ocean (with discussion)', *Environmental and Ecological Statistics* 5 (2), pp. 173–190.

Higdon, D. (2002) 'Space and Space-Time Modeling using Process Convolutions', *Quantitative Methods for Current Environmental Issues*, ed. by C.W. Anderson, V. Barnett, P.C. Chatwin and A.H. EI-Shaarawi, 1st ed., London: Springer, chap. 2, pp. 37–56.

Huynh, B.T., Fievet, N., Gbaguidi, G., Dechavanne, S., Borgella, S., Guézo-Mévo, B., Massougbodji, A., Tuikue Ndam, N., Deloron, P. and Cot, M. (2011) 'Influence of the timing of malaria infection during pregnancy on birth weight and on maternal anemia in Benin', *American Journal of Tropical Medicine and Hygiene* 85 (2), pp. 214–220.

Isaaks, E.H. and Srivastava, R.M. (1989) *An Introduction to applied geostatistics*, New York: Oxford University Press.

Kalilani-Phiri, L., Thesing, P.C., Nyirenda, O.M., Mawindo, P., Madanitsa, M., Membe, G., Wylie, B., Masonbrink, A., Makwakwa, K., Kamiza, S., Muehlenbachs, A., Taylor, T.E. and Laufer, M.K. (2013) 'Timing of Malaria Infection during Pregnancy Has Characteristic Maternal, Infant and Placental Outcomes', *PLoS ONE* 8 (9), e74643.

Kitanidis, P.K. (1987) 'Parametric estimation of covariances of regionalized variables', *Journal of the American Water Resources Association* 23 (4), pp. 557–567.

Kleinschmidt, I. (2001) 'Spatial statistical analysis, modelling and mapping of malaria in Africa', PhD thesis, University of Basel.

Mardia, K.V. and Marshall, R.J. (1984) 'Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression', *Biometrika* 71 (1), pp. 135–146.

Matérn, B. (1986) *Spatial Variation*, 2nd ed., Berlin: Springer.

Müller, W.G. (2007) *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*, 3rd ed., Berlin: Springer-Verlag.

Nansseu; J.R.N., Noubiap; J.J.N., Ndoula; S.T., Zeh; A.F.M. and Monamele, C.G. (2013) 'What Is the Best Strategy for the Prevention of Transfusion-Transmitted Malaria in Sub-Saharan African Countries Where Malaria Is Endemic?', *Malaria Journal* 12, p. 465.

Ouédraogo, A., Tiono, A.B., Diarra, A., Sanon, S., Yaro, J.B., Ouedraogo, E., Bougouma, E.C., Soulama, I., Gansané, A., Ouedraogo, A., Konate, A.T., Nebie, I., Watson, N.L., Sanza, M., Dube, T.J.T. and Sirima, S.B. (2013) 'Malaria Morbidity in High and Seasonal Malaria Transmission Area of Burkina Faso', *PLoS ONE* 8 (1), e50036.

Paciorek, C.J. and Schervish, M.J. (2006) 'Spatial Modelling Using a New Class of Nonstationary Covariance Functions.', *Environmetrics.* 17 (5), pp. 483–506.

Paciorek, C.J. (2003) 'Nonstationary Gaussian Processes for Regression and Spatial Modelling', PhD thesis, Carnegie Mellon University.

Plagemann, C., Kersting, K. and Burgard, W. (2008) 'Nonstationary Gaussian Process Regression using Point Estimates of Local Smoothness', *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II*, ed. by W. Daelemans, B. Goethals and K. Morik, 1st ed., Berlin Heidelberg: Springer-Verlag, chap. 13, pp. 204–219.

Roca-Feltrer, A., Lalloo, D.G., Phiri, K. and Terlouw, D.J. (2012) 'Short Report : Rolling Malaria Indicator Surveys (rMIS): a potential district-level malaria monitoring and evaluation (M & E) tool for program managers.', *American Journal of Tropical Medicine and Hygiene* 86 (1), pp. 96–98.

Schabenberger, O. and Gotaway, C.A. (2005) *Statistical Methods for Spatial Data Analysis*, Boca Raton: Chapman & Hall/CRC.

Shaddick, G. and Zidek, J.V. (2014) 'A case study in preferential sampling: Long term monitoring of air pollution in the UK', *Spatial Statistics* 9, pp. 51–65.

Stanton, M.C. and Diggle, P.J. (2013) 'Geostatistical analysis of binomial data: generalised linear or transformed Gaussian modelling?', *Environmetrics* 24 (3), pp. 158–171.

Stein, M.L. (1999) *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer.

Steketee, R.W., Nahlen, B.L., Parise, M.E. and Menendez, C. (2001) 'The burden of malaria in pregnancy in malaria-endemic areas', *American Journal of Tropical Medicine and Hygiene* 64 ((1, 2) Supplementary), pp. 28–35.

Van der Hoek, W., Konradsen, F., Amerasinghe, P.H., Perera, D., Piyaratne, M.K. and Amerasinghe, F.P. (2003) 'Towards a risk map of malaria for Sri Lanka: The importance of house location relative to vector breeding sites', *International Journal of Epidemiology* 32 (2), pp. 280–285.

Woolhouse, M.E., Dye, C., Etard, J.F., Smith, T., Charlwood, J.D., Garnett, G.P., Hagan, P., Hii, J.L., Ndhlovu, P.D., Quinnell, R.J., Watts, C.H., Chandiwana, S.K. and Anderson, R.M. (1997) 'Heterogeneities in the transmission of infectious agents: implications for the design of control programs.', *Proceedings of the National Academy of Sciences of the United States of America* 94 (1), pp. 338–342.

World Health Organisation (2012) *World Malaria Report 2012*, tech. rep., Geneva: World Health Organisation.

World Health Organisation (2013) *World Malaria Report 2013*, tech. rep., Geneva: World Health Organisation.

World Health Organisation (2015a) *Global technical strategy for malaria 2016-2030*, tech. rep., Geneva: World Health Organisation.

World Health Organisation (2015b) *World Malaria Report 2015*, tech. rep., Geneva: World Health Organisation.

World Health Organisation (2016) *World Malaria Report 2016*, tech. rep., Geneva: World Health Organisation.

Zidek, J.V., Shaddick, G. and Taylor, C.G. (2014) ‘Reducing estimation bias in adaptively changing monitoring networks with preferential site selection’, *The Annals of Applied Statistics* 8 (3), pp. 1640–1670.

Chapter 2

Paper 1. Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure.

Chipeta, M. G.[a,b,c], Terlouw, D. J.[a,c,d], Phiri, K. S.[a] and Diggle, P. J.[b] (2016). Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. ***Environmetrics***, *doi:10.1002/env.2425*

[a]College of Medicine, University of Malawi, Blantyre, Malawi.

[b]Lancaster Medical School, Lancaster University, Lancaster, UK.

[c]Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Blantyre, Malawi.

[d]Liverpool School of Tropical Medicine, Liverpool, UK.

Abstract

The problem of choosing spatial sampling designs for investigating an unobserved spatial phenomenon $\mathcal{S}$ arises in many contexts, for example in identifying households to select for a prevalence survey to study disease burden and heterogeneity in a study region $\mathcal{D}$. We studied randomised inhibitory spatial sampling designs to address the problem of spatial prediction whilst taking account of the need to estimate covariance structure. Two specific classes of design are *inhibitory designs* and *inhibitory plus close pairs designs.* In an inhibitory design, any pair of sample locations must be separated by at least an inhibition distance δ. In an inhibitory plus close pairs design, $n - k$ sample locations in an inhibitory design with inhibition distance δ are augmented by k locations each positioned close to one of the randomly selected $n - k$ locations in the inhibitory design, uniformly distributed within a disc of radius ζ. We present simulation results for the Matérn class of covariance structures. When the nugget variance is non-negligible, inhibitory plus close pairs designs demonstrate improved predictive efficiency over designs without close pairs. We illustrate how these findings can be applied to the design of a rolling Malaria Indicator Survey that forms part of an ongoing large-scale, five-year malaria transmission reduction project in Malawi.

Keywords. *Non-adaptive sampling strategies*; *Spatial statistics*; *Inhibitory designs*; *Prevalence mapping.*

2.1 Introduction.

Geostatistics is concerned with investigation of an unobserved spatial phenomenon $\mathcal{S} = \{S(x) : x \in \mathcal{D} \subset \mathbb{R}^2\}$, where $\mathcal{D}$ is a geographical region of interest. Its particular focus is on investigations in which the available data consist of measurements y_i at a finite set of locations $x_i \in \mathcal{D}$. Typically, each y_i can be regarded as a noisy version of $S(x_i)$. We write $\mathcal{X} = \{x_1, ..., x_n\}$ and call $\mathcal{X}$ the *sampling design*. Geostatistical *analysis* mainly addresses two broad scientific objectives: *estimation* of the parameters that define a stochastic model for the unobserved process $\mathcal{S}$ and the observed data $Y = \{(y_i, x_i) : i = 1, ..., n\}$; *prediction* of the unobserved realisation of $S(x)$ throughout $\mathcal{D}$, or particular characteristics of this realisation, for example its average value. The fundamental geostatistical *design* problem is the specification of $\mathcal{X}$. A key consideration is that sampling designs that are efficient for parameter estimation may be inefficient for prediction, and vice versa (Zimmerman, 2006). In practice, most geostatistical problems focus on spatial prediction, but parameter estimation is an important means to this end. Hence, there is a need to compromise between designing for efficient parameter estimation and designing for efficient prediction given the values of relevant model parameters. In practice, selection of covariates (which must be known at all observed locations) and estimating their effects are also important considerations for study design. However, in this paper we focus on the design implications of the spatial covariance structure of $\mathcal{S}$, this being the distinguishing feature of geostatistical, as opposed to general statistical, methodology.

In a previous paper (Chipeta et al., 2016a), we have discussed *adaptive* geostatistical designs, in which sampling locations are chosen sequentially, either singly or in batches, and at any stage the analysis of already collected data can inform the selection of the next batch of locations. In this paper, we consider *non-adaptive* geostatistical designs, in which the complete design $\mathcal{X}$ must be chosen in advance of any data-collection.

Two examples of non-adaptive designs are *completely random* and *lattice* designs. In a completely random design, the locations x_i are an independent random sample from the uniform distribution on $\mathcal{D}$. In a lattice design, the x_i form a regular (typically square) lattice to cover $\mathcal{D}$. A combination of theoretical and empirical work, from Matérn (1960) onwards, has led to general acceptance that lattice designs should lead to efficient spatial prediction provided model parameters are known. If model parameters are unknown, a completely random design has the advantage that it will include a wider range of inter-point distances, and in particular some small inter-point distances, and so provides more information on the shape of the covariance function of $\mathcal{S}$. However, the resulting uneven spatial distribution of the x_i makes prediction less efficient, given the model parameters. Diggle and Lophaven (2006) described and compared empirically some compromise designs. In their simulations, a lattice design supplemented by some close pairs of points performed well.

A limitation of lattice-based designs is that their absence of a probability sampling frame leaves open the possibility of systematic bias. In the present paper, we, therefore, propose a class of randomised *inhibitory plus close pairs* designs to address the problem of spatial prediction whilst taking account of the need to estimate spatial covariance structure. We evaluate the performance of this class of designs through simulation studies and describe an application to data from a malaria transmission reduction monitoring and evaluation study in the Chikwawa district of southern Malawi.

In Section 2.2 we review the existing literature on non-adaptive geostatistical design strategies. In Section 2.3 we describe our proposed class of designs. In Section 2.4 we describe a class of empirical kriging (EK) optimal designs. Section 2.5 reports on simulation studies of the predictive performance of the proposed design class. We also compare the performance of our proposed designs with EK optimal designs. Section 2.6 describes an application to the sampling design of an ongoing malaria prevalence mapping exercise around the perimeter of the Majete Wildlife Reserve, Chikwawa district, Malawi. Section 2.7 is a concluding discussion. All computations for the paper were run on the High-End Computing Cluster at Lancaster University, using the `R` software environment (R Core Team, 2015).

2.2 Non-adaptive geostatistical design strategies.

Different scientific goals and study settings require different geostatistical design strategies. Ideally, a design $\mathcal{X}$ will be chosen to maximise or minimise a performance criterion that reflects the primary objective of the study (Jardim and Ribeiro, 2007; Nowak, 2010). For example, a possible design criterion when the objective is to predict the value of $S(x)$ throughout the region $\mathcal{D}$ is the spatially averaged mean squared prediction error,

$$MSPE = \int_D \mathrm{E}[\{\hat{S}(x) - S(x)\}^2 dx, \tag{2.1}$$

where $\hat{S}(x) = \mathrm{E}[S(x)|Y;\mathcal{X}]$ is the minimum mean square error predictor of $S(x)$ and expectations are with respect to S. In practice, any such criterion needs to be tempered by application-specific considerations of some kind, for example, different costs and benefits of obtaining data and predictions, respectively, at particular locations.

We review the following strategies for geostatistical designs: designing for efficient parameter estimation; designing for efficient spatial prediction when the covariance function is assumed completely known; and designing for efficient spatial prediction when the covariance function is not known and has to be estimated from the same data. Müller (2007, Chapters 5 – 7) is a relatively recent book-length account of geostatistical design strategies.

Much of the work on spatial sampling design for estimating covariance structures has focused on estimation procedures based on the empirical variogram (Russo, 1984; Warrick and Myers, 1987; Müller and Zimmerman, 1999). Lark (2002) used likelihood estimation procedures under an assumed Gaussian process model. Pettitt and McBratney (1993) studied several sampling designs for estimating parameters using the restricted maximum likelihood (REML) method of parameter estimation. A general consensus from this body of work is that completely random designs are efficient for parameter estimation. However, these designs have often been criticised because they leave large unsampled swaths in the study region $\mathcal{D}$ (Müller, 2007).

Studies of design for efficient spatial prediction with known covariance structure include McBratney, Webster and Burgess (1981), McBratney and Webster (1981), Yfantis, Flatman and Behar (1987), Ritter (1996) and Su and Cambanis (1993). Spatially regular lattice designs, which achieve an even coverage of $\mathcal{D}$, have been shown to be optimal in this case. Other design constructions have also been proposed, collectively known as *spatially balanced* designs, whose common feature is that they result in a more even coverage of $\mathcal{D}$ than does the completely random design. We provide definitions and an overview in Section 2.2.1.

The assumption of a known covariance function is in most cases unrealistic (Müller, 2007). Usually, we have to use the same data for estimation of covariance parameters and for spatial prediction, and effective prediction requires good estimates of the second order characteristics (Guttorp and Sampson, 1994; Müller, Pronzato et al., 2015). Recent work on construction of designs that focus on the goals of efficient spatial prediction in conjunction with parameter estimation includes Zhu (2002), Zhu and Stein (2006), Diggle and Lophaven (2006), Pilz and Spöck (2006), Zimmerman (2006), Banerjee et al. (2008), Bijleveld et al. (2012), Müller, Pronzato et al. (2015) and Chipeta et al. (2016a).

2.2.1 Classes of non-adaptive geostatistical designs.

We now review several design classes that have been used for different analysis objectives: parameter estimation; spatial prediction; and a combination of the two. Design performance is largely influenced by *sample pattern* and *sample density* (Olea, 1984). '*Pattern*' here refers to the geometrical configuration of sample points in a given region, $\mathcal{D}$. '*Density*' refers to the number of sample points per unit area. Both model-dependent and purely geometrical designs have been proposed.

2.2.1.1 Completely randomised designs.

In a completely randomised design, locations $x_i,\ i = 1, \ldots, n$ are chosen independently, each with a uniform distribution over $\mathcal{D}$. This ensures that the design is stochastically independent of the underlying spatial phenomenon of interest $S(x)$, which is a requirement for the validity of standard geostatistical inference methods (Diggle, Menezes and Su, 2010). However, the resulting uneven coverage of $\mathcal{D}$ has a negative impact on spatial prediction. Variants of the completely random design include stratified and cluster random sampling (Cressie, 1991). These design strategies are well established in classical survey sampling; see, for example, Cochran (1977).

2.2.1.2 Completely regular lattice designs.

Design points in this class form a regular lattice pattern over the study region $\mathcal{D}$, thereby ensuring an even coverage. The origin of the lattice should strictly be located at random (Diggle and Ribeiro, 2007), although in practice this is often ignored. These designs are easy to implement and provide well defined directional classes within which variograms can be computed. Regular designs also have the potential of yielding computational savings over irregular designs such as those resulting from random sampling (Cressie, 1991). Regular lattice designs can use square, equilateral triangular or hexagonal grids. A comparison of the three suggests that the equilateral triangular grid design is the most efficient (McBratney, Webster and Burgess, 1981; McBratney and Webster, 1981; Olea, 1984; Yfantis, Flatman and Behar, 1987). However, square lattices are more common in practice. Regular lattice-based designs are commonly applied in remote sensing applications, see, for example, Atkinson (1991), Atkinson, Webster and Curran (1992) and Curran and Atkinson (1998).

2.2.1.3 Other constructions for spatially balanced designs.

Generalised random-tessellation stratified designs (GRTS) are widely used in environmental monitoring surveys. They represent a flexible technique for selecting a spatially balanced, probability sampling design (Stevens and Olsen, 2004; Grafström, Lundström and Schelin, 2012; Brown, Robertson and McDonald, 2015) in which each potential sampling location has a known, non-zero probability of being included in the sample. The design ensures that no points in the target population are too far from a sampled point (i.e., points are spread evenly) (Brown, Robertson and McDonald, 2015) and that few sampled points are close together.

A GRTS design is formulated using a restricted randomisation, referred to as hierarchical randomisation (HR), which randomly orders the spatial addresses (Stevens and Olsen, 2003). The construction proceeds in the following manner (Stevens and Olsen, 2004). Firstly, randomly place a 2×2 square grid over the region and place the cells in random order in a line. Secondly, for each cell, repeat the same process, randomly ordering the sub-cells within each original cell. This second step results in 16 cells in a line. Continue the process until at most one population point occurs in a cell. The random order of the cells is then used to place the points on the line. See Stevens and Olsen (1999), Stevens and Olsen (2003) and Stevens and Olsen (2004) for details.

Grafström, Lundström and Schelin (2012) used a *pivotal method* to construct designs with a high degree of spatial balance. The main purpose of the pivotal method is to construct designs that restrict locations/units that are close in distance from appearing together in the sample, which in turn creates an evenly spread sample. Brown, Robertson and McDonald (2015) extended the GRTS to a balanced acceptance sampling (BAS) design, that allows surveys to be balanced in more than two dimensions. BAS design uses acceptance/rejection sampling algorithm (Flury, 1990), that is if a generated sample point is beyond the edge of the sample space, the sample unit is rejected, otherwise, it is accepted.

Diggle and Lophaven (2006) proposed and developed two different two-step *augmented lattice* designs. These designs supplement a lattice with closely spaced pairs of points which, as noted earlier, are important for estimating certain parameters of the underlying spatial covariance structure, especially when this includes a nugget variance (Diggle and Ribeiro, 2007, Chapter 8) or a smoothness parameter such as the shape parameter of a Matérn correlation function (Zhu and Stein, 2006). In particular, a *lattice plus close pairs* design consists of an initial set of locations in $\mathcal{D}$ that form a $k \times k$ regular lattice at spacing Δ, augmented by a further m locations, each distributed uniformly at random within a disc of radius $\delta = \alpha\Delta$ centred on each of $m \leq k^2$ randomly selected lattice locations. A *lattice plus infill* design class is again initialised with an even coverage of $k \times k$ regular lattice at spacing Δ but is augmented with further locations in a more finely spaced lattice within m randomly selected primary lattice cells.

Royle and Nychka (1998) describe a purely geometric design criterion for spatial prediction. This approach, commonly known as 'space-filling' design, identifies sample locations by minimising a criterion that favours more regular geometrical configurations of sample locations (Nychka and Saltzman, 1998).

2.2.1.4 Summary.

Some general conclusions are the following. Good spatial prediction favours designs that are spatially more regular than a completely random design when model parameters are known. When the analysis objective is parameter estimation, designs with a random configuration of design points are preferable. These two points suggest that some compromise is therefore needed when constructing designs for spatial prediction when model parameters have to be estimated from the same data.

A good geostatistical design strategy also needs to be able to deal with a range of practical constraints. For example, potential sampling points may be limited to a finite set. This holds, for example, in our application to malaria monitoring, where data can only be collected from existing houses, within the study region.

2.3 Inhibitory geostatistical designs.

2.3.1 Design criterion.

We propose a class of *inhibitory* geostatistical designs for spatial prediction when model parameters need to be estimated. We use $[\cdot]$ to mean "the distribution of" and incorporate a stochastic process $\mathcal{S} = \{S(x) : x \in \mathcal{D} \subset \mathbb{R}^2\}$ into a statistical model $[S, Y] = [S][Y|S]$, where $Y = (Y_1, \ldots, Y_n)$ are the measured data values at the points of $\mathcal{X}$ and $S = \{S(x_1), \ldots, S(x_n)\}$. The distribution for estimation inference is then the conditional distribution, $[S|Y]$, which follows from an application of Bayes' theorem as

$$[S|Y] = [S][Y|S] / \int [S][Y|S] \mathrm{d}S \tag{2.2}$$

A typical spatial prediction problem involves making inferences about a functional

$T = \mathcal{T}(\mathcal{S})$ given data $(Y_i, X_i), \quad i = 1, \ldots, n$. We, therefore, extend the above factorisation to $[\mathcal{S}, Y] = [\mathcal{S}|S][S][Y|S]$. In what follows, we use as performance criterion the average prediction variance,

$$APV = \int_{\mathcal{D}} \text{Var}\{S(x)|Y\}\text{d}x \tag{2.3}$$

2.3.2 Simple inhibitory designs.

An inhibitory design consists of n locations chosen at random in $\mathcal{D}$ but with the constraint that no two locations are at a distance of less than some value δ. Formally, the resulting design $\mathcal{X}$ is a realisation of a simple inhibitory point process that is itself a special case of a pairwise interaction point process; see, for example, Diggle (2013, Chapter 6). This construction respects the established principles of random sampling theory while guaranteeing some degree of spatial regularity. All designs $\mathcal{X}$ that meet the inhibitory constraint are equally likely to be picked. Also, the construction can be applied whether or not the potential sampling locations are confined to a finite set of points, although in either case, the value of δ will limit the maximum achievable sample size.

We define the "*packing density*" of the design to be the proportion of the total region covered by n non-overlapping discs of diameter δ, hence $\rho = (n\pi\delta^2)/(4|\mathcal{D}|)$. We use the notation $\mathbf{SI}(n, \delta)$ and compare the performance of designs with fixed sample size n and varying δ. The formal construction of an $\mathbf{SI}(n, \delta)$ design on a region $\mathcal{D}$ proceeds as follows:

1. Draw a sample of locations $x_i : i = 1, \ldots, n$ completely at random in $\mathcal{D}$;

2. Set $i = 1$;

3. Calculate the minimum, d_{min}, of the distances from x_i to all other x_j in the current sample;

4. If $d_{min} \geq \delta$, increase i by 1 and return to step 3 if $i \leq n$, otherwise stop;

5. If $d_{min} < \delta$, replace x_i by a new location drawn completely at random in $\mathcal{D}$ and return to step 4.

2.3.3 Inhibitory design with close pairs.

This class is defined by four scalars, namely: n, the total number of points; δ, the minimum distance between any two locations; k, the number of close pairs and ζ, the radius of the disc from the primary point within which to add a paired point. For a total of n points, this design consists of $n - k$ points in an inhibitory design with inhibition distance δ, augmented by k points each positioned relative to one of the randomly selected $n - k$ points in the inhibitory design according to the uniform distribution over a disc of radius ζ. We use the notation $\mathbf{ICP}(n, k, \delta, \zeta)$. The formal construction of an $\mathbf{ICP}(n, k, \delta, \zeta)$ design on a region $\mathcal{D}$ proceeds as follows:

1. Construct a simple inhibitory design $\mathbf{SI}(n - k,\ \ \delta)$;

2. Sample k from $x_1, \ldots, x_{n-k}$ without replacement and call this set of locations $x_j^*,\ \ j = 1, \ldots, k$;

3. For $j = 1, \dots, k$, x_{n-k+j} is uniformly distributed on the disc with centre x_j^* and radius ζ.

Note that in the $\mathbf{ICP}(n, k, \delta, \zeta)$ design, k must be less than or equal to $n/2$. Also, when comparing an $\mathbf{SI}(n, \delta)$ design with one or more $\mathbf{ICP}(n, k, \delta, \zeta)$ designs, it is appropriate to require all of the inhibitory components to have the same degree of spatial regularity. This requires δ to become a function of k, namely

$$\delta_{(k)} = \delta_0 \sqrt{n/(n-k)}, \tag{2.4}$$

with δ_0 held fixed. For fixed n, the minimum spacing between any two inhibitory points, therefore, increases with k. We also insist that $\zeta \leq \delta_{(k)}/2$. Finally, when the potential sampling locations are restricted to a finite set of points $\{X_i,\ i = 1, \dots, N\}$, the above constructions are modified in an obvious way, with sampling at random from the N potential locations replacing uniform random sampling of points $x \in \mathcal{D}$, with the proviso that it will be impossible to construct an $\mathbf{ICP}(n, k, \delta, \zeta)$ design for some combinations of n, k, δ and ζ.

For fixed sample size n, region $\mathcal{D}$ and an assumed geostatistical model with a specific numerical value for its vector of parameters θ, we numerically optimise the above algorithms to determine the combination of k, δ and ζ that minimise the design criterion in Equation (2.3), using a general-purpose numerical optimiser. Specifically, we use the controlled random search (CRS) procedure for global optimisation (Price, 1976; Price, 1983). The procedure allows for box constraints that we impose on the design parameters of interest above.

2.4 Empirical kriging optimal designs.

In our simulation study (Section 2.5), we compare the performance of inhibitory plus close pairs design with some of the optimal designs we have reviewed in Section 2.1, such as empirical kriging (EK) designs implemented by Zimmerman (2006) and Müller, Pronzato et al. (2015). These designs minimise the empirical kriging criterion:

$$EK(\mathcal{X}) = \max_{x \in \mathcal{D}} \{\mathrm{Var}[\hat{Y}(x) - Y(x)] + \mathrm{tr}\{M_\theta \ \mathrm{Var}[\partial \hat{Y}(x)/\partial\theta]\}\}. \tag{2.5}$$

This adds an explicit additive correction term to the normalised classical prediction variance. In Equation (2.5), $\hat{Y}(x)$ is the posterior mean of $Y(x)$ given data at $\mathcal{X} = \{x_i; i = 1, \dots, n\}$ and M_θ is the covariance matrix of the estimated covariance parameters θ. The Estimation-Adjusted (EA) criterion implemented by Zhu and Stein (2006) is similar in spirit to the EK criterion. Both of these obtain specific designs by a spatial simulated annealing (SSA) search algorithm (van Groenigen and Stein, 1998; van Groenigen, Siderius and Stein, 1999; Lark, 2002). These methods are much more computationally expensive, and the resulting designs depend on the spatial locations of a set of specified potential sampling points in a more complicated way, than do our proposed **ICP**(n, k, δ, ζ) designs. In our simulation study in Section 2.5.3, we follow the SSA algorithm outlined in Müller, Pronzato et al. (2015).

2.5 Simulation studies.

We have carried out simulation studies of our proposed designs to illustrate the gains in predictive efficiency that can be achieved using inhibitory designs when covariance parameters have to be estimated. In our simulation studies, we evaluate our performance criterion (Equation (2.3)) at the estimated parameter values using the plug-in prediction method (Diggle and Ribeiro, 2007). We simulate data on the unit square $[0, 1]^2$, evaluate the integral in Equation (2.3) by numerical quadrature over a 64 × 64 prediction grid, and approximate the expectation of the integral by a Monte Carlo average over $s = 1500$ independent simulations of measurement data Y. We consider two model classes for the data-generation process, namely the linear Gaussian and logistic binomial geostatistical models. Both include an unobserved stationary Gaussian process $S(x)$ with mean zero, variance $\sigma^2 = 1$ and Matérn correlation (Matérn, 1960).

In the linear Gaussian model,

$$Y|S \sim N(\mu, \tau^2) \tag{2.6}$$

where $\mu = S(x)$, whilst in the logistic binomial model,

$$Y|S, U \sim Bin(n, p), \tag{2.7}$$

where $\log(p/1-p) = S(x) + U$ and U is Gaussian white noise with variance τ^2. In both cases, the predictive target is $\mathcal{S}$.

We used a fixed value of the correlation shape parameter, $\kappa = 1.5$, but varied the correlation range parameter ϕ and the nugget variance τ^2.

2.5.1 Linear Gaussian Model.

For each parameter combination, we generated data at $n = 150$ sampling locations. Figure 2.1a shows an inhibitory design without close pairs and $\delta = 0.06$, corresponding to packing density $\rho \approx 0.424$, whilst Figure 2.1b shows a design with $k = 75$ close pairs and $\delta_{(k)} = 0.085$ so that the $n - k = 75$ inhibitory design points also have packing density 0.424. Note that a maximum $\delta = 0.06$ is an arbitrary choice to allow the construction of a more-regular-than-random design.

Figure 2.2 shows the design performance as δ varies between 0.01 and 0.06, $\phi = 0.15, 0.20, 0.25$ and 0.30, and for noise-to-signal ratios $\tau^2 = 0$ and 0.2. Results (not shown) for $\tau^2 = 0.05$, 0.1 and 0.4 show similar trends. These results indicate that designs with larger δ perform better, i.e. spatial predictions become more precise with increasing regularity of the design.

Our comparison of inhibitory designs with and without close pairs indicates that designs with an intermediate number of close pairs give the best performance. However, when τ^2 is close to zero the benefits of close pairs are negligible, see Figure 2.3 panels A – B. In contrast, when τ^2 is larger, close pairs show substantial benefit, see Figure 2.3 panels C – E.

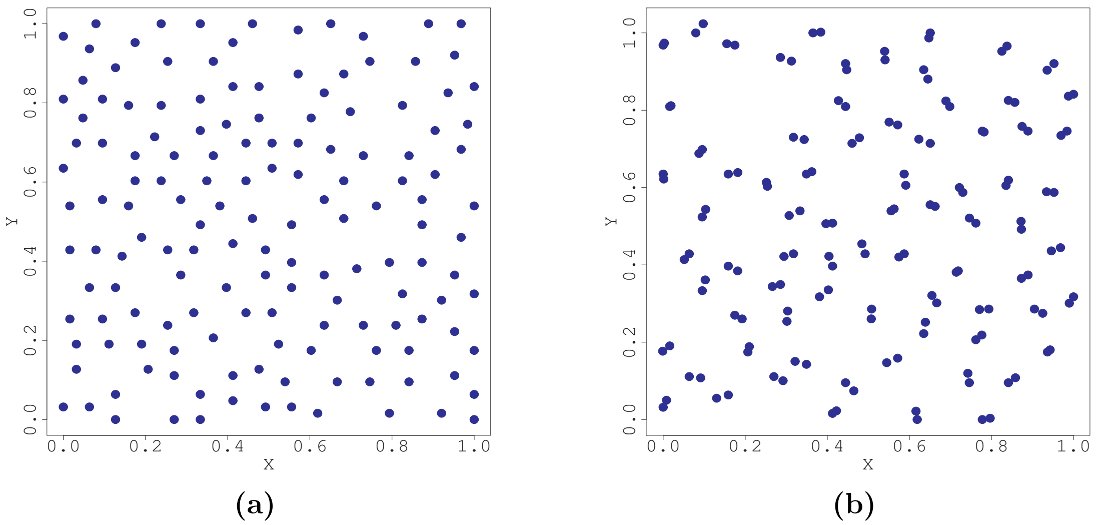


Figure 2.1: Simple inhibitory design, $\delta = 0.06$ (a). Inhibitory design with $k = 75$ close pairs, $\delta_{(k)} = 0.085$ for $n - k$ inhibitory design points (b). The inhibitory distance δ for (b) varies with the number of close pairs k. Sample size $n = 150$ for each of the designs.

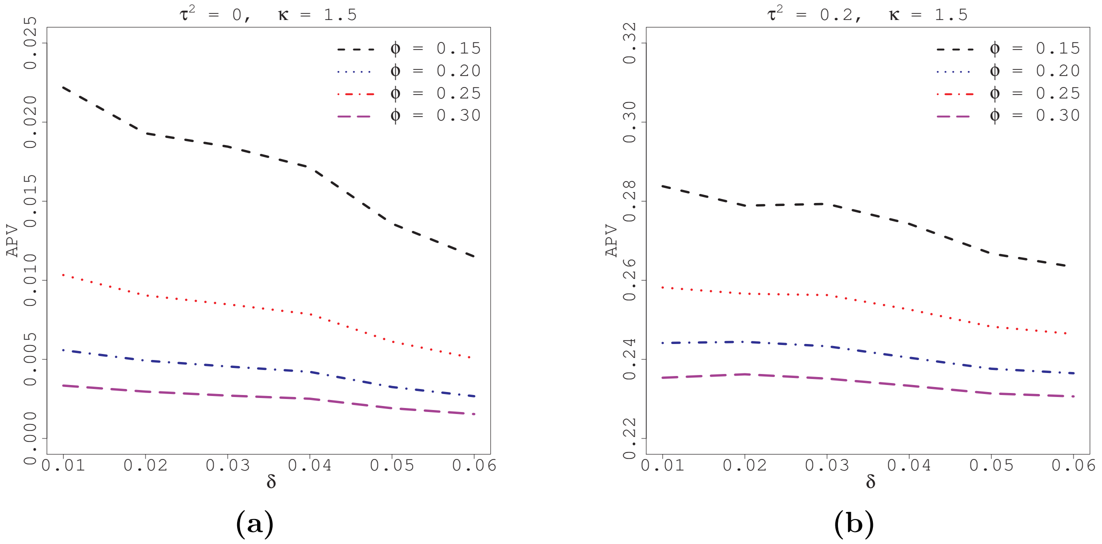


Figure 2.2: Average prediction variance for varying simple inhibitory designs, $\delta = 0.01$ to 0.06, $\kappa = 1.5, \sigma^2 = 1$ and $n = 150$. Panel (a) $\tau^2 = 0$ and panel (b) $\tau^2 = 0.2$.

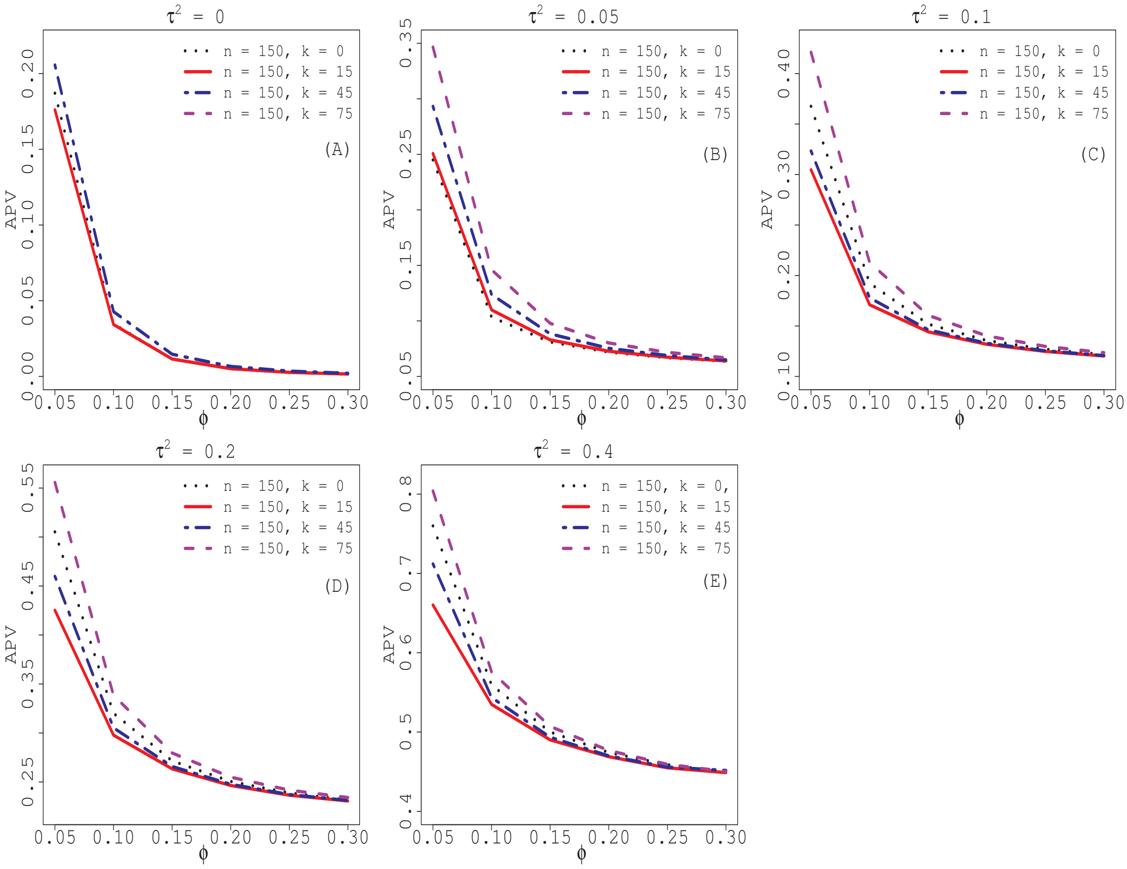
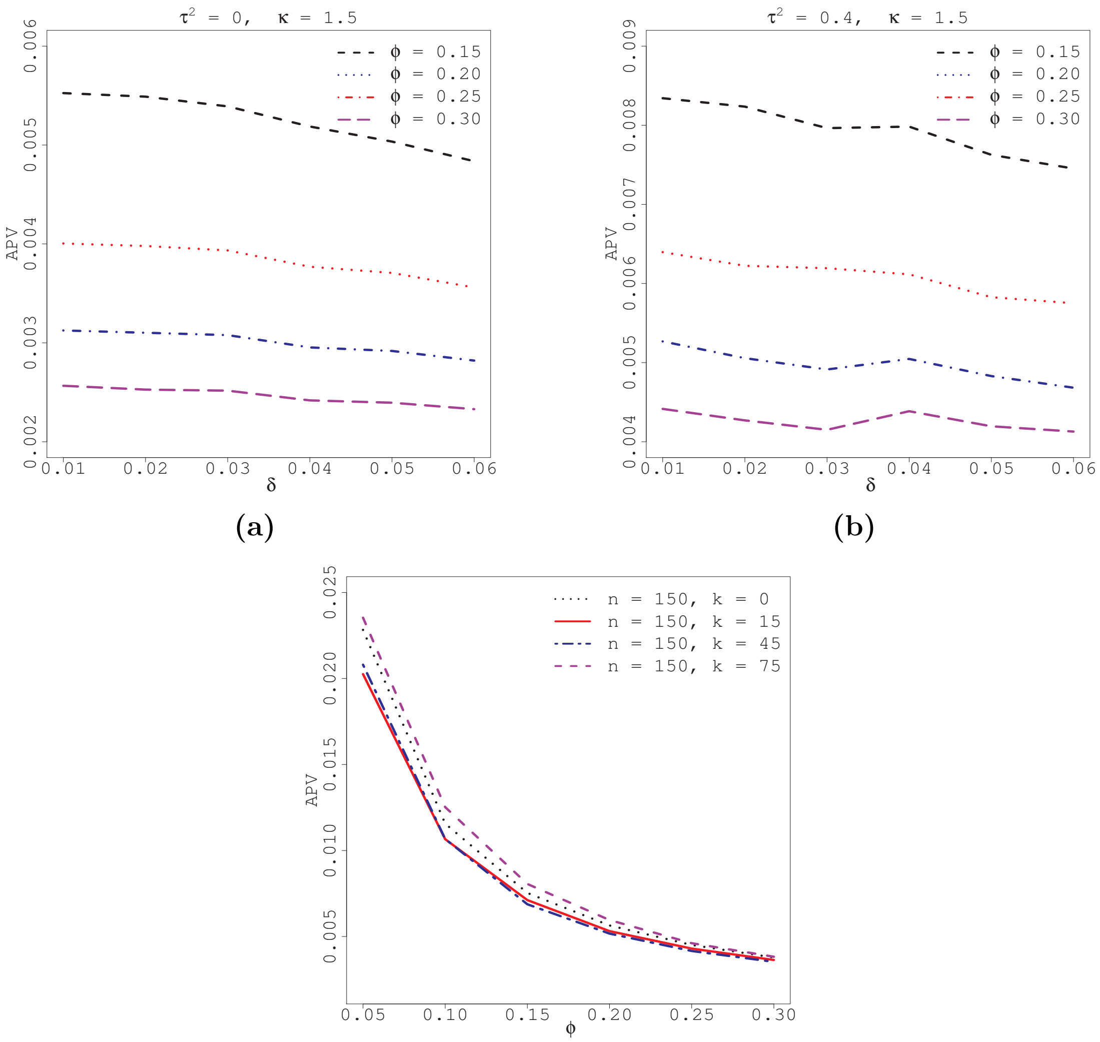


Figure 2.3: Comparing the efficiencies of inhibitory designs: without close pairs, with 15, 45 and 75 close pairs. The fixed total $n = 150$ for each of the designs.

2.5.2 Binomial Model.

We simulated binomial datasets with 10 trials at each of $n = 150$ grid points, and probabilities given by the anti-logit of the simulated values of the Gaussian process. For each combination of parameters, we approximated the expectation in Equation (2.3) by a Monte Carlo average over $s = 1000$ independent simulations of Y. Figures 2.4a to 2.4b show that inhibitory designs with $\delta = 0.06$ give the best results, agreeing with the findings in Section 2.5.1, Figure 2.2. Similarly, Figure 2.4c again shows that inhibitory designs with an intermediate number of close pairs give the best performance when τ^2 is relatively large.



(a) **(b)**

(c)

Figure 2.4: Average prediction variance for varying simple inhibitory designs - Binomial model, $\delta = 0.01$ to 0.06, $\kappa = 1.5, \sigma^2 = 1$ and $n = 150$. Panel (a) $\tau^2 = 0$ and panel (b) $\tau^2 = 0.4$. Panel (c) compares the efficiencies of inhibitory designs with 15, 45 and 75 close pairs. The fixed total $n = 150$ for each of the designs.

2.5.3 ICP vs EK optimal designs.

We simulate data on the unit square $[0,\, 1]^2$ and construct each of the designs using their respective algorithms as described in Section 2.3.3 and Section 2.4, with a fixed sample size $n = 35$. The ICP design has $k = 5$, $\delta_{(k)} = 0.076$ and $\zeta = 0.025$. We consider the linear Gaussian geostatistical model (Equation (2.6)) for the data-generation process. This includes an unobserved stationary Gaussian process $S(x)$

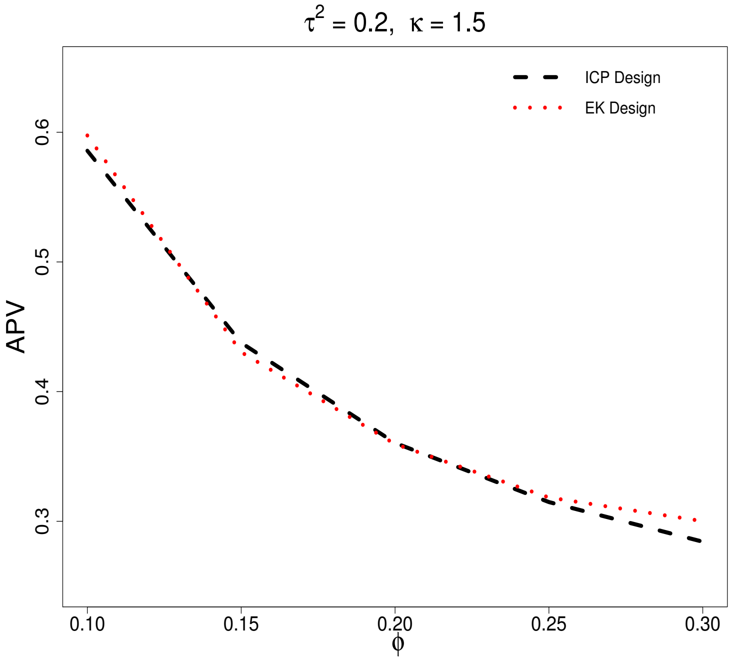


Figure 2.5: Inhibitory plus close pairs design vs Empirical kriging optimal design.

with mean zero, variance $\sigma^2 = 1$ and a Matérn correlation. We evaluate the integral in Equation (2.3) by numerical quadrature over a 7 × 7 prediction grid and approximate the expectation of the integral by a Monte Carlo average over s $=$ 10000 independent simulations of measurement data Y. Figure 2.5 shows results for comparison between numerically optimised ICP and EK optimal designs for θ with fixed variance $\sigma^2 = 1$, fixed noise-to-signal ratio $\tau^2 = 0.2$ and varying $\phi =$ 0.10, 0:15; 0:20; 0:25 and 0.30. In each case, the two optimised designs achieve similar values of the average prediction variance. Here, we have only made a limited set of comparisons due to computational limitations for the EK optimal designs. We elaborate on this point later in the discussion.

2.6 Application: sampling to predict spatial variation in malaria prevalence in the Majete perimeter.

In this section, we illustrate the use of our proposed inhibitory design strategy to construct a survey sample for mapping malaria prevalence in an area surrounding Majete Wildlife Reserve (MWR) within Chikwawa district, Malawi. The MWR is situated in the lower Shire valley at the edge of the African Rift Valley in the southern part of Malawi (15.97° S; 34.76° E). The reserve is crossed by two perennial rivers, the Shire and Mkurumadzi Rivers. Mwanza River runs near the western and southern boundaries of the park. In the wet season, there are also seasonal pools and many seasonal streams. Most rainfall occurs during the wet season, which lasts from November to April. Annually, the precipitation is 680 to 800 mm in the eastern lowlands and 700 to 1000 mm in the western highlands (Wienand, 2013). With an average daily temperature of 28.4 °C, the wet season is slightly warmer than the dry season (average daily temperature 23.3 °C), though the hottest months are September to November, at the end of the dry season (Staub, Binford and Stevens, 2013).

The Majete malaria project (MMP) is a five-year monitoring and evaluation study of malaria prevalence, with an embedded randomised trial of community-level interventions intended to reduce malaria transmission. The study takes place in the “Majete Perimeter”, which is the zone surrounding the MWR. The whole

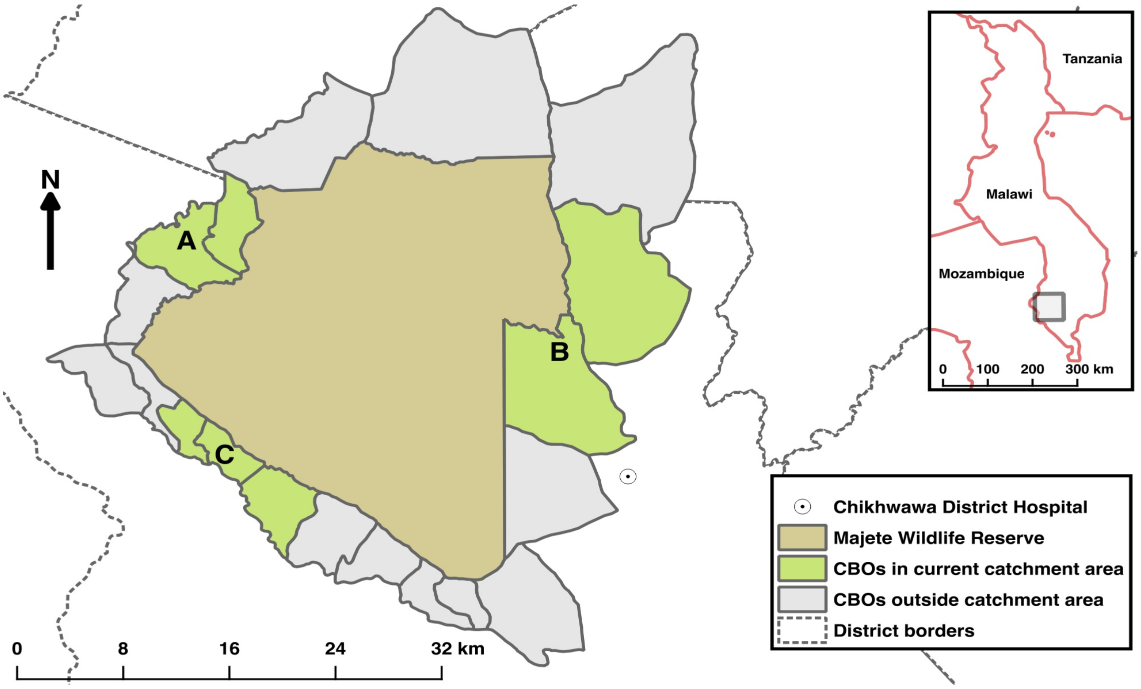


Figure 2.6: Map showing Majete Wildlife Reserve (brown) and borders of the 19 community-based organisations (CBOs) comprising the Majete perimeter. Three focal areas (green), labelled as A, B, and C, mark the communities selected for malaria indicator surveys and the trial. The rest of the CBOs (grey) are outside the project's catchment area.

perimeter is home to a population of approximately 100,000. Figure 2.6 shows the location of the study area, covering the unprotected zone surrounding the game park. The perimeter is subdivided into 19 community-based organizations (CBOs). In the MMP, three sets of these CBOs (CBOs – 1 & 2, CBOs –15 & 16 and CBOs – 6, 7 & 8) define *focal areas* A, B and C respectively. The first stage in the geostatistical design was a complete enumeration of households in the study region, including their geo-location collected using Global Positioning System (GPS) devices on a Samsung Galaxy Tab 3 running Android 4.1 Jellybean operating system. These devices are accurate to within 5 meters.

The sampling unit is a household. We first fit the Binomial model Equation (2.7), with three parameters representing the two variance components and the rate of

Table 2.1: Monte Carlo maximum likelihood estimates and 95 % confidence intervals for the covariance model fitted to malaria prevalence data in Majete focal area B.

Term	Estimate	95 % confidence interval	
Intercept	-1.90986	(-2.19000,	-1.62973)
σ^2	0.53016	(0.31787,	0.88422)
τ^2	0.26328	(0.07426,	0.93341)
ϕ^*	0.31913	(0.13320,	0.76459)

**Distance is given in kilometres.*

decay of spatial correlation with distance, to the “presence/absence” of malaria data from focal area B, then use the resulting estimated covariance model to inform an optimal sampling design for focal area A, whilst allowing for re-estimation of the model parameters. Table 2.1 shows the estimated covariance parameters. With these estimates, we used a general numerical optimiser (controlled random search) to determine the optimal design parameters that minimised the performance criterion in Equation (2.3). From a candidate set of 857 households we sampled a total of 200, the optimal design was found with $k = 24$ close paired locations, $\delta_{(k)} = 0.123$ km and $\zeta = 0.08$ km, see Figure 2.7. The blue dots represent the 176 inhibitory sample locations, red dots represent the 24 close pair locations and the black dots are the remaining 657 candidate locations. Note that the total sample size of 200 locations we used here is an arbitrary choice, chosen for illustration purposes only. The sampling locations provide a good spatial coverage of the study area, which is advantageous for efficient spatial prediction, whilst the inclusion of the close pairs is advantageous for parameter estimation.

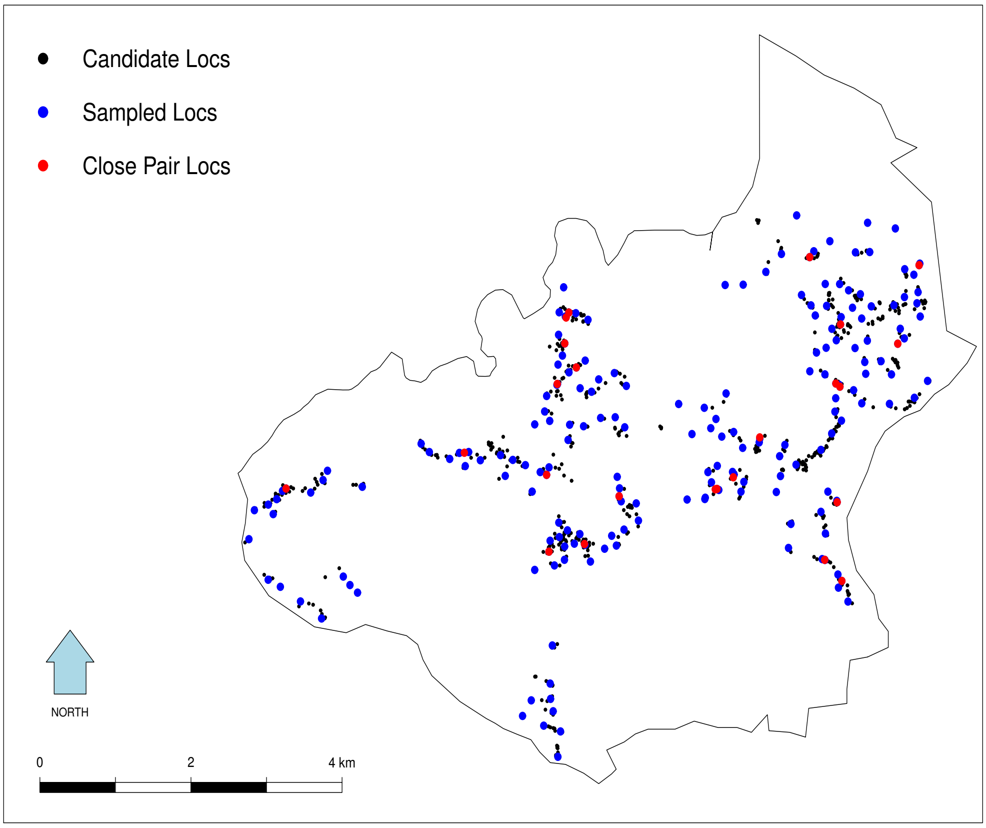


Figure 2.7: Inhibitory (blue dots) plus close pairs design locations (red dots) and all potential sampling locations (black dots), in focal area A

2.7 Discussion.

Parameter values are usually unknown in practice. Designing for efficient spatial prediction with estimated parameters involves a compromise. In this paper, we have proposed and demonstrated a class of inhibitory sampling designs for accurate spatial prediction with estimated covariance model parameters. The design strategies described in Section 2.3 are specifically intended to deliver efficient mapping of the complete surface, $S(x)$, over the region of interest. We considered inhibitory designs with and without close pairs of sampling locations. Inhibitory designs are random designs that generate spatially regular configurations of design points.

Our proposed designs incorporate the widely accepted concept that spatial prediction is improved by using a more-regular-than-random configuration of sampling locations (Olea, 1984). Our simulation studies show that when the same data are used for both parameter estimation and spatial prediction, the optimum inhibitory design includes a small proportion of close pairs (between 10 % and 30 % in our examples). This is consistent with previously expressed views that in order to compromise between prediction accuracy and efficient parameter estimation, optimal geostatistical designs should include close pairs in an otherwise spatially regular design (Lark, 2002; Diggle and Lophaven, 2006; Müller, 2007). However, our results also show that with our proposed class of designs, clear benefits for including close pairs are only realised when the nugget variance is relatively large. In our case, we conjecture that this is a consequence of the fact that inhibitory designs avoid the rigidity of lattice designs, resulting in a more varied set of inter-point distances. This is consistent with findings of Zimmerman (2006). He found that the EK-optimal design resembled the optimal design for prediction with known covariance parameters (which is spatially very regular) when the nugget effect was small and the spatial correlation is strong, whereas when the nugget effect is large (50 % of total variance) the EK-optimal design consists of small clusters of sites regularly dispersed throughout the study area, regardless of the strength of spatial correlation.

Our comparison of ICP and EK optimal designs showed that they exhibit similar performance in terms of prediction variance. This is consistent with previous findings that, for a fixed design $\mathcal{X}$, the influence of the correction term in Equation (2.5) diminishes with increasing sample size n. Müller, Pronzato et al. (2015) showed that for a design with $n \geq 10$, $\max_{\mathrm{x}\in\mathcal{D}}\mathrm{Var}[\hat{Y}(x) - Y(x)]$ and $EK(\mathcal{X}_n)$ yield similar values, implying that the effect of the correction term in Equation (2.5) becomes negligible as n increases. We suggest that, in the presence of a substantial nugget effect, the essential feature of both ICP and EK designs that results in their similar performance is their inclusion of small clusters of points in an otherwise regularly spaced design. For a large n, designs that minimise the classical prediction variance resemble the EK-optimal designs. However, as noted earlier and also in Zhu and Stein (2006) and Müller, Pronzato et al. (2015), spatial simulated annealing based EK-/EA- optimal designs are computationally very costly to construct, with each run taking at least 8 hours of central processor unit time. ICP designs can, therefore, be found more easily, quickly and inexpensively, with each run taking less than 30 minutes of central processor unit time. The computations that were reported in the paper were run on the High-End Computing Cluster at Lancaster University, using the `R` software environment (R Core Team (2015); see also https://www.r-project.org/). ICP designs can be implemented by the average practitioner more easily than similarly performing EK-/EA-optimal designs.

We have approached the sampling design problem assuming an underlying stochastic process with a stationary covariance structure. This is a common assumption in geostatistical applications. However, when explanatory variables are available their spatial distribution will also affect design performance. Numerical optimisation of a performance criterion such as Equation (2.3) in the presence of explanatory variables involves no additional principles.

Acknowledgements.

We thank the Majete integrated malaria control project (MMP) for providing the data used here. The MMP study was generously supported by Dioraphte Foundation, The Netherlands. Michael Chipeta holds a studentship from the UK Economic and Social Research Council. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

Funding.

Michael Chipeta is supported by ESRC-NWDTC Ph.D. studentship (grant number ES/J500094/1). Dr. Dianne Terlouw, Prof. Kamija Phiri and Prof. Peter Diggle are supported by the Majete integrated malaria control project grant.

References

Atkinson, P.M. (1991) ‘Optimal ground-based sampling for remote sensing investigations: estimating the regional meant’, *International Journal of Remote Sensing* 12 (3), pp. 559–567.

Atkinson, P.M., Webster, R. and Curran, P.J. (1992) ‘Cokriging with Ground Based Radiometry’, *Remote Sensing Environment* 41, pp. 45–60.

Banerjee, S., Gelfand, A.E., Finley, A.O. and Sang, H. (2008) ‘Gaussian predictive process models for large spatial data sets’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 70 (4), pp. 825–848.

Bijleveld, A.I., van Gils, J.A., van der Meer, J., Dekinga, A., Kraan, C., van der Veer, H.W. and Piersma, T. (2012) ‘Designing a benthic monitoring programme with multiple conflicting objectives’, *Methods in Ecology and Evolution* 3 (3), pp. 526–536.

Brown, J., Robertson, B.L. and McDonald, T. (2015) ‘Spatially Balanced Sampling: Application to Environmental Surveys’, *Procedia Environmental Sciences* 27, pp. 6–9.

Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2016a) 'Adaptive geostatistical design and analysis for prevalence surveys', *Spatial Statistics* 15, pp. 70–84.

Cochran, W.G. (1977) *Sampling Techniques*, 3rd ed., New York: John Wiley & Sons, Ltd.

Cressie, N. (1991) *Statistics for Spatial Data*, New York: Wiley.

Curran, P.J. and Atkinson, P.M. (1998) 'Geostatistics and remote sensing', *Progress in Physical Geography* 22 (1), pp. 61–78.

Diggle, P.J. (2013) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns.* 3rd ed., Boca Raton: CRC Press.

Diggle, P.J. and Lophaven, S. (2006) 'Bayesian geostatistical design', *Scandinavian Journal of Statistics* 33 (1), pp. 53–64.

Diggle, P.J., Menezes, R. and Su, T.-L. (2010) 'Geostatistical inference under preferential sampling', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59 (2), pp. 191–232.

Diggle, P.J. and Ribeiro, J.P. (2007) *Model-based Geostatistics*, New York: Springer.

Flury, B.D. (1990) 'Acceptance-Rejection Sampling Made Easy.', *Society for Industrial and Applied Mathematics* 32 (3), pp. 474–476.

Grafström, A., Lundström, N. and Schelin, L. (2012) 'Spatially Balanced Sampling through the Pivotal Method', *Biometrics* 68 (2), pp. 514–520.

Guttorp, P. and Sampson, P.D. (1994) ‘Methods for estimating heterogeneous spatial covariance functions with environmental applications’, *Handbook of Statistics* 12 (236), pp. 661–689.

Jardim, E. and Ribeiro, P.J. (2007) ‘Geostatistical assessment of sampling designs for Portuguese bottom trawl surveys’, *Fisheries Research* 85 (3), pp. 239–247.

Lark, R.M. (2002) ‘Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood’, *Geoderma* 105 (1-2), pp. 49–80.

Matérn, B. (1960) *Spatial Variation*, tech. rep., Stockholm: Statens Skogsforsningsinstitut.

McBratney, A.B., Webster, R. and Burgess, T.M. (1981) ‘The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalized Variables–I: Theory and method.’, *Computers & Geosciences* 7 (4), pp. 331–334.

McBratney, A.B. and Webster, R. (1981) ‘The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalised Variables–II: Program and Examples.’, *Computers & Geosciences* 7 (4), pp. 335–365.

Müller, W.G. (2007) *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*, 3rd ed., Berlin: Springer-Verlag.

Müller, W.G., Pronzato, L., Rendas, J. and Waldl, H. (2015) ‘Efficient prediction designs for random fields’, *Applied Stochastic Models in Business and Industry* 31 (2), pp. 178–194.

Müller, W.G. and Zimmerman, D.L. (1999) ‘Optimal designs for Variogram estimation’, *Enviromentrics* 10, pp. 23–37.

Nowak, W. (2010) ‘Measures of parameter uncertainty in geostatistical estimation and geostatistical optimal design’, *Mathematical Geosciences* 42 (2), pp. 199–221.

Nychka, D. and Saltzman, N. (1998) ‘Design of Air-Quality Monitoring Networks’, *Case Studies in Environmental Statistics SE - 4*, Lecture Notes in Statistics 132, ed. by D. Nychka, W. Piegorsch and L. Cox, pp. 51–76.

Olea, R.A. (1984) ‘Sampling design optimization for spatial functions’, *Journal of the International Association for Mathematical Geology* 16 (4), pp. 369–392.

Pettitt, A.N. and McBratney, A.B. (1993) ‘Sampling Designs for Estimating Spatial Variance Components’, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 42 (1), pp. 185–209.

Pilz, J. and Spöck, G. (2006) ‘Spatial sampling design for prediction taking account of uncertain covariance structure’, *7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.* Lisbon, Portugal, pp. 109–118.

Price, W.L. (1976) ‘A controlled random search procedure for global optimisation’, *The Computer Journal* 20 (4), pp. 367–370.

Price, W.L. (1983) ‘Global optimization by controlled random search’, *Journal of Optimization Theory and Applications* 40 (3), pp. 333–348.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*, https://www.R-project.org/, Vienna, Austria.

Ritter, K. (1996) ‘Asymptotic optimality of regular sequence designs’, *The Annals of Statistics* 24 (5), pp. 2081–2096.

Royle, J. and Nychka, D. (1998) ‘An algorithm for the construction of spatial coverage designs with implementation in SPLUS’, *Computers & Geosciences* 24 (5), pp. 479–488.

Russo, D. (1984) ‘Design of an Optimal Sampling Network for Estimating the Variogram’, *Soil Science Society of America Journal* 48 (4), pp. 708–716.

Staub, C.G., Binford, M.W. and Stevens, F.R. (2013) ‘Elephant herbivory in Majete Wildlife Reserve, Malawi’, *African Journal of Ecology* 51, pp. 536–543.

Stevens, D.L. and Olsen, A.R. (1999) ‘Spatially Restricted Surveys over Time for Aquatic Resources.’, *International Biometric Society* 4 (4), pp. 415–428.

Stevens, D.L. and Olsen, A.R. (2003) ‘Variance estimation for spatially balanced samples of environmental resources’, *Environmetrics* 14 (6), pp. 593–610.

Stevens, D.L. and Olsen, A.R. (2004) ‘Spatially Balanced Sampling of Natural Resources’, *Journal of the American Statistical Association* 99 (465), pp. 262–278.

Su, Y.S.Y. and Cambanis, S. (1993) ‘Sampling Designs for Estimation of a Random Process’, *Stochastic Processes and their Applications* 46, pp. 47–89.

Van Groenigen, J.W., Siderius, W. and Stein, A. (1999) ‘Constrained optimisation of soil sampling for minimisation of the kriging variance’, *Geoderma* 87 (3-4), pp. 239–259.

Van Groenigen, J.W. and Stein, A. (1998) 'Constrained Optimization of Spatial Sampling using Continuous Simulated Annealing', *Journal of Environmental Quality* 27 (5), pp. 1078–1086.

Warrick, A.W. and Myers, D.E. (1987) 'Optimization of sampling locations for variogram calculations', *Water Resources Research* 23 (3), pp. 496–500.

Wienand, J. (2013) 'Woody vegetation change and elephant water point use in Majete Wildlife Reserve: implications for water management strategies', PhD thesis, Stellenbosch University.

Yfantis, E.A., Flatman, G.T. and Behar, J.V. (1987) 'Efficiency of Kriging Estimation for Square , Triangular , and Hexagonal Grids', *Mathematical Geology* 19 (3), pp. 183–205.

Zhu, Z. (2002) 'Optimal Sampling Design and Parameter Estimation of Gaussian Random Fields', PhD thesis, University of Chicago.

Zhu, Z. and Stein, M.L. (2006) 'Spatial sampling design for prediction with estimated parameters', *Journal of Agricultural, Biological, and Environmental Statistics* 11 (1), pp. 24–44.

Zimmerman, D.L. (2006) 'Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction', *Environmetrics* 17 (6), pp. 635–652.

Chapter 3

Paper 2. Adaptive Geostatistical Design and Analysis for Prevalence Surveys.

Chipeta, M. G.[a,b,c], Terlouw, D. J.[a,c,d], Phiri, K. S.[a] and Diggle, P. J.[b] Adaptive geostatistical design and analysis for prevalence surveys, ***Spatial Statistics.***, Volume 15, February 2016, Pages 70–84, http://dx.doi.org/10.1016/j.spasta.2015.12.004.

[a]College of Medicine, University of Malawi, Blantyre, Malawi.

[b]Lancaster Medical School, Lancaster University, Lancaster, UK.

[c]Malawi-Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi.

[d]Liverpool School of Tropical Medicine, Liverpool, UK.

Abstract

Non-adaptive geostatistical designs (NAGDs) offer standard ways of collecting and analysing geostatistical data in which sampling locations are fixed in advance of any data collection. In contrast, adaptive geostatistical designs (AGDs) allow collection of geostatistical data over time to depend on information obtained from previous information to optimise data collection towards the analysis objective. AGDs are becoming more important in spatial mapping, particularly in poor resource settings where uniformly precise mapping may be unrealistically costly and the priority is often to identify critical areas where interventions can have the most health impact. Two constructions are: *singleton* and *batch* adaptive sampling. In singleton sampling, locations x_i are chosen sequentially and at each stage, x_{k+1} depends on data obtained at locations $x_1, \ldots, x_k$. In batch sampling, locations are chosen in batches of size $b > 1$, allowing each new batch, $\{x_{(k+1)}, \ldots, x_{(k+b)}\}$, to depend on data obtained at locations $x_1, \ldots, x_{kb}$. In most settings, batch sampling is more realistic than singleton sampling. We propose specific batch AGDs and assess their efficiency relative to their singleton adaptive and non-adaptive counterparts using simulations. We then show how we are applying these findings to inform an AGD of a rolling Malaria Indicator Survey, part of a large-scale, five-year malaria transmission reduction project in Malawi.

Keywords. *Adaptive sampling strategies*; *Spatial statistics*; *Geostatistics*; *Malaria*; *Prevalence mapping.*

3.1 Introduction.

Geostatistics has its origins in the South African mining industry (Krige, 1951), and was subsequently developed by Georges Matheron and colleagues into a self-contained methodology for solving prediction problems arising principally in mineral exploration; Chilès and Delfiner (2012) is a recent book-length account. Within the general statistics research community, the term geostatistics more generally refers to the branch of spatial statistics that is concerned with investigating an unobserved spatial phenomenon $\mathcal{S} = \{S(x) : x \in \mathcal{D} \subset \mathbb{R}^2\}$, where $\mathcal{D}$ is a geographical region of interest, using data in the form of measurements y_i at locations $x_i \in \mathcal{D}$. Typically, each y_i can be regarded as a noisy version of $S(x_i)$. We write $\mathcal{X} = \{x_1, \ldots, x_n\}$ and call $\mathcal{X}$ the *sampling design.*

Geostatistical analysis can address either or both of two broad objectives: *estimation* of the parameters that define a stochastic model for the unobserved process $\mathcal{S}$ and the observed data $\{(y_i, x_i) : i = 1, ..., n\}$; *prediction* of the unobserved realisation of $S(x)$ throughout $\mathcal{D}$, or particular characteristics of this realisation, for example its average value.

A key consideration for geostatistical design is that sampling designs that are efficient for parameter estimation are generally inefficient for prediction, and vice versa - see, for example, Diggle and Ribeiro (2007) and Müller (2007). Since parameter values are usually unknown in practice, design for prediction, therefore, involves a compromise. Furthermore, the diversity of potential predictive targets requires design strategies to be context-specific. Another important distinction is between *non-adaptive* sampling designs that must be completely specified prior to data-collection, and *adaptive* designs, for which data are collected over a period of time and later sampling locations can depend on data collected from earlier locations.

In this paper we formulate, and evaluate through simulation studies, a class of adaptive design strategies that address two compromises: between efficient parameter estimation and efficient prediction; and between theoretical advantages and practical constraints. The motivation for our work is the mapping of spatial variation in malaria prevalence in rural communities through a series of "rolling malaria indicator surveys," henceforth rMIS (Roca-Feltrer et al., 2012). rMIS is a malaria transmission monitoring and evaluation tool conducted on a monthly basis. Adaptive design is especially relevant here because resource constraints make it difficult to achieve uniformly precise predictions throughout the region of interest, hence as data accrue over the study region $\mathcal{D}$ it becomes appropriate to focus progressively on sub-regions of $\mathcal{D}$ where precise prediction is needed to inform public health action, for example, to prioritise sub-regions for early intervention.

In Section 3.2 we review the existing literature on adaptive geostatistical design and set out the methodological framework within which we will specify and evaluate adaptive design strategies. Section 3.3 describes our proposed class of adaptive designs for efficient prediction. Section 3.4 gives the results of a simulation study in which we compare the predictive efficiency of our proposed design strategy with simpler, non-adaptive strategies. Section 3.5 is an application to the design of an ongoing prevalence mapping exercise around the perimeter of the Majete Wildlife Reserve, Chikwawa district, southern Malawi through an rMIS that will be conducted monthly over a two-year period. Section 3.6 is a concluding discussion.

3.2 Methodological framework.

3.2.1 Geostatistical models for prevalence data.

The standard geostatistical model for prevalence data can be formulated in a hierarchical form as follows (Diggle, Tawn and Moyeed, 1998). For $i = 1, ..., n$, let Y_i be the number of positive outcomes out of n_i individuals tested at location x_i in a region of interest $\mathcal{D} \subset \mathbb{R}^2$, and $d(x_i) \in \mathbb{R}^p$ a vector of associated covariates. The model assumes that $Y_i \sim \text{Binomial}(n_i, p(x_i))$ where $p(x)$ is the prevalence of disease at a location x. The model further assumes that

$$\log[p(x)/\{1 - p(x)\}] = d(x)'\beta + S(x) \tag{3.1}$$

where $S(x)$ is a stationary Gaussian process with zero mean, variance σ^2 and correlation function $\rho(u) = \text{Corr}\{(S(x), S(x')\}$, where u is the distance between x and x'.

Fitting the standard model involves computationally intensive Monte Carlo methods, but software implementations are available; we use the `R` package `PrevMap` (Giorgi and Diggle, 2015). Stanton and Diggle (2013) show that provided the n_i

are at least 100 and $|p(x) - 0.5|$ is at most 0.4, reliable predictions can be obtained using the following computationally simpler non-hierarchical approximate model. Define the *empirical logit transform*,

$$Y_i^* = \log\{(Y_i + 0.5)/(n_i - Y_i + 0.5)\}$$

and assume that

$$Y_i^* = d(x_i)'\beta + S(x_i) + Z_i, \tag{3.2}$$

where the Z_i are mutually independent zero-mean Gaussian random variables with variance τ^2. Using this approximate method, predictive inferences need to be back-transformed from the logit to the prevalence scale.

In what follows, we will assume a Matérn (1960) correlation structure for $S(x)$,

$$\rho(u;\phi;\kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}\mathcal{K}_{\kappa}(u/\phi), \tag{3.3}$$

where $\phi > 0$ is a scale parameter that controls the rate at which correlation decays with increasing distance, $\mathcal{K}_{\kappa}(\cdot)$ is a modified Bessel function of order $\kappa > 0$, and $S(x)$ is m times mean-square differentiable if $\kappa > m$. In the simulation studies reported in Section 3.4, we use the computationally simpler, approximate method to compare different designs and do not include covariates. For the analyses of the Majete data reported in Section 3.5, we use the standard model Equation (3.1).

3.2.2 Likelihood-based inference under adaptive design.

Almost all geostatistical analyses are conducted under the assumption that the sampling design, $\mathcal{X}$, is stochastically independent of S. This justifies basing inference on the likelihood function corresponding to the conditional distribution of Y given $\mathcal{X}$, which typically gives information on all quantities of interest. Diggle, Menezes and Su (2010) discuss the inferential challenges that result when the independence assumption does not hold, in which case the data $(\mathcal{X}, Y)$ should strictly be considered jointly as a realisation of a marked point process. Diggle, Menezes and Su (2010) call this *preferential sampling*; see also Pati, Reich and Dunson (2011), Gelfand, Sahu and Holland (2012), Shaddick and Zidek (2014), and Zidek, Shaddick and Taylor (2014) .

In adaptive design, $\mathcal{X}$ and S are not independent but are conditionally independent given Y, which simplifies the form of the likelihood function. To see why, let $\mathcal{X}_0$ denote an initial sampling design chosen independently of S, and Y_0 the resulting measurement data. Similarly denote by $\mathcal{X}_1$ the set of additional sampling locations added as a result of analysing the initial dataset $(\mathcal{X}_0, Y_0)$, Y_1 the resulting additional measurement data, and so on. After k additions, the complete dataset consists of $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \cup ... \cup \mathcal{X}_k$ and $Y = (Y_0, Y_1, ..., Y_k)$. Using the notation $[\cdot]$ to mean "the distribution of", the associated likelihood for the complete dataset is

$$[\mathcal{X}, Y] = \int_S [\mathcal{X}, Y, S] dS. \tag{3.4}$$

We consider first the case $k = 1$. The standard factorisation of any multivariate distribution gives

$$[\mathcal{X}, Y, S] = [S, \mathcal{X}_0, Y_0, \mathcal{X}_1, Y_1] = [S][\mathcal{X}_0|S][Y_0|\mathcal{X}_0, S][\mathcal{X}_1|Y_0, \mathcal{X}_0, S][Y_1|\mathcal{X}_1, Y_0, \mathcal{X}_0, S]. \tag{3.5}$$

On the right-hand side of Equation (3.5), note that by construction, $[\mathcal{X}_0|S] = [\mathcal{X}_0]$ and $[\mathcal{X}_1|Y_0, X_0, S] = [\mathcal{X}_1|Y_0, X_0]$. It then follows from Equation (3.4) and Equation (3.5) that

$$[\mathcal{X}, Y] = [\mathcal{X}_0][\mathcal{X}_1|\mathcal{X}_0, Y_0] \times \int_S [Y_0|\mathcal{X}_0, S][Y_1|\mathcal{X}_1, Y_0, \mathcal{X}_0, S][S]dS \tag{3.6}$$

The first term of the right-hand side of Equation (3.6) is the conditional distribution of $\mathcal{X}$ given Y_0. The second term simplifies to

$$[Y_0|\mathcal{X}_0][Y_1|\mathcal{X}_1, Y_0, \mathcal{X}_0] = [Y_0, Y_1|\mathcal{X}_0, \mathcal{X}_1] = [Y|\mathcal{X}].$$

It follows that

$$[\mathcal{X}, Y] = [\mathcal{X}|Y_0] \times [Y|\mathcal{X}]. \tag{3.7}$$

Equation (3.7) shows that the conditional likelihood, $[Y|\mathcal{X}]$, can legitimately be used for inference although, depending on how $[\mathcal{X}|Y_0]$ is specified, it may be inefficient. The argument leading to Equation (3.7) extends to $k > 1$ with essentially only notational changes.

3.3 An adaptive design strategy.

3.3.1 Performance criteria.

In practice, each geostatistical prediction exercise will have its own, context-specific primary objective. We provide a framework for a general discussion here. For clarity, we repeat some basic terminology and let $\mathcal{S} = \{S(x) : x \in \mathcal{D}\}$ denote the realisation of the process $S(x)$ over $\mathcal{D}$. Also, let Y denote the data obtained from the sampling design $\mathcal{X} = \{x_1, ..., x_n\}$, and $Y = (Y_1, ..., Y_n)$ the corresponding measurement data. Denote by $T = \mathcal{T}(\mathcal{S})$, called the *predictive target*, represent the property of $\mathcal{S}$ that is of primary interest. A generic measure of the predictive accuracy of a design $\mathcal{X}$ is its mean square error, $MSE(\mathcal{X}) = \mathrm{E}[(T - \hat{T})^2]$, where $\hat{T} = \mathrm{E}[T|Y; \mathcal{X}]$ is the minimum mean square error predictor of T for any given design $\mathcal{X}$. Note that in the expression for $MSE(\mathcal{X})$ the expectation is with respect to both $\mathcal{S}$ and Y, whereas in the expression for $\hat{T}$ it is with respect to $\mathcal{S}$ holding Y fixed at its observed value.

One obvious predictive target is $S(x)$ for arbitrary location $x \in \mathcal{D}$. Another, which may be more relevant when the practical goal is to decide whether or not to launch a public health intervention, is a complete map $T(x) = I(S(x) > c)$, where $I(\cdot)$ is the indicator function and $\boldsymbol{c}$ is a policy-relevant threshold; see, for example, Figure 3 of Zouré et al. (2014). Spatially neutral versions of these targets can be defined by integration over $\mathcal{D}$, hence

$$IMSE(\mathcal{X}) = \int_{\mathcal{D}} \mathrm{E}[(T(x) - \hat{T}(x))^2] dx.$$

We emphasise that in any particular application, other measures of performance may be more appropriate. However, for a comparative evaluation of different general design strategies, we adopt $IMSE(\mathcal{X})$ as a sensible generic measure.

3.3.2 Some non-adaptive geostatistical designs.

Two standard non-adaptive designs are a *completely random* design, in which the sample locations x_i form an independent random sample from the uniform distribution on $\mathcal{D}$ and a *completely regular* design in which the x_i form a regular square or, less commonly, triangular lattice. Geostatistical design problems can be classified according to whether the primary objective is parameter estimation or spatial prediction and, in the latter case, whether model parameters are assumed known or unknown. Our focus is on design for efficient prediction when model parameters are unknown, this being the ultimate goal of most geostatistical analyses. Completely regular designs typically give efficient prediction when the target is the spatial average of $S(x)$, i.e. $T = \int_{\mathcal{D}} S(x)dx$, and model parameters are known; see, for example, Matérn (1960, Chapter 5); Bellhouse and Herzberg (1984), Fernández, Real et al. (2005), Marchant, Lark and Wheeler (2005), Müller (2007) and Diggle and Ribeiro (2007). When parameters are unknown, less regular designs have been shown to be preferable in particular settings see, for example, Diggle and Lophaven (2006), although a general theory of optimal geostatistical design is lacking.

Most of the previous research on design considerations for prediction assume a known covariance structure for the data, see, for example, Benhenni and Cambanis (1992), Ritter (1996) and Müller (2005). Su and Cambanis (1993) address the problem of estimating parameters from a random process with a finite number of observations, and measure the design performance by integrated mean square error. They show that random designs are asymptotically optimal. McBratney, Webster and Burgess (1981) address the problem of choosing the spacing of a regular rectangular or triangular lattice design to achieve an acceptable value of the maximum of the prediction variance over the region of interest. Yfantis, Flatman and Behar (1987) compare three regular sampling designs, namely the square, equilateral triangle and regular hexagonal lattices. They conclude that the hexagonal design is the best when the nugget effect is large and the sampling density is sparse.

Royle and Nychka (1998) and Nychka and Saltzman (1998) use a geometrical approach that does not depend on the covariance structure of the underlying process $S(x)$. In this approach, sample points are located in a way that minimises a criterion that is a function of the distances between sampled and non-sampled locations. Royle and Nychka (1998) show that the resulting *space-filling* designs generally perform well.

In contrast to the spatial designs for efficient prediction reviewed above, Russo (1984), Müller and Zimmerman (1999) and Bogaert and Russo (1999) consider variogram-based parameter estimation. The variogram of $S(x)$ is the function $\gamma(u) = \frac{1}{2}\text{Var}\{S(x) - S(x')\}$ where u is the distance between x and x'. Müller and Zimmerman (1999) regard a design as optimal if it minimises a suitable measure of the "size" of the covariance matrix of the resulting parameter estimates.

Typically, the same dataset is used for covariance structure estimation and prediction of $S(x)$ at unsampled locations, in which case it is desirable to use a design that compromises between these two analysis objectives. Zhu and Stein (2006) address the problem of spatial sampling design for prediction of stationary isotropic Gaussian processes with estimated parameters of the covariance structure. They employ a two-step algorithm that uses an initial set of locations $\mathcal{X}_0$ to find the best design for prediction with known covariance parameters and then, conditional on $\mathcal{X}_0$, uses the rest to find the best design for estimation of those covariance parameters. Pilz and Spöck (2006) address a similar design problem but using a model-based approach in choosing an optimal design for spatial prediction in the presence of uncertainty in the covariance structure. Using a Bayesian approach, Diggle and Lophaven (2006) consider designs that are efficient for spatial prediction when parameters are unknown. They looked at two different design scenarios, namely: *retrospective* design, using as performance criterion the average prediction variance (APV),

$$APV = \int_{\mathcal{D}} \text{Var}\{S(x)|\,Y\}\text{d}x, \tag{3.8}$$

and *prospective* design, with performance criterion the expectation of APV, with respect to the process $S(x)$. They concluded that in either situation, the inclusion of close pairs in an otherwise regular lattice design is generally a good choice.

3.3.3 A class of adaptive designs.

Our proposed approach to adaptive geostatistical design is as follows.

1. Specify the finite set, $\mathcal{X}^*$ say, of n^* potential sampling locations $x_i \in \mathcal{D}$. In our motivating application, this consists of the locations of all households in their respective villages in the Majete perimeter area. In other applications, any point $x \in \mathcal{D}$ may be a potential sampling location, in which case we take $\mathcal{X}^*$ to be a finely spaced regular lattice to cover $\mathcal{D}$.

2. Use a non-adaptive design to choose an initial set of sample locations, $\mathcal{X}_0 = \{x_i \in \mathcal{D} : i = 1, ..., n_0\}$.

3. Use the corresponding data Y_0 to estimate the parameters of an assumed geostatistical model.

4. Specify a criterion for the addition of one or more new sample locations to form an enlarged set $\mathcal{X}_0 \cup \mathcal{X}_1$. A simple example would be for $\mathcal{X}_1$ to be the elements of $\mathcal{X}^*$ with the largest values of the prediction variance amongst all points not already included in $\mathcal{X}_0$.

5. Repeat steps 3 and 4 with augmented data Y_1 at the points in $\mathcal{X}_1$.

6. Stop when the required number of points has been sampled, a required performance criterion has been achieved or no more potential sampling points are available.

Within this general approach, in addition to choosing a suitable addition criterion in step 4, we need to choose the number and locations of points in the initial design,

$\mathcal{X}_0$, and the number to be added at each subsequent stage called the *batch size.* A batch size $b = 1$ must be optimal theoretically, but is often infeasible in practice. For example, in our application to prevalence mapping in the Majete Wildlife Reserve perimeter area, the associated sampling involves field work in challenging terrain and remote villages to obtain the measurements Y. Restricting each field-trip to collection of a single measurement would be a hopelessly inefficient use of limited resources.

3.3.4 Types of adaptive designs.

We develop two main types of adaptive geostatistical designs namely: *singleton* and *batch* adaptive designs.

In *singleton adaptive sampling*, $b = 1$, i.e. locations are chosen sequentially, allowing x_{k+1} to depend on data obtained at all earlier locations $x_1, \ldots, x_k$. In singleton adaptive sampling, one possible addition criterion is to choose x_{k+1} to be the location x with the largest prediction variance of $S(x)$ given the data from $x_1, \ldots, x_k$.

In *batch adaptive sampling*, $b > 1$. A naive extension of the above addition criterion, choosing $(x_{k+1}, ..., x_{k+b})$ to be the b available locations with the largest prediction variances of $S(x)$, is likely to fail because it does not penalise sampling from multiple locations x at which the corresponding $S(x)$ are highly correlated.

3.3.5 Algorithm for adaptive geostatistical design.

For the predictive target $T = S(x)$ at a particular location x, given an initial set of sampling locations $\mathcal{X}_0 = (x_1, ..., x_{n_0})$ the available set of additional sampling locations is $A_0 = \mathcal{X}^* \setminus \mathcal{X}_0$. For each $x \in A_0$, denote by $PV(x)$ the prediction variance, $\text{Var}(T|Y_0)$. For the Gaussian model Equation (3.2),

$$PV(x) = \sigma^2(1 - r'V^{-1}r),$$

where $r = (r_1, \ldots, r_{n_0})$ with $V = R + \nu^2 I$, R is the n by n matrix with elements $r_{ij} = \rho(||x_i - x_j||)$, $\nu^2 = \tau^2/\sigma^2$ and I is the identity matrix (Diggle and Ribeiro, 2007, p136).

We propose to incorporate a *minimum distance* addition criterion, whereby we choose new locations $x_{n_0+1}, x_{n_0+2}, ..., x_{n_0+b}$ with the b largest values of $PV(x)$ subject to the constraint that no two locations are separated by a distance of less than δ.

For a formal specification, we use the following notation:

- $\mathcal{X}^*$ is the set of all potential sampling locations, with number of elements of n^*;
- $\mathcal{X}_0$ is the initial sample, with number of elements n_0;
- b is the batch size;

- $n = n_0 + kb$ is the total sample size;
- $\mathcal{X}_j, j \geq 1$, is the set of locations added in the j^{th} batch, with number of elements b;
- $A_j = \mathcal{X}^* \setminus (\mathcal{X}_0 \cup ... \cup \mathcal{X}_j)$ is the set of available locations after addition of the j^{th} batch.

The algorithm then proceeds as follows.

1. Use a non-adaptive design to determine $\mathcal{X}_0$.
2. Set j=0
3. For each $x \in A_j$, calculate $PV(x)$:

   (i) choose $x^* = \arg\max_{A_j} PV(x)$,

   (ii) if $||x^* - x_i|| > \delta$, for all $i = 1, ..., n_0 + jb$, add x^* to the design,

   (iii) otherwise, remove x^* from A_j
4. Repeat step 3 until b locations have been added to form the set $\mathcal{X}_{j+1}$.
5. Set $A_j = A_{j=1} \setminus \mathcal{X}_j$ and we update j to $j + 1$.
6. Repeat steps 3 to 5 until the total number of sampled locations is n or $A_j = \emptyset$.

3.4 Simulation studies.

We conducted simulation studies of our proposed AGD method so as to compare its performance with standard examples of non-adaptive geostatistical designs (NAGDs). Sampling in non-adaptive designs is based on *a priori* information and is fixed before the study is implemented (Thompson and Collins, 2002). Two examples of NAGD are: *random* and *inhibitory* design. Inhibitory designs use a constrained form of simple random sampling (Diggle, 2013) whereby the distance between any two sampled locations is required to be at least δ. In this way, we retain the objective of a randomised design whilst guaranteeing a relatively even spatial coverage of the study region.

In each case, data were generated as a realisation of Gaussian process $S(x)$ on a 64 by 64 grid covering the unit square, giving a total of $n^* = 4096$ potential sampling locations. We specified $S(x)$ to have expectation $\mu = 0$, variance $\sigma^2 = 1$ and Matérn correlation function (Equation (3.3)), with $\phi = 0.05$ and $\kappa = 1.5$, and no measurement error, i.e. $\tau^2 = 0$. In each run of the simulation, we used the adaptive design algorithm outlined in Section 3.3.5 to sample a total of $n = 100$ locations. We varied the initial sample size n_0 between 30 and 90 and considered batch sizes $b = 1$ (singleton adaptive sampling), 5 and 10.

3.4.1 Adaptive vs non-adaptive sampling.

For the non-adaptive sampling of each realisation, and for the initial sample in adaptive sampling, we used an inhibitory design with $\delta = 0.03$. We evaluated

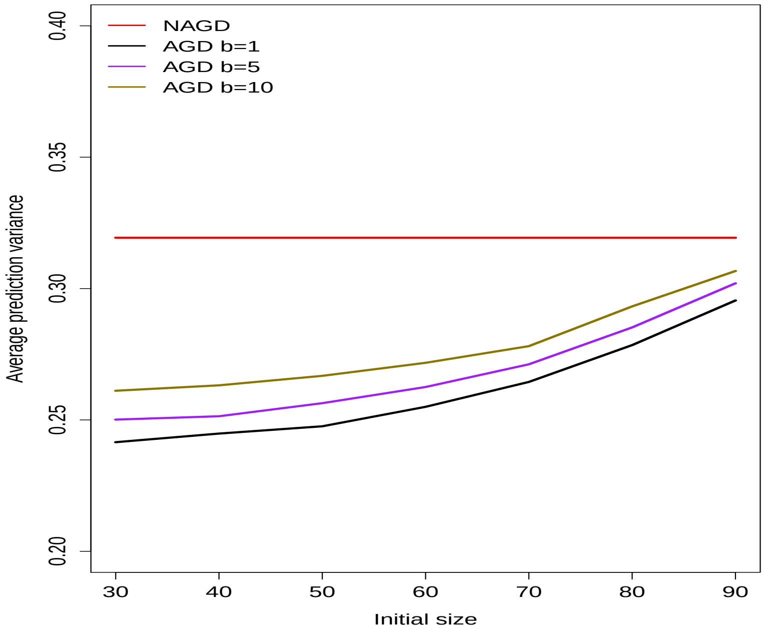


Figure 3.1: Non-adaptive (NAGD) vs **minimum distance** batch adaptive (AGD) sampling, with $\delta = 0.03$ and AGD batch sizes $b = 1, 5$ and 10; Initial size (n_0) ranges from 30 to 90. See text for details of the simulation model.

each design by its spatially averaged prediction variance, i.e. APV as defined at Equation (3.8), in turn, averaged over 100 replicate simulations. When the initial sample size is $n_0 = 30$, Figure 3.1 shows singleton adaptive sampling to have the lowest APV, achieving a value APV = 0.24. As the size of the batch increases, APV also increases but remains substantially lower than the value APV=0.33 achieved by non-adaptive sampling.

As the initial size n_0 increases towards $n = 100$, the APV for any of the AGDs necessarily approaches that of the NAGD. For example, Figure 3.1 shows the value of APV ≈ 0.30 when $n_0 = 90$ and $b = 10$. For $b = 1$ and 5, APV generally remains low whilst steadily approaching that of NAGD when n_0 increases towards n.

3.5 Application: rolling malaria indicator surveys for malaria prevalence in the Majete perimeter.

In this Section, we illustrate the use of our proposed sampling methodology to construct a malaria prevalence map for part of an area of the community surrounding Majete Wildlife Reserve within Chikwawa district (16° 1′ S; 34 ° 47′ E), in the lower Shire valley, southern Malawi. The Shire river (the biggest river in Malawi) runs throughout the length of Chikwawa district, causing perennial flooding in the rainy season. Chikwawa is situated in a tropical climate zone with a mean annual temperature of 26 °C, a single rainy season from November to April and an annual rainfall of approximately 770 mm. The district has extensive rice and sugar-cane irrigation schemes.

The area surrounding Majete Wildlife Reserve forms the region for a five-year monitoring and evaluation study of malaria prevalence, with an embedded randomised trial of community-level interventions intended to reduce malaria transmission. The whole Majete perimeter is home to a population of ≈ 100,000. Within this population, three distinct administrative units known as focal areas A, B and C have been selected to form the study region. These are spread over 61 villages with ≈ 6,600 households and a population of ≈ 24,500. Here, we illustrate adaptive sampling design methodology using data from focal area B, see Figure 3.2. Note that the sampling unit in the Majete study is the household.

The first stage in the geostatistical design was a complete enumeration of households in the entire study region, including their geo-location collected using Global

Positioning System (GPS) devices on a Samsung Galaxy Tab 3 running Android 4.1 Jellybean operating system. These devices are accurate to within 5 meters. In the on-going rMIS, approximately 90 households are sampled per month per focal area, so that each household will be visited twice over the two years of the study. Malaria prevalence is highly seasonal. The adaptive design problem therefore consists of deciding which households to sample in each of the first 12 months so as to optimise the precision of the resulting sequence of 12 prevalence maps. In year 2, the sampling design will be re-visited to take account of both statistical considerations and any practical obstacles encountered during the first year. Here, to illustrate the methodology, we use data from the first wave of sampling.

Ethical approval for the study was obtained from Malawi's College of Medicine Research Ethics Committee (COMREC) and Liverpool School of Tropical Medicine Research Ethics Committee (LSTM-REC). The informed consent process involves two stages. The first stage is group-consent, whereby a group of potential participants, for example, the inhabitants of a single village, receive an information sheet and are given the opportunity to ask any questions that they may have regarding the objectives and procedures of the study. In the second stage individual informed consent is obtained from each participant or (if they are aged < 15) from one of their parents or a legal guardian. Two copies of a consent form are completed; one is kept confidentially and securely by the study team and the second is kept by the participant.

3.5.1 Data.

An initial malaria indicator survey was conducted over the period April to June 2015. The survey recruited children aged less than 5 years and women of child-bearing age, 15 to 49 years, in 10 village communities in order to monitor the burden of malaria. An inhibitory sampling design was used to sample an initial 100 households per focal area. Selection of the households was as follows. Households were randomly selected within each village from a list of enumerated households, whilst ensuring a good spatial coverage of the focal area by insisting that the distance between any two sampled households is not less than 0.1 kilometres. Figure 3.2 shows the sampled household locations (white dots) in their respective villages, with black dots indicating all households in each village. Data collected include the outcome of malaria rapid diagnostic test, age, gender of each individual and socio-economic status of each household.

For predictive mapping, any covariates included in the model must be available at all prediction locations. We, therefore, used two digital elevation model (DEM) derivatives, elevation and normalized difference vegetation index (NDVI), which are readily available throughout the study region. Data for these covariates were derived using the Advanced Space-borne Thermal Emission and Reflection Radiometer (ASTER) Global DEM version 2. ASTER GDEM V2 has a spatial resolution of 30 meters. The data were downloaded from the United States Geological Survey (USGS) through their 'Global Data Explorer' `http://gdex.cr.usgs.gov/gdex/`.

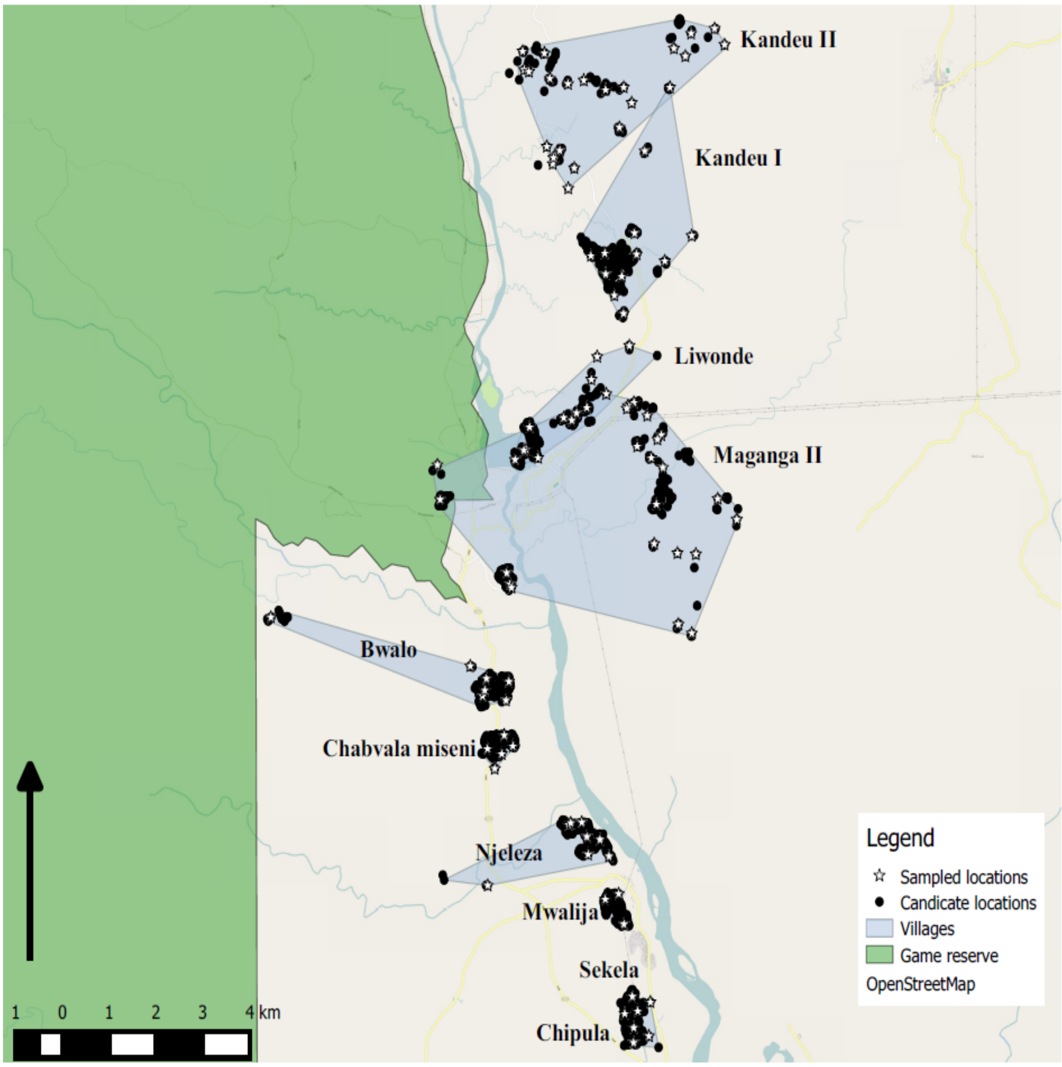


Figure 3.2: Households within the Majete Wildlife Reserve perimeter in focal area B (black dots) and sampled household locations (white dots) shown in their respective villages.

3.5.2 Results.

We emphasise that at this early stage of the Majete study the data are too sparse for a definitive prevalence analysis but are sufficient to illustrate the practical implementation of our proposed AGD method. The response from each individual in a sampled household is the binary outcome of a rapid diagnostic test (RDT) for the presence/absence of malaria from a finger-prick blood sample. Out of the 100 households in the initial sample, 72 had at least one individual who met the

Table 3.1: Monte Carlo maximum likelihood estimates and 95 % confidence intervals for the model fitted to the Majete malaria data.

Term	Estimate	95 % Confidence Interval
Intercept	-5.4827	(-7.6760, -3.2893)
Elevation	0.02651	(0.0162, 0.0368)
NDVI	4.6130	(0.1581, 9.0680)
Elev. × NDVI	-0.0405	(-0.0588, -0.0223)
σ^2	0.6339	(0.4438, 0.9055)
ϕ^*	0.2293	(0.1042, 0.5049)

**Distance is given in kilometres.*

inclusion criteria (see Section 3.5.1 above). The total number of eligible individuals in these 72 households was 126, with household size ranging from 1 to 8 individuals. For covariate selection we used ordinary logistic regression, retaining covariates with nominal p-values less than 0.05. This resulted in the set of covariates shown in Table 3.1, with terms for elevation, NDVI and the interaction between the two. We then fitted the binomial logistic model (Equation (3.1)) to obtain the Monte Carlo maximum likelihood estimates of the parameters and associated 95 % confidence intervals, as also shown in Table 3.1. Each evaluation of the log-likelihood used 10,000 simulated values, obtained by conditional simulation of 110,000 values and sampling every 10^{th} realization after discarding a burn-in of 10,000 values.

From Table 3.1, elevation and NDVI show positive marginal associations with malaria, with a negative interaction. Focal area B is divided through its length by the Shire river. The north-east part has relatively high elevation and NDVI values. Prevalence is generally low in the south-west of the region, whereas the north-east has pockets of comparatively high malaria prevalence. This suggests that heterogeneity in malaria prevalence over focal area B involves other risk factors (social or environmental) that are not available in the current data.

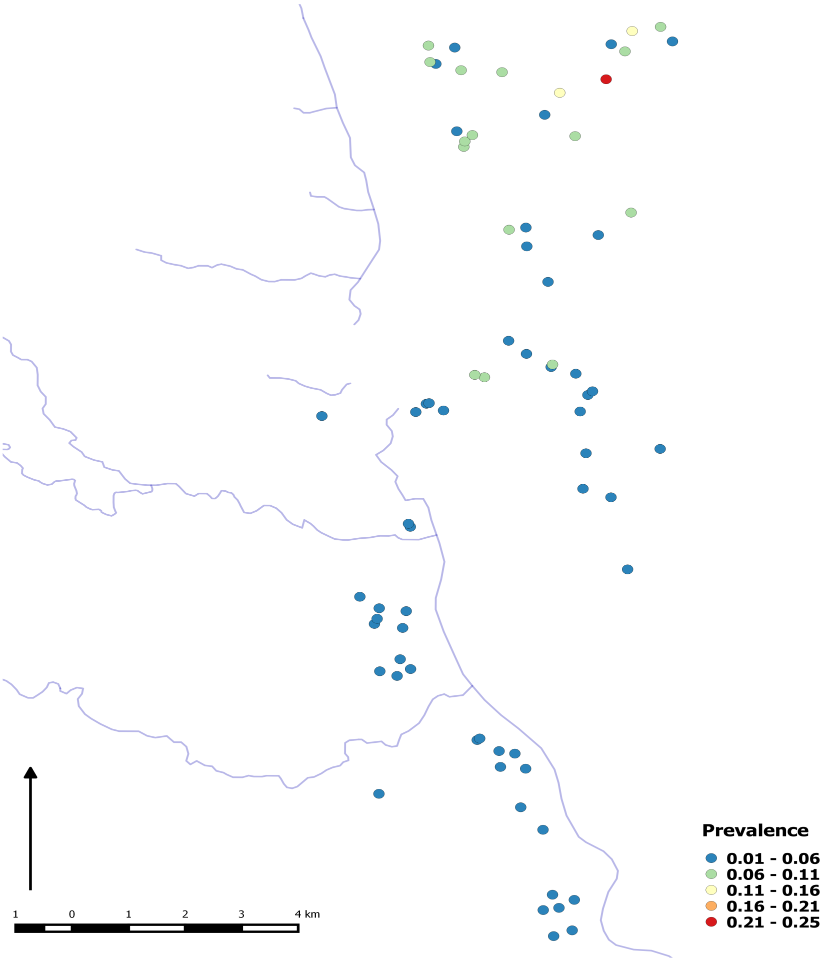


Figure 3.3: Predictions of $d(x)'\beta + S(x)$ at observed locations in focal area B. The blue lines show Shire and Matope rivers.

Figure 3.3 shows the predicted prevalence at each of the observed locations. Households at high altitude and under dense vegetation cover have generally high malaria prevalence. For this study, the elevation of households varied from 60 to 460 meters above sea level. Rivers and streams that are fast flowing in nature are not generally favourable for mosquito larvae; the Shire river is a big and fast flowing river. Sampling was done at the time of peak malaria transmission at the end of the rainy season. This could potentially explain the low prevalence in the southern part of the study region. Also, the high prevalence area in the north-east is generally more remote with poorer access to health facilities.

3.5.3 Adaptive sampling in practice.

We now use the *minimum distance batch adaptive sampling* approach explained in Section 3.3.5 to determine new locations that can and should be added to the existing sample in an adaptive manner. We first calculate the prediction variance at each household using the data from the 72 initial sample locations, shown as red dots in Figure 3.4. Prediction variances range between 0.0003 and 0.06, and are relatively small at locations closer to the observed locations, although this depends on the number of eligible individuals at each location. We then choose a sample of 90 additional locations using random sampling as well as the algorithm outlined in Section 3.3.5 above for comparison sake. The black dots in Figure 3.6 show 90 new locations determined using random sampling. The blue dots in Figure 3.8 show 90 new locations determined using the minimum distance threshold $\delta = 0.15$ kilometres. The new sampling locations are well spread across the study region, which is beneficial for area-wide spatial prediction. Also, although we have imposed a minimum distance of 0.15 kilometres between any two sampled locations in order to penalise highly correlated multiple sample locations, the new sample locations nevertheless include some pairs of old and new locations in which the new location has been chosen to be relatively close to an initial location with high prediction variance; recall that the number of eligible individuals per household varied between 1 and 8, hence the prediction variance at a sampled location is itself highly variable. Also, as noted earlier, closely spaced pairs are helpful for effective spatial prediction when the true model parameters are not known, which is the case in most geostatistical problems.

In Figure 3.5 we show the prediction variance surface for the inset sub-region in Figure 3.4. In Figure 3.7 and Figure 3.9 we show the same information after addition of 90 randomly and adaptively selected locations in Figure 3.6 and Figure 3.8, respectively. The adaptive sampling design criterion ensures that data are collected only from locations that will deliver useful additional information in order to understand the spatial heterogeneity throughout the study region. A comparison of the two prediction variance surfaces after addition of the 90 locations shows the extent to which the adaptive design out-performs non-adaptive random sampling.

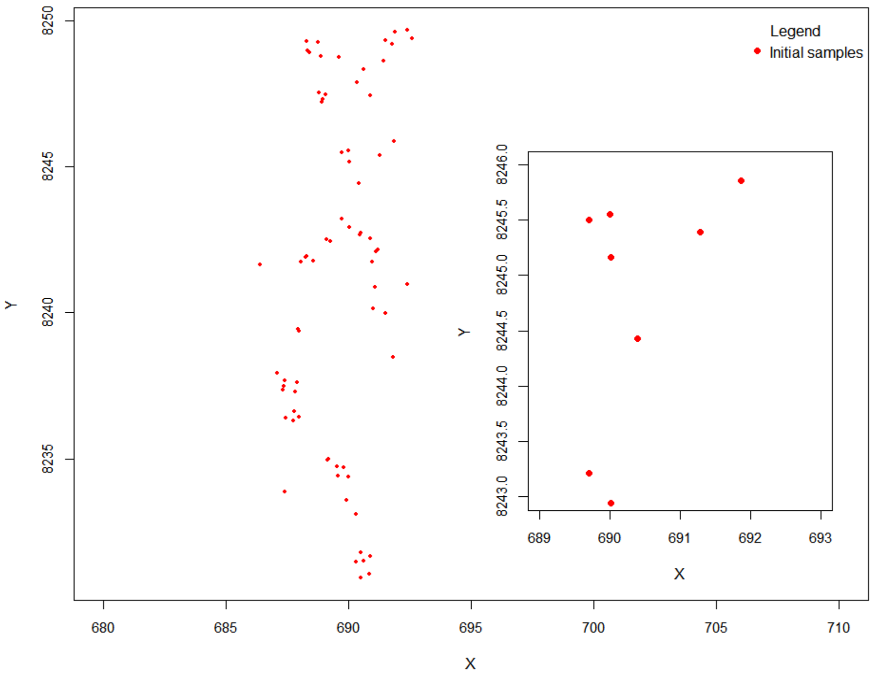


Figure 3.4: Initial inhibitory sampling design locations (red dots) in focal area B. The inset shows a subset of locations.

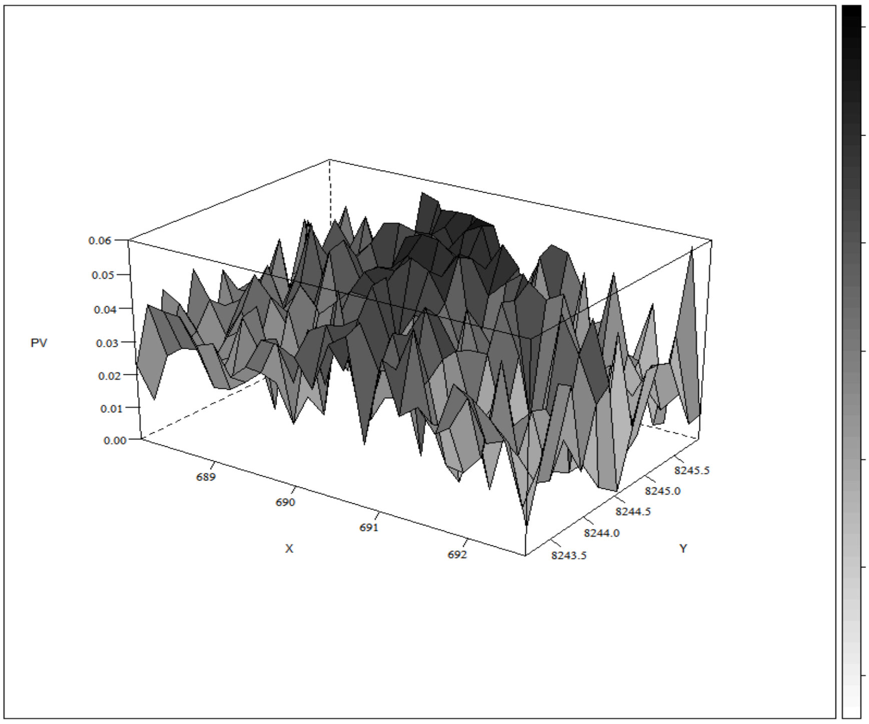


Figure 3.5: Prediction variance surface for the inset sub-region from Figure 3.4.

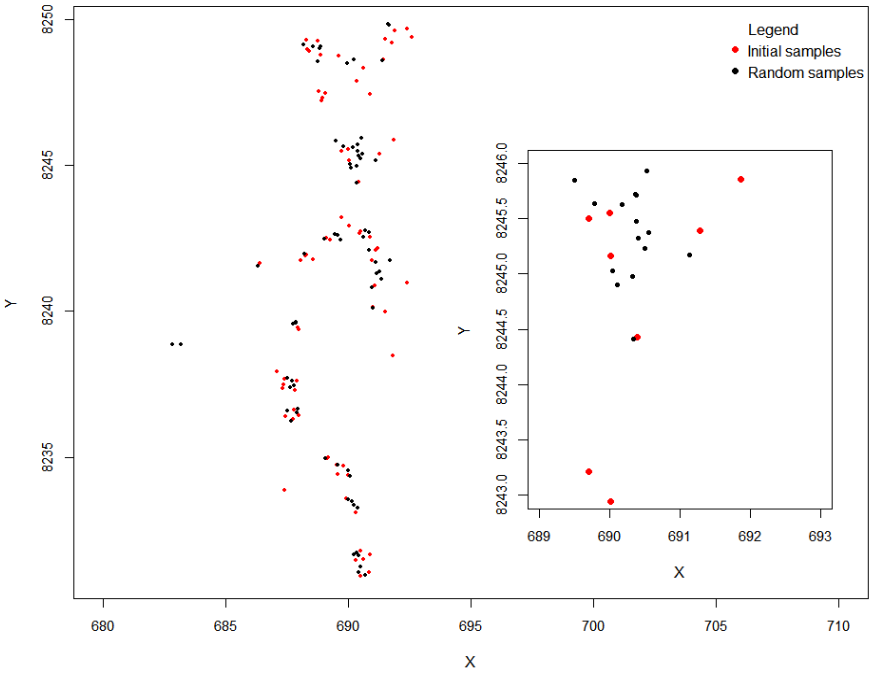


Figure 3.6: Initial inhibitory sampling design locations (red dots) and random sampling design locations (black dots) in focal area B. The inset shows a subset of locations.

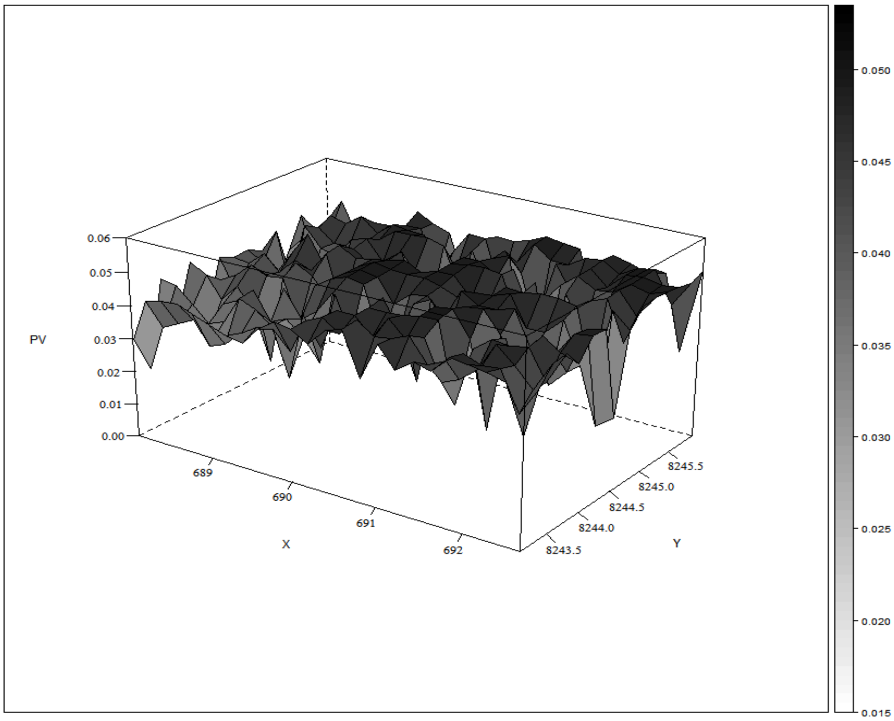


Figure 3.7: Prediction variance surface for the inset sub-region from Figure 3.6.

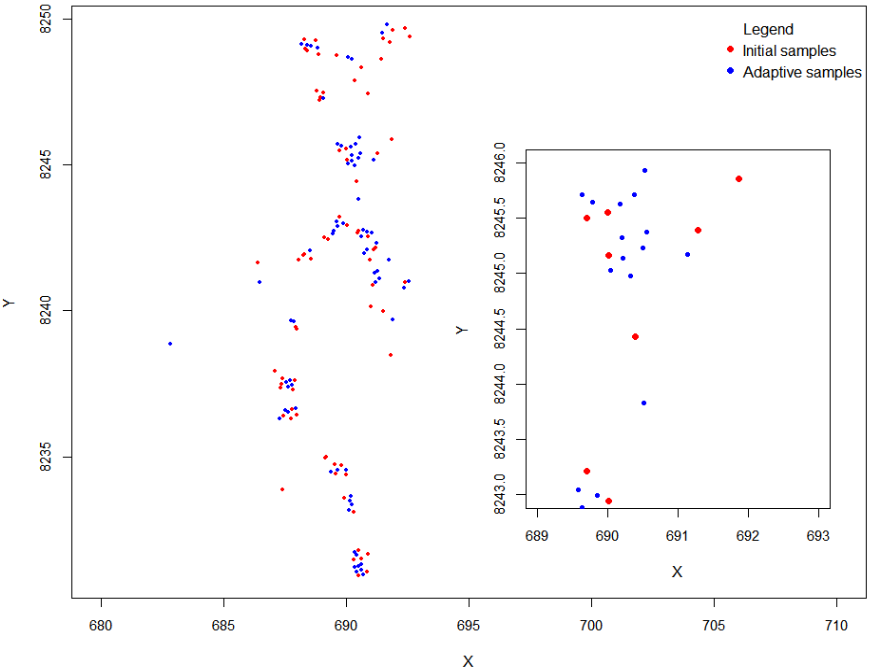


Figure 3.8: Initial inhibitory sampling design locations (red dots) and adaptive sampling design locations (blue dots) in focal area B. The Inset shows a subset of locations.

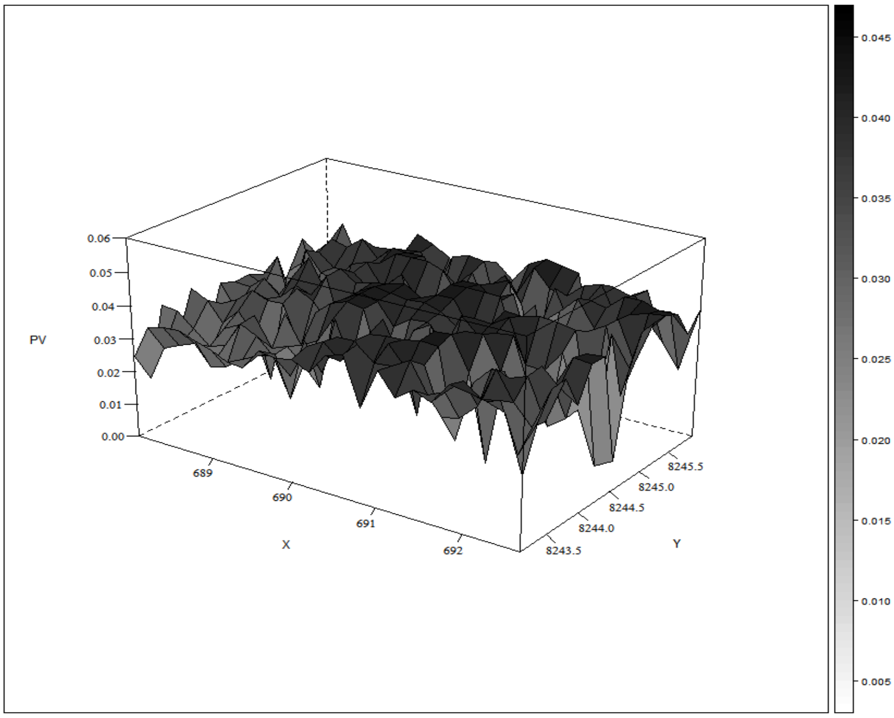


Figure 3.9: Prediction variance surface for the inset sub-region from Figure 3.8.

3.6 Discussion.

In any particular application, the objectives of the study can and should inform the design strategy. We have developed an adaptive design strategy within a model-based geostatistics (MBG) framework for survey-based disease mapping in poor resource settings. The particular design strategy described in Section 3.3.5 is intended to deliver efficient mapping of the complete surface, $S(x)$, over the region of interest. The same principles, but with a context-specific performance criterion replacing the point-wise prediction variance, can be used in other settings. For example, if the aim is to detect and subsequently evaluate sub-regions that appear to meet a policy-determined intervention threshold so as to use scarce resources to best effect, accurate prediction in low-prevalence sub-regions is relatively unimportant and an adaptive design should result in the progressive concentration of sampling into areas of relatively high prevalence.

In our application to malaria prevalence mapping, we used an initial set of rMIS data to map disease prevalence in focal area B and analysed the resulting data to define a follow-up sample of new locations with the aim of reducing as much as possible the average prediction variance. We used a large batch size, $b = 90$ because of the high cost in staff and travel time of re-visiting the study region more often than monthly. Smaller batch sizes, if feasible, would potentially lead to greater gains in efficiency. The optimum choice of the minimum distance δ between sampled locations should relate to the scale of the spatial correlation, i.e. the parameter ϕ in the Matérn model (Equation (3.3)), as its purpose is to prevent redundant duplication of highly correlated data points. The exact nature of this relationship appears to be intractable although, in principle, simulations could be used to find a near-optimum value of δ for any assumed spatial correlation structure.

Our use of average prediction variance as a spatially neutral optimisation criterion in the Majete application reflects our lack of prior knowledge about the spatial variation in prevalence. It is possible that in the later stages of this five-year study, the optimisation criterion will be changed, for example to more precisely delineate areas of persistent high risk.

A fundamental feature of our approach is that we distinguish between a measured value Y_i at a location x_i and the corresponding value $S(x_i)$ of an underlying spatial process $S(x)$ which is the focus of scientific interest. The difference between Y_i and $S(x_i)$ is considered to be measurement error. In some versions of geostatistical analysis, this difference is interpreted as short-range spatial variation and would therefore be considered to form part of the predictive target. Note, however, that with prevalence data of the kind considered in the Majete application, each measurement necessarily includes binomial sampling variation.

The adaptive sampling design approach is of potentially wide application to disease mapping in low-resource settings, where accurate registry data typically do not exist. Mapping exercises are an important component of any control or elimination programme. Collecting data adaptively allows for local identification and targeting of areas with high transmission, incidence or prevalence, and an understanding of which household-level and community-level factors influence these properties. Knowledge of these properties can inform area-wide health policy making and identify locations of greatest need where interventions that would be considered too costly or complicated to implement across an entire population can be targeted in order to optimise their public health impact.

The choice of the initial sampling design $\mathcal{X}_0$ is an important step for adaptive sampling. The initial sample size, n_0, needs to be large enough to allow the fitting of a geostatistical model, whose estimate parameter values then drive the adaptive sampling. In the Majete application, we prescribed $n_0 = 100$ but, in the event, found eligible study participants in 72 of the sampled households. We recommend re-estimation of the model parameters after each batch of locations has been added.

In the Majete application, the irregular spatial distribution of households across the study region meant that the initial set of 72 sampled locations achieved a good compromise between even coverage of the study region and the inclusion of close pairs, which is generally helpful for efficient parameter estimation. In other contexts, and specifically where there is essentially no restriction of the placement of sampling locations, it would be better to use an initial design that forces the inclusion of some close pairs, as recommended in Diggle and Lophaven (2006).

As with classical survey sampling, in applications where there is good prior knowledge of large-scale heterogeneity pre-stratification of the study region into subregions can bring substantial gains in efficiency (Wang, Haining and Cao, 2010; Hu and Wang, 2011; Gao et al., 2015). In such cases, further benefits can be obtained by using adaptive designs within each stratum. However, a detailed discussion of stratified designs is beyond the scope of the present paper.

In conclusion, the proposed adaptive sampling design approach provides a systematic approach to the collection of exposure and outcome data over time so as to optimise progress towards achievement of the analysis objective. Adaptive designs are particularly well-suited to spatial mapping studies in low-resource settings where uniformly precise mapping may be unrealistically costly and the priority is often to identify critical areas where interventions can have the greatest health impact. Development of adaptive geostatistical design methodology is, therefore, timely for monitoring and evaluating interventions in tropical diseases with high burden such as malaria, in areas where accurate disease registries do not exist and resources are severely limited. Malaria, in particular, is a leading cause of death in most of sub-Saharan Africa, especially among children under 5 years of age. Malaria monitoring and control programmes can benefit from the availability of accurate prevalence maps. Geostatistical analysis in conjunction with adaptive sampling is an effective, practical strategy for producing accurate local-scale maps that can pick up short-term changes in disease burden and that are complementary to the national-scale maps that have been produced, for example, by Hay, Guerra, Tatem et al. (2004), Guerra et al. (2007), Hay, Guerra, Gething et al. (2009) and Gething et al. (2012).

Acknowledgements.

We thank the participants and Majete integrated malaria control project staff involved in the ongoing data collection of the presented household prevalence surveys, part of which is here.

Funding.

Michael Chipeta is supported by ESRC-NWDTC Ph.D. studentship (grant number ES/J500094/1). Dr. Dianne Terlouw, Prof. Kamija Phiri and Prof. Peter Diggle are supported by the Majete integrated malaria control project grant funded by the Dioraphte foundation.

References

Bellhouse, D.R. and Herzberg, A.M. (1984) 'Equally spaced design points in polynomial regression: a comparison of systematic sampling methods with the optimal design of experiments.', *The Canadian Journal of Statistics* 12 (2), pp. 77–90.

Benhenni, K. and Cambanis, S. (1992) 'Sampling designs for estimating integrals of stochastic processes', *Annals of Statistics* 20, pp. 161–194.

Bogaert, P. and Russo, D. (1999) 'Optimal spatial sampling design for the estimation of the variogram based on a least squares approach', *Water Resources Research* 35 (4), pp. 1275–1289.

Chilès, J.-P. and Delfiner, P. (2012) *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed., New Jersey: John Wiley & Sons, Inc.

Diggle, P.J. (2013) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns.* 3rd ed., Boca Raton: CRC Press.

Diggle, P.J. and Lophaven, S. (2006) 'Bayesian geostatistical design', *Scandinavian Journal of Statistics* 33 (1), pp. 53–64.

Diggle, P.J., Menezes, R. and Su, T.-L. (2010) 'Geostatistical inference under preferential sampling', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59 (2), pp. 191–232.

Diggle, P.J. and Ribeiro, J.P. (2007) *Model-based Geostatistics*, New York: Springer.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998) 'Model-based geostatistics (with discussion)', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47 (3), pp. 299–350.

Fernández, J.A., Real, C., Couto, J.A., Aboal, J.R. and Carballeira, A. (2005) 'The effect of sampling design on extensive bryomonitoring surveys of air pollution', *Science of the Total Environment* 337 (1-3), pp. 11–21.

Gao, B.-B., Wang, J.-F., Fan, H.-M., Xu, K., Hu, M.-G. and Chen, Z.-Y. (2015) 'A stratified optimization method for a multivariate marine environmental monitoring network in the Yangtze River estuary and its adjacent sea', *International Journal of Geographical Information Science* 29 (8), pp. 1332–1349.

Gelfand, A.E., Sahu, S.K. and Holland, D.M. (2012) 'On the Effect of Preferential Sampling in Spatial Prediction.', *Environmetrics* 23 (7), pp. 565–578.

Gething, P.W., Elyazar, I.R.F., Moyes, C.L., Smith, D.L., Battle, K.E., Guerra, C.A., Patil, A.P., Tatem, A.J., Howes, R.E., Myers, M.F., George, D.B., Horby, P., Wertheim, H.F.L., Price, R.N., Mueller, I., Baird, J.K. and Hay, S.I. (2012) 'A long neglected world malaria map: Plasmodium vivax endemicity in 2010', *PLoS Neglected Tropical Diseases* 6 (9), e1814.

Giorgi, E. and Diggle, P.J. (2015) ‘PrevMap : an R Package for Prevalence Mapping’, *Journal of Statistical Software (to appear)*, pp. 1–27.

Guerra, C.A., Hay, S., Lucioparedes, L.S., Gikandi, P.W., Tatem, A.J., Noor, A.M. and Snow, R.W. (2007) ‘Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project’, *Malaria Journal* 6, p. 17.

Hay, S., Guerra, C.A., Gething, P.W., Patil, A.P., Tatem A.J., Noor, A.M., Kabaria, C.W., Manh, B.H., Elyazar, I.R.F., Brooker, S., Smith, D.L., Moyeed, R.A. and Snow, R.W. (2009) ‘A world malaria map: Plasmodium falciparum endemicity in 2007’, *PLoS Medicine* 6 (3), e1000048.

Hay, S., Guerra, C.A., Tatem, A.J., Noor, A.M. and Snow, R.W. (2004) ‘The global distribution and population at risk of malaria: past, present, and future’, *The Lancet Infectious Diseases* 4 (6), pp. 327–336.

Hu, M.G. and Wang, J.F. (2011) ‘A spatial sampling optimization package using MSN theory’, *Environmental Modelling and Software* 26 (4), pp. 546–548.

Krige, D.G. (1951) ‘A Statistical Approach to Some Mine Valuation and Allied Problems on the Witwatersrand.’, *Journal of the Chemical, Metallurgical and Mining Society of South Africa* 52, pp. 119–139.

Marchant, B.P., Lark, R.M. and Wheeler, H.C. (2005) *Developing methods to improve sampling efficiency for automated soil mapping*, tech. rep., Home-Grown Cereals Authority.

Matérn, B. (1960) *Spatial Variation*, tech. rep., Stockholm: Statens Skogsforsningsinstitut.

McBratney, A.B., Webster, R. and Burgess, T.M. (1981) 'The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalized Variables–I: Theory and method.', *Computers & Geosciences* 7 (4), pp. 331–334.

Müller, W.G. (2005) 'A comparison of spatial design methods for correlated observations', *Environmetrics* 16, pp. 495–505.

Müller, W.G. (2007) *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*, 3rd ed., Berlin: Springer-Verlag.

Müller, W.G. and Zimmerman, D.L. (1999) 'Optimal designs for Variogram estimation', *Enviromentrics* 10, pp. 23–37.

Nychka, D. and Saltzman, N. (1998) 'Design of Air-Quality Monitoring Networks', *Case Studies in Environmental Statistics SE - 4*, Lecture Notes in Statistics 132, ed. by D. Nychka, W. Piegorsch and L. Cox, pp. 51–76.

Pati, D., Reich, B.J. and Dunson, D.B. (2011) 'Bayesian geostatistical modelling with informative sampling locations', *Biometrika* 98, pp. 35–48.

Pilz, J. and Spöck, G. (2006) 'Spatial sampling design for prediction taking account of uncertain covariance structure', *7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.* Lisbon, Portugal, pp. 109–118.

Ritter, K. (1996) 'Asymptotic optimality of regular sequence designs', *The Annals of Statistics* 24 (5), pp. 2081–2096.

Roca-Feltrer, A., Lalloo, D.G., Phiri, K. and Terlouw, D.J. (2012) 'Short Report : Rolling Malaria Indicator Surveys (rMIS): a potential district-level malaria monitoring and evaluation (M & E) tool for program managers.', *American Journal of Tropical Medicine and Hygiene* 86 (1), pp. 96–98.

Royle, J. and Nychka, D. (1998) 'An algorithm for the construction of spatial coverage designs with implementation in SPLUS', *Computers & Geosciences* 24 (5), pp. 479–488.

Russo, D. (1984) 'Design of an Optimal Sampling Network for Estimating the Variogram', *Soil Science Society of America Journal* 48 (4), pp. 708–716.

Shaddick, G. and Zidek, J.V. (2014) 'A case study in preferential sampling: Long term monitoring of air pollution in the UK', *Spatial Statistics* 9, pp. 51–65.

Stanton, M.C. and Diggle, P.J. (2013) 'Geostatistical analysis of binomial data: generalised linear or transformed Gaussian modelling?', *Environmetrics* 24 (3), pp. 158–171.

Su, Y.S.Y. and Cambanis, S. (1993) 'Sampling Designs for Estimation of a Random Process', *Stochastic Processes and their Applications* 46, pp. 47–89.

Thompson, S.K. and Collins, L.M. (2002) 'Adaptive sampling in research on risk-related behaviors.', *Drug and Alcohol Dependence* 68, S57–S67.

Wang, J., Haining, R. and Cao, Z. (2010) 'Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning', *International Journal of Geographical Information Science* 24 (4), pp. 523–543.

Yfantis, E.A., Flatman, G.T. and Behar, J.V. (1987) 'Efficiency of Kriging Estimation for Square , Triangular , and Hexagonal Grids', *Mathematical Geology* 19 (3), pp. 183–205.

Zhu, Z. and Stein, M.L. (2006) ‘Spatial sampling design for prediction with estimated parameters’, *Journal of Agricultural, Biological, and Environmental Statistics* 11 (1), pp. 24–44.

Zidek, J.V., Shaddick, G. and Taylor, C.G. (2014) ‘Reducing estimation bias in adaptively changing monitoring networks with preferential site selection’, *The Annals of Applied Statistics* 8 (3), pp. 1640–1670.

Zouré, H.G.M., Noma, M., Tekle, A.H., Amazigo, U.V., Diggle, P.J., Giorgi, E. and Remme, J.H.F. (2014) ‘The geographic distribution of onchocerciasis in the 20 participating countries of the African Programme for Onchocerciasis Control: (2) pre-control endemicity levels and estimated number infected’, *Parasites & Vectors* 7, p. 326.

Chapter 4

Paper 3. Invited Discussion of "Optimal Design in Geostatistics under Preferential Sampling," by Ferreira and Gamerman.

Chipeta, Michael[a]; Diggle, Peter J.[a,b] Comment on Article by Ferreira and Gamerman. ***Bayesian Analysis***. 10 (2015), no. 3, 737–739. doi:10.1214/15-BA944A. http://projecteuclid.org/euclid.ba/1432213203.

[a] Institute of Infection and Global Health, University of Liverpool, UK.

[b] Medical School, Lancaster University, UK.

4.1 Introduction.

This paper is a welcome addition to the growing literature on preferential sampling in a geostatistical setting. Earlier papers cited by the authors have shown that preferential sampling materially affects parameter estimation and prediction. The authors now demonstrate that the same applies to design, or more specifically to the optimal augmentation of an initial set of geostatistical data that has been sampled preferentially. Almost in passing, the paper also sets out an algorithm for Bayesian inference under preferential sampling that is a useful contribution in its own right. Might we look forward to an `R` package implementation of this?

Our comments fall into two categories: theoretical remarks on what we call *adaptive design*, including an explanation of why this does not necessarily require consideration of preferential sampling issues; practical constraints that may limit the scope for theoretically optimal designs to be used in practice, especially in low-resource settings.

4.2 Adaptive geostatistical design and preferential sampling.

The topic of geostatistical design is multi-faceted. One useful distinction is between adaptive and non-adaptive designs. A *non-adaptive* design is one that is completely determined before any data are collected. An *adaptive* design is one in which an initial design is augmented in a way that depends on the analysis of interim data.

We make two theoretical comments that follow from the definition of preferential sampling given in Diggle, Menezes and Su (2010).

Firstly, an adaptive design need not be preferential. To see why, it is sufficient to consider a two-stage adaptive design, $X = (X_0, X_1)$ with associated measurement data (Y_0, Y_1), where subscripts 0 and 1 identify initial and follow-up stages, respectively. Similarly, write $S = (S_0, S_1)$ for the corresponding decomposition of the latent process S. Quite generally, we can factorise the joint distribution of (X, Y, S) as

$$[X, Y, S] = [S, X_0, Y_0, X_1, Y_1] = [S][X_0|S][Y_0|X_0, S][X_1|Y_0, X_0, S][Y_1|X_1, Y_0, X_0, S]. \tag{4.1}$$

On the right-hand side of Equation (4.1), if the initial design is non-preferential, $[X_0|S] = [X_0]$, whilst by construction $[X_1|Y_0, X_0, S] = [X_1|Y_0, X_0]$. It then follows that

$$\begin{aligned} [X, Y] &= [X_0][X_1|X_0, Y_0] \times \int_S [Y_0|X_0, S][Y_1|X_1, Y_0, X_0, S][S]dS \\ &= [X|Y_0] \times [Y|X] \end{aligned} \tag{4.2}$$

and the log-likelihood is a sum of two components, $\log[X|Y_0] + \log[Y|X]$. This shows that the conditional likelihood, $[Y|X]$, can legitimately be used for inference although, depending on how $[X|Y_0]$ is specified, to do so may be inefficient. The argument leading to Equation (4.2) is closely related to the proof that if data are "missing at random" the missingness mechanism can be ignored when using likelihood-based inference (Rubin, 1976), and extends to multi-stage adaptive designs with essentially only notational changes.

Secondly, shared dependence of a design X and the latent process S on observed covariates does not necessarily render X preferential. Specifically, if Z denotes the covariate process, then $[X, S|Z] = [S|Z][X|S, Z]$. The requirement for the design to be non-preferential is that $[X|S, Z] = [X|Z]$, which in general is a weaker requirement than $[X|S] = [X]$. This illustrates, not for the first time, that spatial statistical inference can be greatly simplified by judicious selection of spatially referenced covariates.

4.3 Some practical constraints on geostatistical design.

The paper makes a number of explicit and implicit assumptions that together provide a very reasonable framework for theoretical analysis, but it is worth bearing in mind that in any particular application, the design problem may be constrained in various ways. These include the following.

1. *Is the spatial integral of the predictive variance an appropriate measure of predictive performance*

This would not be the case if, for example, $S(x)$ represents pollution and the main objective is to monitor compliance with environmental standards; see Fanshawe and Diggle (2012).

2. *Sampling may not be equally costly at every location*

   Put another way, should the design be constrained by the number of locations to be sampled, or by the total sampling effort in the field? An obvious example of this is when travel-time represents a non-negligible proportion of field-effort; see, for example, Figures 2 and 4 of Diggle, Thomson et al. (2007), where the sampled points follow the routes of field-trips, leading to a highly aggregated pattern that is far from optimal from a purely theoretical perspective.

3. *Large-scale spatial heterogeneity*

   The latent process S may exhibit different patterns of small-scale and large-scale spatial variation, in which case it may be desirable to compromise between designs that are locally and globally optimal. A pragmatic strategy might then be to pre-stratify the study region into relatively homogeneous sub-regions and apply optimal design theory separately to each sub-region.

4. *The number of potential sampling points may be finite*

   This applies to disease prevalence surveys when the sampling unit is either a household or a well-defined community. We are currently working on the adaptive design of an ongoing malaria prevalence mapping project around the perimeter of the Majete national park, Malawi, where the first task has been

> to enumerate and geo-locate each household in each village within the study region. In the course of the project, we expect to sample all households, but the order in which they are sampled (in a sequence of monthly field-trips) will be chosen adaptively with the aim of optimising the estimation of the complete spatio-temporal variation in malaria prevalence, which is known to include a strong seasonal component.

None of these comments are intended to detract from the value of the paper on its own terms. Theoretical studies of this kind help to further our understanding of important, and often subtle, methodological issues around modelling and inference for preferentially sampled geostatistical data.

References

Diggle, P.J., Thomson, M.C., Christensen, O.F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J.H., Boussinesq, M. and Molyneux, D.H. (2007) ‘Spatial modelling and the prediction of Loa loa risk: decision making under uncertainty’, *Annals of Tropical Medicine and Parasitology* 101 (6), pp. 499–509.

Diggle, P.J., Menezes, R. and Su, T.-L. (2010) ‘Geostatistical inference under preferential sampling’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59 (2), pp. 191–232.

Fanshawe, T.R. and Diggle, P.J. (2012) ‘Adaptive Sampling Design for Spatio-Temporal Prediction’, *Spatio-temporal design: Advances in efficient data acquisition*, ed. by J. Mateu and W.G. Müller, 1st ed., Chichester, UK: John Wiley & Sons, Ltd, chap. 11, pp. 249–268.

Rubin, D.B. (1976) ‘Inference and Missing Data’, *Biometrika* 63 (3), pp. 581–592.

Chapter 5

Paper 4. Adaptive geostatistical sampling enables efficient identification of malaria hotspots in rural Chikwawa, Malawi.

Kabaghe, A. N.[a,b], **Chipeta, M. G.**[b,c,d], McCann, R. S.[b,f], Phiri, K. S.[b], van Vugt, M.[a], Takken, W.[f], Diggle, P. J.[c], and Terlouw, D. J. [b,d,e] (2016). Adaptive geostatistical sampling enables efficient identification of malaria hotspots in rural Chikwawa, Malawi. ***PLoS ONE***.

[a]Academic Medical Centre, University of Amsterdam, Netherlands.

[b]College of Medicine, University of Malawi, Blantyre, Malawi.

[c]Lancaster Medical School, Lancaster University, Lancaster, UK.

[d]Malawi-Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi.

[e]Liverpool School of Tropical Medicine, Liverpool, UK.

[f]Wageningen University and Research, Netherlands.

Abstract

Background: In the context of malaria elimination, interventions will need to target high-burden areas to further reduce transmission. Current tools to monitor and report disease burden lack the capacity to continuously detect fine-scale spatial and temporal variations of disease distribution exhibited by malaria. These tools use random sampling techniques that are inefficient for capturing underlying heterogeneity while health facility data in resource-limited settings are inaccurate. Continuous community surveys of malaria burden provide real-time results of local spatio-temporal variation. Adaptive sampling design improves prediction of the outcome of interest compared to current random sampling technique. We present findings of continuous malaria prevalence surveys using an adaptive sampling design.

Methods: We conducted repeated cross-sectional surveys guided by an adaptive sampling design to monitor the prevalence of malaria parasitaemia and anaemia in children below five years old in the communities living around Majete Wildlife Reserve in Chikwawa district, southern Malawi. We fitted a geostatistical model to predict malaria prevalence in the area.

Findings: We conducted five rounds of sampling, and tested 876 children aged 6–59 months from 1,377 households over a 12-month period. Malaria prevalence prediction maps showed spatial heterogeneity and presence of hotspots; predictors of malaria include age, socio-economic status and ownership of insecticide-treated mosquito nets.

Interpretation: Continuous malaria prevalence surveys using adaptive sampling increased malaria prevalence prediction accuracy. Results from the surveys were readily available after data collection. The tool can assist local managers to target malaria control interventions in areas with the greatest health impact and is ready for assessment in other diseases.

Funding: The study was funded by Dioraphte Foundation, The Netherlands.

5.1 Background.

In the context of malaria elimination, limited resources and significant decline in malaria incidence and prevalence (Bhatt et al., 2015; World Health Organisation, 2015b), interventions will need to target high disease burden areas to further reduce transmission (Bousema, Griffin et al., 2012; Alemu, Worku and Berhane, 2013; Walker et al., 2016). Current tools for monitoring or reporting malaria burden lack the capacity to detect high malaria transmission areas, often called *"hotspots"*, and to report continuous changes of disease burden over time (World Health Organisation, 2015a). Malaria exhibits spatial and temporal heterogeneity in both stable and endemic transmission settings (Alemu, Worku and Berhane, 2013; Baidjoe et al., 2016). National malaria control programmes (NMCP) rely on national surveys such as Malaria Indicator Surveys (MIS) and Demographic and Health Surveys (DHS) or use health facility malaria case reports and/or registers to monitor malaria burden and the progress of malaria control. The surveys generally are cross-sectional, use random population samples, do not report real-time results and are repeated after long periods of time (at least two years). They lack spatial and temporal heterogeneity information for malaria prevalence and only produce data at a national and regional level rather than sub-district level. In resource-limited settings, facility case registers, where available, provide unreliable data (Snow et al., 1999; Chilundo, Sundby and Aanestad, 2004), under-represent the burden of disease in the community, are incomplete, prone to errors, and may misreport the number of cases due to lack of diagnostic capacity (Chilundo, Sundby and Aanestad, 2004; Amexo et al., 2004; Rowe et al., 2009; Afrane et al., 2013).

Continuous disease surveys allow continuous monitoring of changes in spatial and temporal disease distribution at national, regional and district levels; the surveys are potential tools to accurately monitor disease control progress in low-resource settings where surveillance systems are weak (Rowe, 2009). Continuous malaria prevalence surveys allow continuous analysis of data, mapping of malaria prevalence, and reporting short-term changes in disease prevalence and intervention coverage (Giorgi, Sesay et al., 2015). Monthly cross-sectional prevalence surveys report results within a short duration (Roca-Feltrer et al., 2012). Use of such surveys would assist district managers to identify high disease transmission "hotspot" areas for early targeted intervention (Bousema, Griffin et al., 2012).

Recent developments in geostatistical modelling offer opportunities to implement more accurate predictive methods for disease burden (Reid et al., 2010; Patil et al., 2011). Geostatistical modelling can be used to map disease risk and visualise spatial and temporal changes of disease burden and intervention coverage. The random sampling of clusters used currently in surveys lacks the accuracy for detecting fine-scale heterogeneity of disease burden. These sampling methods may under-represent heterogeneously distributed and hard to reach populations in resource-limited settings (Kondo et al., 2014). An adaptive geostatistical design (AGD) would allow gain in statistical sampling efficiency by focusing on areas where prediction of the measure of interest is imprecise. Chipeta et al. (2016a) previously demonstrated AGD on simulated data and reported potential for improved prediction of malaria prevalence compared to non-adaptive (random) sampling. Adaptive designs allow sampling to focus on sub-regions where precise prediction is needed to inform public health action.

We describe the first field application of adaptive sampling design in continuous malaria prevalence surveys for a 12 month period, and we present malaria prevalence maps from the study site in Chikwawa district, Malawi.

5.2 Methods.

5.2.1 Study setting.

We conducted the study in villages surrounding Majete Wildlife Reserve (MWR) in Chikwawa district, southern Malawi from April 2015 to April 2016. Malaria transmission is intense and peaks from December to March during the rainy season (Mzilahowa et al., 2012). The study area is within the catchment area of the Majete Malaria Project (MMP), a five-year, community-based malaria control project. We conducted the surveys in 61 villages with approximately 6,600 households and a total population of approximately 25,000. The area was divided into three administrative units, which, for convenience purposes are referred to as *focal areas*: A, B and C; see Figure 5.1 from which villages and households within villages were sampled.

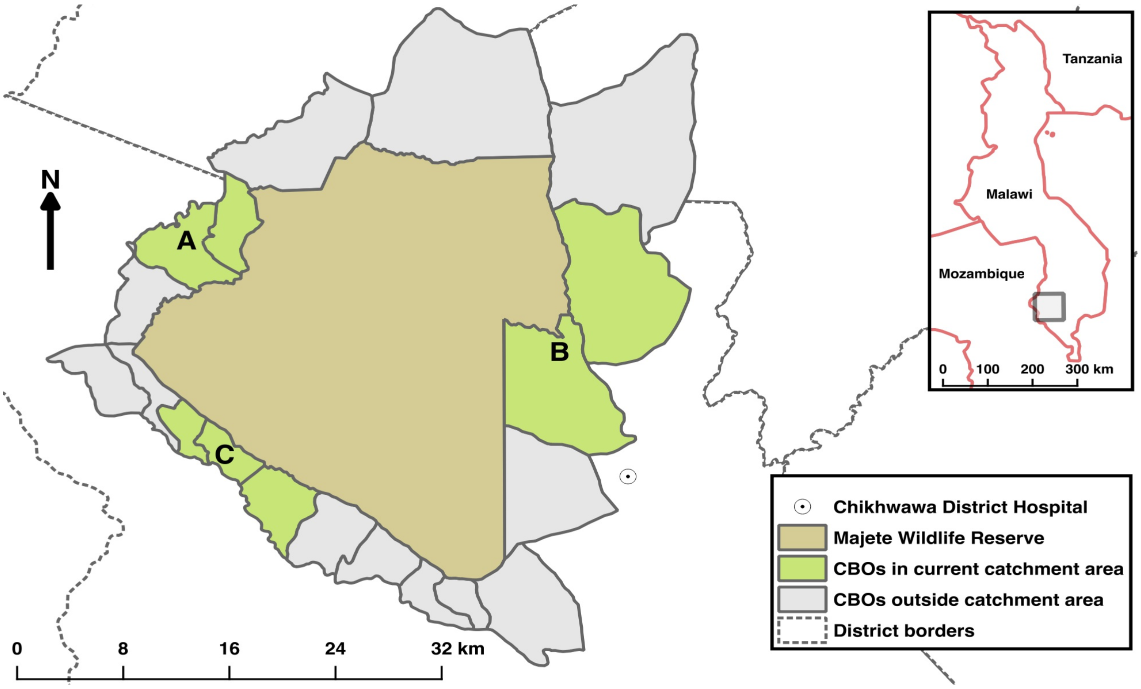


Figure 5.1: Map showing Majete Wildlife Reserve (brown) and borders of the 19 community-based organisations (CBOs) comprising the Majete perimeter. Three focal areas (green), labelled as A, B, and C, mark the communities selected for malaria indicator surveys. The rest of the CBOs (grey) are outside the project's catchment area.

5.2.2 Study design.

The sampling unit in the study was the household. We used an adaptive repeated cross-sectional Malaria Indicator Survey (rMIS) design (Roca-Feltrer et al., 2012; Chipeta et al., 2016a) to collect data. In this design, on any sampling occasion, the choice of sampling households was informed by prevalence results from an analysis of the data collected on earlier occasions and a different set of households was chosen on each occasion. The adaptive design problem consisted of deciding which households to sample in each round of sampling to optimise the precision of the resulting sequence of area-wide prevalence maps.

The first stage in the geostatistical design of the study was a complete enumeration of households in the study region from August to November 2014; Geo-location data were collected using Global Positioning System (GPS) devices on Samsung Galaxy Tab 3 running Android 4.1 Jellybean Operating System, accurate to within 5 meters on Open Data Kit (ODK) platform. We used the enumeration data to sample the first 100 households in each of the three focal areas using a spatially inhibitory random sampling design (Chipeta et al., 2016b), to achieve approximately uniform coverage of each of the focal areas in the study-area. The second round of sampling also followed a spatially inhibitory sample. At the end of these two initial and each subsequent sampling period, a standard operating procedure was followed in checking data for consistency and completeness before uploading them to an off-site database server. The accumulating data up to that period were analysed immediately and the prevalence prediction results fed into an adaptive sampling algorithm to inform the choice of new sampling locations in the next sampling round. We sampled 90 households per two months per focal area in each of the subsequent sampling rounds. Figure 5.2 shows a map of focal area B with an inset to demonstrate adaptive sampling in practice. Adaptive geostatistical designs are explained in more details in Chipeta et al. (2016a).

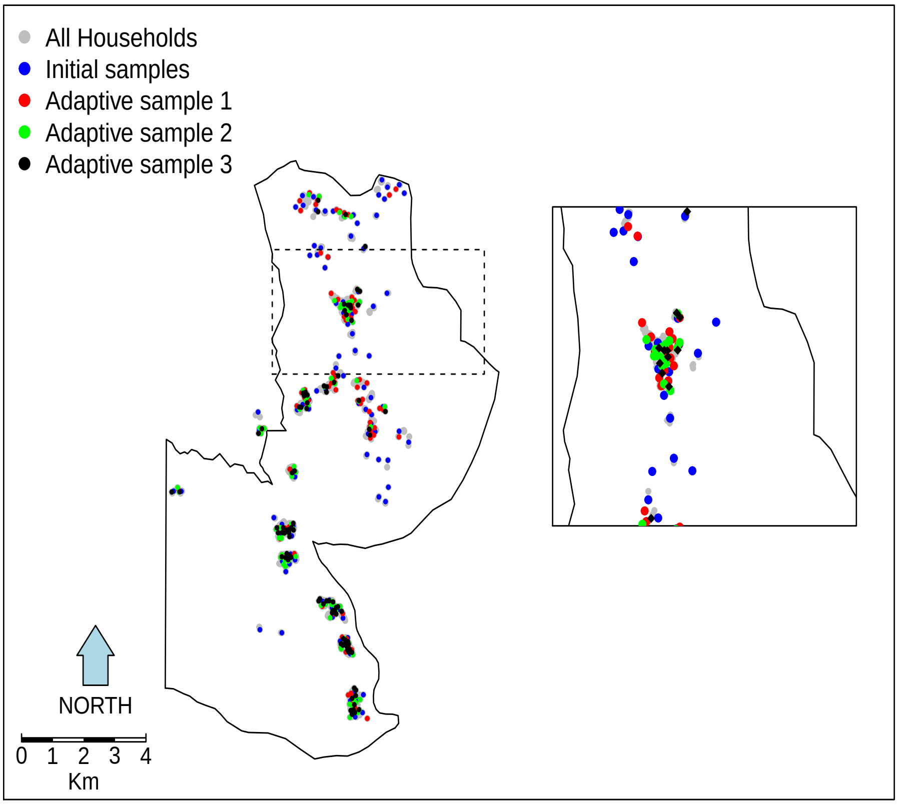


Figure 5.2: Adaptive sampling in practice, initial spatially inhibitory design samples augmented with adaptive design samples. The inset shows a zoomed-in subset of locations.

5.2.3 Participants.

We invited children 6–59 months and women 15–49 years old who slept in the sampled household the previous night to participate. If the head of household consented we interviewed, tested for malaria and anaemia and recorded temperature, weight, height and mid-upper arm circumference (MUAC) measurements. Note that MUAC measurements were done for children only. Households that did not have any eligible participants were only interviewed; no clinical assessment or blood tests were done.

5.2.4 Procedures.

We developed electronic forms and training material adapted from the global malaria indicator survey tool-kit (Roll Back Malaria Partnership, 2016). Research teams comprising a research nurse and 2 to 4 research assistants invited sampled household members to central locations where consent was obtained from the head of household. Teams followed up household members who did not present to the central location to conduct the survey. Sampled households that were unoccupied, or had been demolished, were replaced by the nearest household. The teams administered a questionnaire and tested eligible participants for malaria and anaemia using a rapid diagnostic test (RDT; SD Bioline Ag P.f (HRP)) and haemocue 301 (Haemocue, Angelholm, Sweden), respectively. Participants with RDT positive results or low haemoglobin reading (less than 11g/dl) were managed according to Malawi national treatment guidelines or referred to a health facility, respectively.

5.2.5 Statistical analysis.

The primary outcome from each individual was a binary indicator for a positive or negative malaria test by malaria RDT in children aged 6–59 months. Age of each individual, availability of at least one ITN and socio-economic status (SES) were considered, as defined in Table 5.2. For SES, an indicator of household wealth taking discrete values from 1 (poor) to 5 (wealthy), was derived by an application of principal component analysis as discussed in Vyas and Kumaranayake (2006). Data for elevation were derived using the Advanced Space-borne Thermal Emission and Reflection Radiometer Global Digital Elevation Model (ASTER GDEM) version 2, which has a spatial resolution of 30 meters. The data were downloaded from the United States Geological Survey (USGS, http://gdex.cr.usgs.gov/gdex/). Normalised difference vegetation index (NDVI) data were calculated based on images from the Landsat 8 satellite, also downloaded from the USGS (http://earthexplorer.usgs.gov/). For NDVI measure, we calculated and used mean values for the above sampling period.

We used the geostatistical binary probit model for binary response data in the following manner. Let i and j denote the indices of the i^{th} household and j^{th} individual within that household. The response variable Y_{ij} is a binary indicator taking value 1 if the individual has been tested positive for malaria and 0 otherwise. Conditionally on a zero-mean stationary Gaussian process $S(x_i)$, Y_{ij} are mutually independent Bernoulli variables with probit link function $\Phi^{-1}(\cdot)$,

$$Y_{ij}|d_{ij}, S(x_i) \overset{ind}{\sim} Bernoulli(p_{ij})$$

$$\Phi^{-1}(p_{ij}) = d'_{ij}\beta + S(x_i), \tag{5.1}$$

where d_{ij} is a vector of covariates, both at individual- and household- level, with associated regression coefficients. For details, see Beron and Vijverberg (2004), Rue and Held (2005) and Berrett and Calder (2012). The Gaussian process $S(x)$ has isotropic Matérn covariance function (Matérn, 1986) with variance σ^2, scale parameter ϕ and shape parameter κ.

The target for predictive inference is $T = \mathcal{T}(\mathcal{S})$, i.e. malaria prevalence prediction for unobserved locations in the study region. Additionally, we delineate sub-regions of the study region where prevalence $p(x)$ is likely to exceed a policy intervention/national threshold, exceedance probability, in which case the target becomes $T = \{x : p(x) > \boldsymbol{c}\}$ for pre-specified $\boldsymbol{c}$. All analyses were done in `R` statistical environment (R Core Team, 2015).

5.2.6 Ethical consideration.

Ethical clearance for the study was obtained from the College of Medicine research ethics committee (COMREC) in Malawi (P.09/14/1631). Permissions were obtained from the Ministry of Health (MoH) and the district health authorities in Chikwawa district. Prior to the start of the study, a series of meetings were held in participating communities to explain the nature and purpose of the study. We obtained individual informed consent and in the case of children, from their parents or legal guardians.

5.2.7 The role of the funding source.

Dioraphte Foundation, the funder of the study, had no role in study design, data collection, data analysis, data interpretation or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

5.3 Results.

We conducted five sampling rounds within 12 months and completed data-collection from 1,377 (87.8 %) of the 1,568 sampled households (Table 5.1). Consent was refused from 41 (2.6 %) households. Data-collection was not completed in a further 149 (9.5 %) households, mainly because the house was vacated between the initial enumeration and the time of household sampling. From the total sampled households, 1,044 (67.5 %) had either children 6–59 months, women 15–49 years or both eligible children and women. A total of 876 children aged 6–59 months were tested for malaria and anaemia; we excluded results from women of child-bearing age in the analysis as malaria prevalence surveys are based on children. It took an average of 4–8 weeks to complete data collection per sampling round; results of each sampling round were available within 1–2 weeks after completion of data collection and cleaning.

For covariate selection we used the ordinary probit regression model, retaining covariates with nominal p-values less than 0.05; these ignore the effects of spatial

Table 5.1: Characteristics of sampled households within the Majete Wildlife Reserve perimeter.

	N	%
Total sampled households	1,568	-
Households completed	1,377	87.8
Refused consent	41	2.6
Children 6–59 months in sampled households	1,016	-
Children 6–59 months enrolled	876	86.2*
Household wealth quintile		
Lowest	390	28.3
Second	196	14.2
Middle	258	18.7
Fourth	267	19.4
Top	266	19.3

*Percentage of eligible children from sampled households who actually took part in the survey.

correlation and are likely to be anti-conservative, thereby avoiding false exclusion of potentially important covariates. This resulted in the set of covariates shown in Table 5.2, with terms for social economic status (SES), availability of at least one ITN, NDVI and elevation. The σ^2 and ϕ are variance of the Gaussian process and scale of the spatial correlation respectively. We then fitted the geostatistical binary probit model (Equation (5.1)) to obtain the Bayesian estimates of the parameters and associated 95 % highest posterior density (HPD), as also shown in Table 5.2. Each evaluation of the Markov chain Monte Carlo used 2,000 simulated values, obtained by conditional simulation of 21,000 values and sampling every 10^{th} realisation after discarding a burn-in of 1,000 values.

From Table 5.2, an increase in SES, age and ownership of at least one ITN are all associated with a reduction in the probability of a positive RDT. Elevation was negatively associated with the probability of a positive RDT whereas NDVI shows

Table 5.2: Bayesian estimates and 95 % highest posterior density intervals for the geostatistical binary probit model fitted to the Majete malaria data for children 6–59 months.

Term	Estimate	95 % HPD[1]
Intercept	0.6647	(0.1538, 1.0850)
SES[2]	-0.0737	(-0.1087, -0.0337)
ITN[3]	-0.1829	(-0.3166, -0.0337)
Age	-0.4921	(-0.6045, -0.3903)
Elevation	-0.0009	(-0.0015, -0.0004)
NDVI[4]	0.0524	(-1.1358, 0.9811)
σ^2	0.4693	(0.2154, 0.8109)
ϕ^*	2.3869	(0.7629, 4.9778)

**Distance is given in kilometres.*

[1] HPD = Highest Posterior Density; [2] SES = Social Economic Status; [3] ITN = Insecticide-Treated Net (availability of at least one in household); [4] NDVI = Normalised Difference Vegetation Index.

a positive, but non-significant, association.

Here, we present maps of malaria prevalence in children aged 6–59 months in focal area B. Prevalence maps for focal areas A and C are provided in the supplementary material Section 5.6. Overall, prevalence is higher in focal area B compared to focal areas A and C; however, Figure 5.3 (left panel) shows that prevalence is generally low in the south-west of the region, whereas the north-east has pockets of comparatively high malaria prevalence. Hotspots in focal areas A and C are mainly localised. Figure 5.3 (right panel) shows the map of exceedance probabilities that prevalence is over the national threshold of 30 %. Figure 5.4 shows the contributions of the linear regression to the predicted log-odds of prevalence at each of the observed locations in focal area B.

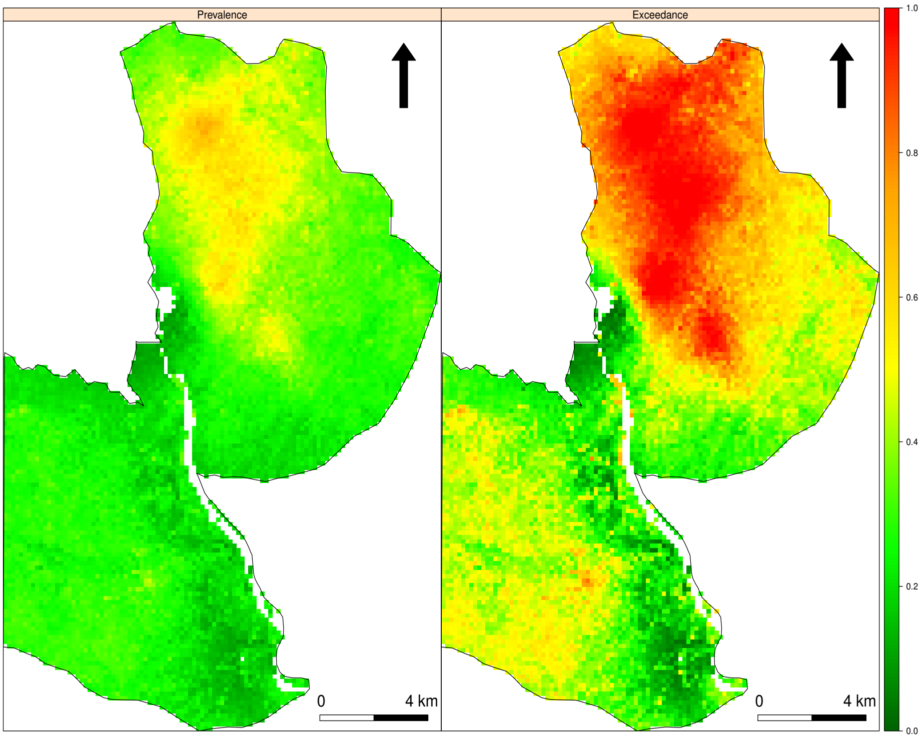


Figure 5.3: Malaria prevalence in children 6–59 months in focal area B (left panel). The right-hand panel shows the map of exceedance probabilities $P(x; 0.3)$ for the Bayesian prediction.

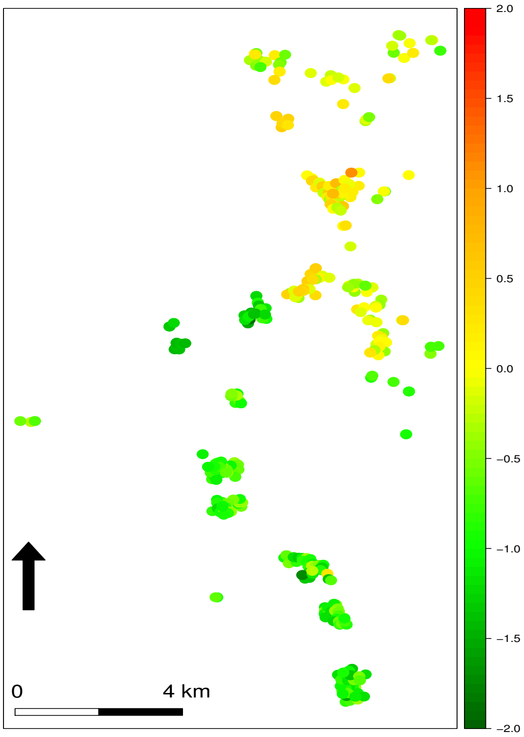


Figure 5.4: Contributions of the linear regression and of the unexplained spatial variation to the predicted log-odds of malaria prevalence in children 6–59 months at each of the observed locations in focal area B.

5.4 Discussion.

We have modelled malaria prevalence in children aged 6–59 months in a rural area of southern Malawi using individual, household and environmental data as covariates, and allowing for spatial correlation. Adaptive sampling prior to each round of data collection was used to identify areas where increased sampling effort should be focused to maximise the increase in overall predictive accuracy. Malaria prevalence predictions at observed locations show disease burden at the finest scale possible, and we detected multiple malaria hotspots across the study regions. To our knowledge, this is the first time an adaptive sampling technique has been implemented to monitor spatial distribution of malaria or any disease in a human population.

Other studies map disease prevalence heterogeneity using national and community surveys (conducted at different time points), expert opinion, facility data or a combination of these data sources (Kazembe, Kleinschmidt and Sharp, 2006; Kazembe, Kleinschmidt, Holtz et al., 2006; Burton et al., 2011; Gosoniu et al., 2012; Noor et al., 2014). With an adaptive sampling technique, we avoided reporting results based on multiple data sources which differ in accuracies, collection times and sampled areas. Health facility disease registers in resource-limited settings contain low quality, incomplete and unreliable data (Chilundo, Sundby and Aanestad, 2004; Rowe et al., 2009; Afrane et al., 2013). These data are inadequate to monitor fine changes in spatial and temporal malaria prevalence variations. Using continuous surveys based on AGD readily provides results of representative cross-sectional surveys soon after data collection. The continuous surveys monitor short-term spatial and temporal changes of disease burden to enable managers to detect and target areas requiring scaling up of interventions. The uptake and impact of malaria control interventions can also be monitored.

Compared to the recommended national MIS, the continuous prevalence surveys using AGD are not as logistically demanding. The surveys can potentially be conducted by district personnel throughout a prolonged period to complement the 2-yearly MIS. The actual data collection required small teams and took a short period of time to complete. Cost-effectiveness of implementing continuous surveys using AGD will be assessed and discussed in a separate paper, though a previous study in the same geographic area reported continuous malaria surveys using random sampling was affordable and logistically simple compared to national MIS (Roca-Feltrer et al., 2012).

The current recommended 2-yearly national MIS are cross-sectional surveys using a two-stage sample design based on geographical clusters known as enumeration areas. The sampling process is: 1a) random probability sampling of clusters, 1b) household enumeration of sampled clusters, 2) then random probability sampling of households in the clusters. Cluster sampling under-represents disease burden for heterogeneously distributed diseases and hard to reach populations (Kondo et al., 2014). The national MIS reports univariate malaria prevalence at district or regional level and without a confidence interval. Comparing disease prevalence between surveys would be inaccurate as sampled points are different and the proportions are crude (unadjusted without confidence intervals). Furthermore, the national MIS reports data from a single time point though malaria prevalence exhibits spatial and temporal variations.

By combining AGD and continuous malaria prevalence surveys, we maximise the precision of malaria prevalence predictions at the local level. Adaptive samples add value to continuous prevalence surveys. Rather than continuously selecting random samples, subsequent samples depend on previous prevalence results calculated from contributions of individual, household and environmental predictors; this allows for models to be refined as data becomes available. The subsequent samples focus on areas of relatively high uncertainty to enable more precise delineation of areas where disease prevalence is above or below a given threshold $\boldsymbol{c}$; for example, predictive probabilities of the exceedance of policy-relevant or national thresholds. AGDs also provide a more complete picture of spatial variations (Chipeta et al., 2016a). This approach can potentially empower both local and national programme managers to invest limited resources and efforts on high priority areas for elimination (Bousema, Griffin et al., 2012; Roca-Feltrer et al., 2012; Alemu, Worku and Berhane, 2013; Walker et al., 2016).

We demonstrate the first application of adaptive sampling for continuous spatial diseases surveillance in this small study population. This approach can potentially monitor temporal disease variations and will need to be implemented at a larger scale for this assessment. For large scale implementation, technical personnel are required to manage data collection, analysis and continuous sampling.

Our innovative approach for the discovery of malaria hotspots can be further fine-tuned by estimates of *Plasmodium* transmission intensities through monitoring of mosquito populations. The combined result is instrumental for effective application of malaria interventions (Bousema, Griffin et al., 2012).

Algorithms are being developed and will be available as an `R` package on the comprehensive `R` archive network (CRAN) website. The modules can be developed for real-time monitoring of disease prevalence. For example, the Meningitis Environmental Risk Information Technologies (MERIT) initiative developed such a module for meningitis epidemics prediction (MERIT Initiative, 2012; Stanton, Agier et al., 2014).

AGD enables more efficient estimation of spatial variation than traditional simple random sampling strategies (Chipeta et al., 2016a), whilst retaining the objectivity of probability-based sampling. In AGD the initial sample is a probability sample (Chipeta et al., 2016a), albeit one that is restricted to induce a degree of spatial regularity into sampled locations, and therefore achieves its increase in efficiency without risk of introducing subjective bias.

The repeated cross-sectional AGD methods are generally versatile and may apply to diseases with similar heterogeneity patterns (Schur et al., 2011; Grimes and Templeton, 2016). For example, high disease burden for neglected tropical diseases (NTDs) areas such as onchocerciasis, schistosomiasis etc. can be identified and targeted for interventions such as mass drug administration (MDA).

5.5 Conclusion.

AGD are automated algorithms that help in sampling optimisation decisions for prevalence surveys. Applying AGD to continuous disease surveys provides fine-scale disease prevalence prediction in resource-limited settings and can be a reliable surveillance tool for both district and national level programme managers. AGD results were readily available during the survey and identified several hotspots in each of the focal areas. This disease monitoring approach is ready to be assessed on a larger scale and for other diseases.

Declaration of interest.

All authors declare no conflicting interests.

Acknowledgement.

The Majete Malaria Project (MMP) was generously supported by Dioraphte Foundation, The Netherlands. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders. We thank Chikwawa District Health Office, the participants and Majete integrated malaria control project staff involved in the ongoing data collection of the presented household prevalence surveys, part of which is here.

5.6 Supplementary material.

5.6.1 Generalised linear modelling: non-spatial probit regression model.

Probit regression modelling assuming no spatial dependence showed socio-economic status, availability of at least 1 insecticide-treated bed net in the household, child's age, elevation and normalised difference vegetation index to be significant explanatory variables for malaria prevalence in children 6–59 months in Majete Wildlife Reserve perimeter. This model (Equation (5.2)) is considered as "non-spatial probit model for malaria prevalence":

$$\pi_i = \Phi(X_i^T \beta) \tag{5.2}$$

where π_i is the probability of a positive malaria RDT, Φ is the cumulative distribution function, the $X's$ are the linear predictors and β are coefficients. Socio-economic status, bed nets, age and elevation are all negatively associated with malaria prevalence whereas NDVI is positively associated with malaria prevalence, see Table 5.3.

Supplementary Table 5.3: Parameter estimates from non-spatial probit model for malaria prevalence in children 6–59 months in Majete Wildlife Reserve perimeter.

Term	Estimate	Std Error	Z value	P-value
Intercept	0.19176	0.22873	0.84	0.40181
SES	-0.09046	0.02203	-4.11	<0.001
ITN	-0.25797	0.06732	-3.83	<0.001
Age	-0.45688	0.06183	-7.39	<0.001
Elevation	-0.00123	0.00018	-6.71	<0.001
NDVI	1.58551	0.51214	3.10	0.002

5.6.2 Geostatistical modelling of malaria prevalence: geostatistical binary probit model.

Prevalence mapping and spatial predictions were obtained based on the geostatistical binary probit model for binary response data (Equation (5.1)). At location x_i, the response variable Y_{ij} is a binary indicator taking value 1 if the individual has been tested positive for malaria and 0 otherwise. The index i represents the household, and the index j represents an individual within the household. Y_{ij} are mutually independent Bernoulli variables with probit link function $\Phi^{-1}(\cdot)$, conditional on an unobserved spatial stochastic process $S(x)$, hence the conditional mean number of positive rapid diagnostic tests (RDT) at location x_i depends on explanatory variables (d_{ij} i.e. both at individual- and household- level) observed at location x_i and on $S(x_i)$, and p_{ij} is the probability that an individual j at location x_i will have a positive RDT. We modelled $S(x)$ as a Gaussian process with mean zero, variance σ^2 and Matérrn correlation structure:

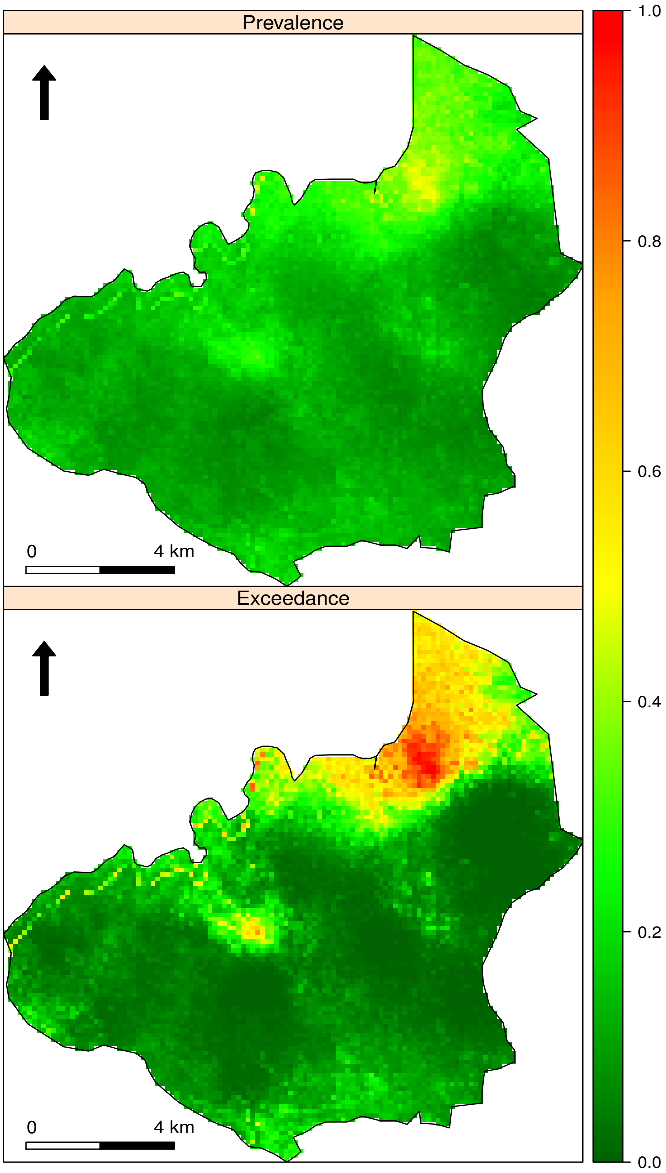
$$\mathbf{Corr}[S(x), S(x')] = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}\mathcal{K}_{\kappa}(u/\phi) \tag{5.3}$$

where u is the distance between x and x', ϕ is the scale of spatial correlation and κ is the smoothness parameter. The term $S(x)$ in Equation (5.1) captures the residual spatial variation after adjusting for the covariates.

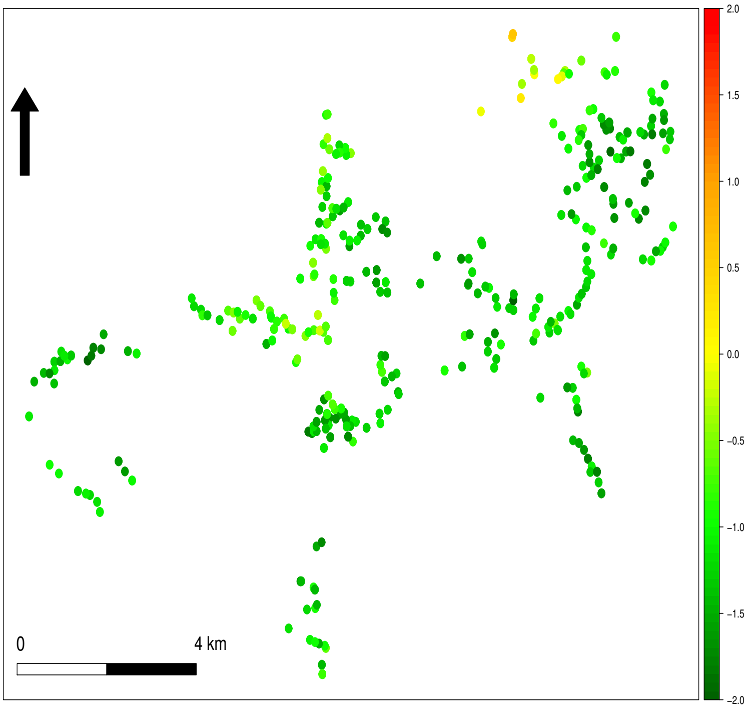
5.6.3 Malaria prevalence in children 6–59 months in focal area A.

Figure 5.5 (top panel) shows the prevalence of positive malaria RDT in children in focal area A. The north-eastern and central parts of the area show localised malaria hotspots. Overall, prevalence is generally low in focal area A, also in comparison with focal areas B and C, focal area A has the lowest malaria prevalence. Figure 5.5 (bottom panel) shows the map of exceedance probabilities that prevalence is over the national threshold of 30 %. The clear distinction in prevalence between the upper and lower parts of north-east of focal area A is interesting. It shows an example where increased sampling effort is needed to in order to look into and understand other factors driving prevalence, i.e. social behaviour or environmental factors.

Figure 5.6 shows the log odds of the predicted prevalence at each of the observed locations in focal area A as explained by linear regression explanatory variables, see Table 5.3 above, and the contribution of the stochastic process $S(x)$. Households in the north-west of the area show higher prevalence than the rest of the area.



Supplementary Figure 5.5: Malaria prevalence in children 6–59 months in focal area A (top panel). The bottom panel shows the map of exceedance probabilities $P(x; 0.3)$ for the Bayesian prediction.

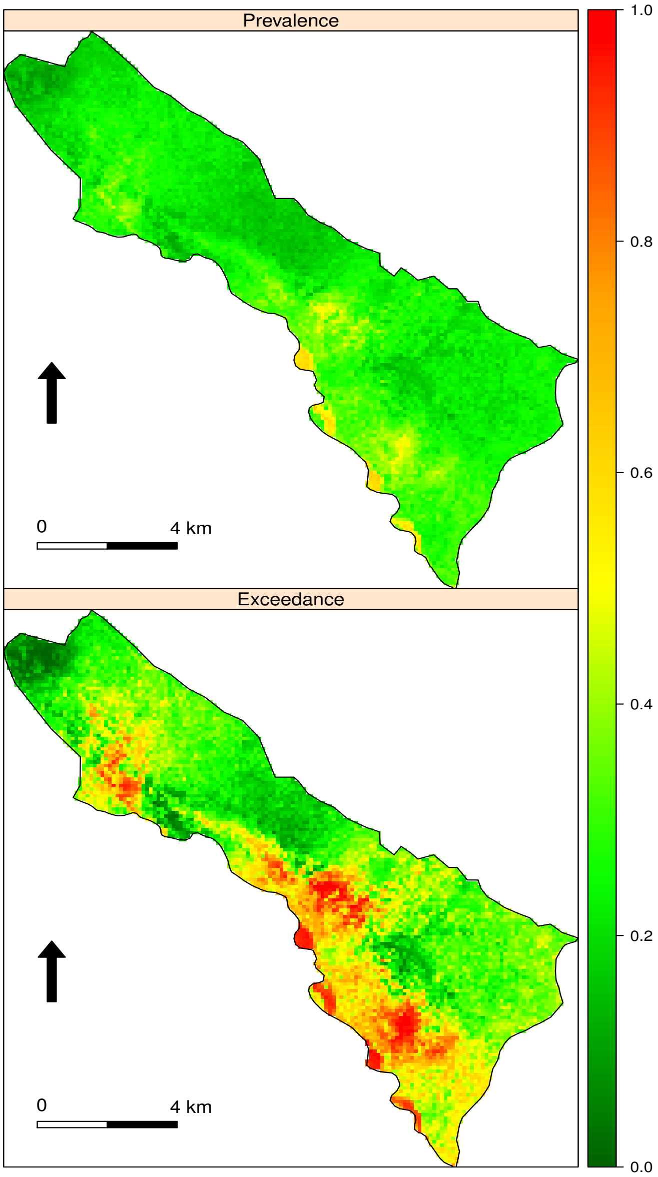


Supplementary Figure 5.6: Contributions of the linear regression and of the unexplained spatial variation to the predicted log-odds of malaria prevalence in children 6–59 months at each of the observed locations in focal area A.

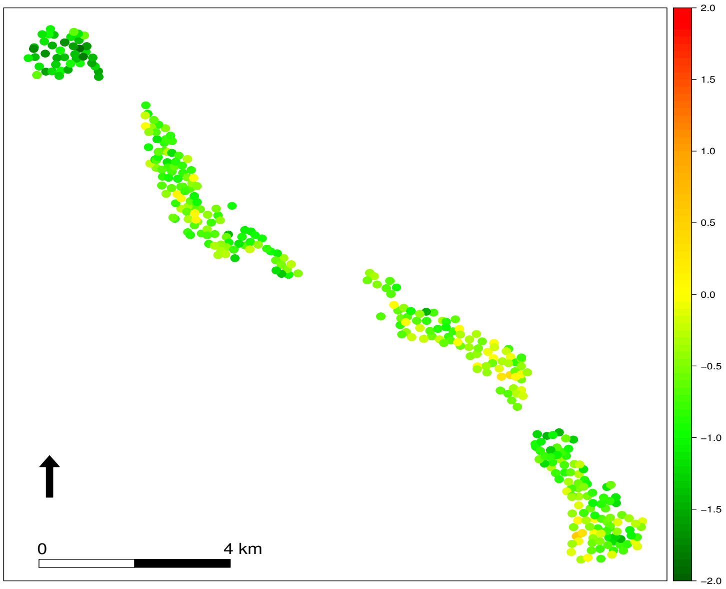
5.6.4 Malaria prevalence in children 6–59 months in focal area C.

Figure 5.7 (top panel) shows the prevalence of positive malaria RDT in children in focal area C. The area has several localised malaria hotspots throughout its length. Overall, prevalence is relatively lower in focal area C, in comparison with focal area B but higher than focal area A. Figure 5.7 (bottom panel) shows the map of exceedance probabilities that prevalence is over 30 %. Similar to other focal areas, focal area C has adjacent areas with largely antithetical prevalence.

Figure 5.8 shows the log odds of the predicted prevalence at each of the observed locations in focal area C as explained by linear regression explanatory variables and the contribution of the stochastic process $S(x)$. Households at higher altitude show lower prevalence as compared to households at a lower altitude. A large proportion of the households shows a high prevalence in focal area C.



Supplementary Figure 5.7: Malaria prevalence in children 6–59 months in focal area C (top panel). The bottom panel shows the map of exceedance probabilities $P(x; 0.3)$ for the Bayesian prediction.



Supplementary Figure 5.8: Contributions of the linear regression and of the unexplained spatial variation to the predicted log-odds of malaria prevalence in children 6–59 months at each of the observed locations in focal area C.

References

Afrane, Y.A., Zhou, G., Githeko, A.K. and Yan, G. (2013) 'Utility of Health Facility-based Malaria Data for Malaria Surveillance', *PLoS ONE* 8 (2), e54305.

Alemu, K., Worku, A. and Berhane, Y. (2013) 'Malaria infection has spatial, temporal, and spatiotemporal heterogeneity in unstable malaria transmission areas in northwest Ethiopia', *PLoS ONE* 8 (11), e79966.

Amexo, M., Tolhurst, R., Barnish, G. and Bates, I. (2004) 'Malaria misdiagnosis: Effects on the poor and vulnerable', *Lancet* 364 (9448), pp. 1896–1898.

Baidjoe, A.Y., Stevenson, J., Knight, P., Stone, W., Stresman, G., Osoti, V., Makori, E., Owaga, C., Odongo, W., China, P., Shagari, S., Kariuki, S., Drakeley, C., Cox, J. and Bousema, T. (2016) 'Factors associated with high heterogeneity of malaria at fine spatial scale in the Western Kenyan highlands', *Malaria Journal* 15, p. 307.

Beron, K.J. and Vijverberg, W.P.M. (2004) 'Probit in a Spatial Context: A Monte Carlo Analysis', *Advances in Spatial Econometrics: Methodology, Tools and Applications.* Ed. by L. Anselin, S.J. Rey and R.J.G.M. Florax, 1st ed., Berlin Heidelberg: Springer, chap. 8, pp. 169–195.

Berrett, C. and Calder, C.A. (2012) 'Data augmentation strategies for the Bayesian spatial probit regression model', *Computational Statistics and Data Analysis* 56 (3), pp. 478–490.

Bhatt, S., Weiss, D.J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K.E., Moyes, C.L., Henry, A., Eckhoff, P.A., Wenger, E.A., Briet, O., Penny, M.A., Smith, T.A., Bennett, A., Yukich, J., Eisele, T.P., Griffin, J.T., Fergus, C.A., Lynch, M., Lindgren, F., Cohen, J.M., Murray, C.L.J., Smith, D.L., Hay, S.I., Cibulskis, R.E. and Gething, P.W. (2015) 'The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015', *Nature* 526 (7572), pp. 207–211.

Bousema, T., Griffin, J.T., Sauerwein, R.W., Smith, D.L., Churcher, T.S., Takken, W., Ghani, A., Drakeley, C. and Gosling, R. (2012) 'Hitting Hotspots: Spatial Targeting of Malaria for Control and Elimination', *PLoS Medicine* 9 (1), e1001165.

Burton, D.C., Flannery, B., Onyango, B., Larson, C., Alaii, J., Zhang, X., Hamel, M.J., Breiman, R.F. and Feikin, D.R. (2011) 'Healthcare-seeking behaviour for common infectious disease-related illnesses in rural Kenya: A community-based house-to-house survey', *Journal of Health, Population and Nutrition* 29 (1), pp. 61–70.

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

Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2016a) 'Adaptive geostatistical design and analysis for prevalence surveys', *Spatial Statistics* 15, pp. 70–84.

Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2016b) 'Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure', *Enviromentrics (in press)*, pp. 1–11.

Giorgi, E., Sesay, S.S.S., Terlouw, D.J. and Diggle, P.J. (2015) 'Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models', *Journal of the Royal Statistical Society. Series A: Statistics in Society* 178 (2), pp. 445–464.

Gosoniu, L., Msengwa, A., Lengeler, C. and Vounatsou, P. (2012) 'Spatially explicit Burden estimates of malaria in Tanzania: Bayesian geostatistical modeling of the malaria indicator survey data', *PLoS ONE* 7 (5), e23966.

Grimes, J.E.T. and Templeton, M.R. (2016) 'Geostatistical modelling of schistosomiasis prevalence', *The Lancet Infectious Diseases* 15 (8), pp. 869–870.

Kazembe, L.N., Kleinschmidt, I. and Sharp, B.L. (2006) 'Patterns of malaria-related hospital admissions and mortality among Malawian children: an example of spatial modelling of hospital register data', *Malaria Journal* 5, p. 93.

Kazembe, L., Kleinschmidt, I., Holtz, T. and Sharp, B. (2006) 'Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data', *International Journal of Health Geographics* 5, p. 41.

Kondo, M.C., Bream, K.D.W., Barg, F.K. and Branas, C.C. (2014) ‘A random spatial sampling method in a rural developing nation.’, *BMC Public Health* 14, p. 338.

Matérn, B. (1986) *Spatial Variation*, 2nd ed., Berlin: Springer.

MERIT Initiative (2012) *Meningitis Environmental Risk Information Technologies*, http://merit.hc-foundation.org/ProgramActivities2012.html.

Mzilahowa, T., Hastings, I.M., Molyneux, M.E. and McCall, P.J. (2012) ‘Entomological indices of malaria transmission in Chikhwawa district, Southern Malawi.’, *Malaria Journal* 11, p. 380.

Noor, A.M., Kinyoki, D.K., Mundia, C.W., Kabaria, C.W., Mutua, J.W., Alegana, V.A., Fall, I.S. and Snow, R.W. (2014) ‘The changing risk of Plasmodium falciparum malaria infection in Africa: 2000-10: A spatial and temporal analysis of transmission intensity’, *The Lancet* 383 (9930), pp. 1739–1747.

Patil, A.P., Gething, P.W., Piel, F.B. and Hay, S.I. (2011) ‘Bayesian geostatistics in health cartography: the perspective of malaria’, *Trends in Parasitology* 27 (6), pp. 246–253.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*, https://www.R-project.org/, Vienna, Austria.

Reid, H., Haque, U., Clements, A.C.A., Tatem, A.J., Vallely, A., Ahmed, S.M., Islam, A. and Haque, R. (2010) ‘Mapping malaria risk in Bangladesh using Bayesian geostatistical models’, *American Journal of Tropical Medicine and Hygiene* 83 (4), pp. 861–867.

Roca-Feltrer, A., Lalloo, D.G., Phiri, K. and Terlouw, D.J. (2012) 'Short Report : Rolling Malaria Indicator Surveys (rMIS): a potential district-level malaria monitoring and evaluation (M & E) tool for program managers.', *American Journal of Tropical Medicine and Hygiene* 86 (1), pp. 96–98.

Roll Back Malaria Partnership (2016) *MIS Toolkit*, http://www.malariasurveys.org/.

Rowe, A.K. (2009) 'Potential of integrated continuous surveys and quality management to support monitoring, evaluation, and the scale-up of health interventions in developing countries', *American Journal of Tropical Medicine and Hygiene* 80 (6), pp. 971–979.

Rowe, A.K., Kachur, S.P., Yoon, S.S., Lynch, M., Slutsker, L. and Steketee, R.W. (2009) 'Caution is required when using health facility-based data to evaluate the health impact of malaria control efforts in Africa', *Malaria Journal* 8, p. 209.

Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications.* London: Chapman & Hall/CRC.

Schur, N., Hürlimann, E., Garba, A., Traoré, M.S., Ndir, O., Ratard, R.C., Tchuem Tchuenté, L.-A., Kristensen, T.K., Utzinger, J. and Vounatsou, P. (2011) 'Geostatistical Model-Based Estimates of Schistosomiasis Prevalence among Individuals Aged less than 20 Years in West Africa', *PLoS Neglected Tropical Diseases* 5 (6), e1194.

Snow, R.W., Craig, M., Deichmann, U. and Marsh, K. (1999) 'Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population', *Bulletin of the World Health Organization* 77 (8), pp. 624–640.

Stanton, M.C., Agier, L., Taylor, B.M. and Diggle, P.J. (2014) 'Towards real-time spatiotemporal prediction of district level meningitis incidence in sub-Saharan Africa', *Journal of the Royal Statistical Society. Series A: Statistics in Society* 177 (3), pp. 661–678.

Vyas, S. and Kumaranayake, L. (2006) 'Constructing socio-economic status indices: How to use principal components analysis', *Health Policy and Planning* 21 (6), pp. 459–468.

Walker, P.G.T., Griffin, J.T., Ferguson, N.M. and Ghani, A.C. (2016) 'Estimating the most efficient allocation of interventions to achieve reductions in Plasmodium falciparum malaria burden and transmission in Africa: a modelling study', *The Lancet Global Health* 4 (7), e474–e484.

World Health Organisation (2015a) *Global technical strategy for malaria 2016-2030*, tech. rep., Geneva: World Health Organisation.

World Health Organisation (2015b) *World Malaria Report 2015*, tech. rep., Geneva: World Health Organisation.

Chapter 6

General discussion, conclusions and future research.

This thesis has described methods for non-adaptive and adaptive geostatistical designs. The need to develop user-friendly and affordable tools for monitoring transmission and to control progress of malaria in resource-limited settings motivated the development of these design approaches. However, the methods are generally applicable and therefore are relevant to other diseases and other scientific areas. The methods focus on constructing designs for prevalence surveys that are efficient for computing spatial predictions while taking the uncertainties of the parameters in the Gaussian geostatistical model into account.

In the current chapter, we further discuss each of the papers presented in early chapters. As each paper already contains its own discussion, we give only a short summary of each, outlining the main contributions, the implications of our results in malaria transmission control and discussing possible future research aimed at improving and broadening the range of applicability of the developed methodology.

Throughout the development of the methods in this thesis, we made stationary Gaussian assumptions for the underlying random field for the geostatistical model in Equation (1.1), one consequence of which is that a linear function of the data that gives the best minimum mean square error (MMSE) prediction of the value of the surface at an arbitrary location x_0, say. The stationarity assumption implies that the spatial correlation is a function of the distance between locations and independent of locations themselves. These are reasonable and widely used assumptions; see, for example, Journel and Huijbregts (1978), De Oliveira, Kedem and Short (1997) and Diggle and Ribeiro (2002).

In Chapter 2, we proposed a class of inhibitory non-adaptive geostatistical designs. We developed designs that compromise between efficient parameter estimation and spatial prediction. The basic idea was to construct a design criterion based on the variance of the predictive distribution, in which parameter uncertainties are included. Through simulation studies, comparison to existing optimal designs and an application to malaria prevalence data from the Majete malaria project in Chikwawa district, southern Malawi, we showed that inhibitory designs with an intermediate number of close pairs give the best performance.

ICP designs not only give the best performance, but they are also easy to implement by an average practitioner as compared to other existing optimal design strategies. In terms of design computation time, ICP designs showed a 16 fold improvement (in computational speed) over the existing optimal simulated annealing-based designs on a high end computing (HEC) cluster at Lancaster University. During simulation studies, we were able to compute ICP designs on a Lenovo Z70 laptop personal computer in under 30 minutes. Hence, we recommend that where the same data are going to be used for parameter estimation and spatial prediction, ICP designs should be used.

In Chapter 3, we proposed adaptive geostatistical designs for prevalence surveys. These designs allow collection of geostatistical data over time to depend on information obtained from previous information to optimise data collection towards the analysis objective(s). Through simulation studies and an application to malaria prevalence data from the Majete malaria project in Chikwawa district, southern Malawi, we showed that using an adaptive design for fine-scale risk mapping can lead to more accurate prevalence predictions. This approach finds wide applicability in resource-limited settings, where accurate registry data typically are not available.

One of the main contributions of this paper is that it develops a geostatistical design methodology that allows probabilistic decision making in the identification of sampling sites. Adaptive designs allow model fitting and refining as data becomes available, with future sampling locations chosen accordingly. This is advantageous in the sense that a good choice of model may not be known in practice until data collection has started. Furthermore, the paper demonstrates the practical implementation of the adaptive design methodology. It outlines algorithms that can easily be adapted for implementation in different software and scenarios, ready for use.

We used the spatial integral of the predictive variance Equation (3.8) as an appropriate measure of predictive performance. This holds for the application we demonstrate in the paper. However, there are other criteria that can be used. Maximum prediction variance,

$$MPV = \max_{x \in \mathcal{D}} [\text{Var}(S(x))] \tag{6.1}$$

is another criterion that has been used to measure the predictive performance of a design, see, for example, Zimmerman (2006) and Fanshawe and Diggle (2012). The criterion we use in this paper would not hold if, for example, we need to delineate areas where disease prevalence is above or below a given threshold $\boldsymbol{c}$ or if $S(x)$ represents pollution (as in the case of the Galicia lead pollution example) and the main objective is to monitor compliance with environmental standards. Fanshawe and Diggle (2012) use as a criterion,

$$-\int_{\mathcal{D}} \{P(S(x) > c) - p_0\}^2 dx, \tag{6.2}$$

where $\boldsymbol{c}$ and p_0 are fixed values, to be specified in advance, the aim being to most clearly delineate regions that lie above or below a policy-relevant threshold, $\boldsymbol{c}$. In our example, this criterion can be used to identify areas where disease prevalence or pollution is above a certain threshold $\boldsymbol{c}$ that would trigger deployment of targeted interventions by programme implementers. This criterion can potentially be used to further extend adaptive methods in constructing designs that allow focused sampling effort in sub-areas with more interesting outcomes than other sub-areas. Additionally, in environmental monitoring network design applications, a mobile monitoring network can be used. This gives rise to a dynamic sampling design problem, for which an adaptive design methodology would also be appropriate.

We develop and demonstrate the methodology in a malaria monitoring application. The AGD sampling, as currently demonstrated, requires a complete enumeration census of locations that must be geo-referenced in order to create a sampling frame, from which samples can be drawn. This, for small scale studies (i.e. sub-district level) is feasible and not an overly arduous process. However, when expanding to regional or national level, the lack of such a sampling frame becomes a major challenge to the implementation of such sampling design methods. One possible solution to this challenge would be to employ a two-stage stratified sampling procedure, in which the study area is divided into strata then apply *adaptivity* at either stratum level or within the stratum. We elaborate on this point later in Section 6.1.

In Chapter 4, we presented a commentary on a paper by Ferreira and Gamerman (2015) which addressed a topical issue in geostatistics, namely the effect of preferential sampling of the locations at which a spatial process is measured. It has been widely shown and discussed that when the choice of spatial sampling locations is consciously or unconsciously biased in some way, say by practical demands (for example, in environmental monitoring network applications), this can lead to preferential sampling. This happens when the process that determines the locations (for example, monitoring sites) and the process being modelled (for example, pollution concentration) are stochastically dependent in particular ways (Diggle and Giorgi, 2017). Recent studies include Diggle, Menezes and Su (2010), Pati, Reich and Dunson (2011), Gelfand, Banerjee and Finley (2012), Shaddick and Zidek (2014) and Zidek, Shaddick and Taylor (2014). The selective sampling materially affects parameter estimation as well as prediction, both of which may become biased. In our discussion, we introduced adaptive designs, which need not be preferential. We also discussed practical considerations that may constrain the geostatistical design problem in particular applications, including an application in which we developed and applied adaptive geostatistical designs.

Chapter 5 presented the first field epidemiological application of AGD sampling, involving a sequence of continuous malaria prevalence surveys to measure the disease's spatial heterogeneity in declining transmission setting, in rural Malawi. The geostatistical binary probit model in Section 5.2.5 incorporated spatial correlation as well as individual-specific, household-specific and environmental covariates to produce fine-scale spatial prediction for malaria prevalence in children 6–59 months. With this model, we were able to detect malaria hotspot areas and the underlying malaria spatial heterogeneity over the study area.

With the current advances in geostatistical modelling, the growing need for accurate high spatial resolution for fine scale mapping of disease burden in view of decreasing malaria transmission, and the need for prudent allocation of resources, AGD sampling is a preferred option for continuous disease prevalence monitoring. These novel methods can inform more efficient design and analysis of surveys aimed at understanding geographical variations in intervention coverage and health outcomes, especially at sub-district scales in low-resource settings.

This study provides researchers and programme implementers with a valuable alternative for monitoring fine-scale disease prevalence needed to identify and target high burden areas and hotspots. The approach outlined in this paper can be used to complement periodical national disease surveys such as malaria indicator surveys (MIS), demographic and health surveys (DHS) and Multiple Indicator Cluster Surveys (MICS) to monitor malaria transmission and control progress both at sub-national and sub-district levels. Most malaria spatial modelling is still carried out in a research setting rather than in programme implementation. National control programmes especially in resource-limited settings should incorporate these new technologies to guide targeting of interventions. This Chapter contributes to current advances in easy-to-use geostatistical modelling and its application in disease transmission control and monitoring.

6.1 Future work

The stationarity assumption we make in the thesis is widely accepted. However, in practice, local covariance may vary with spatial location, thus a stationary covariance model may not be appropriate. For example, in our application, when malariological indices are modelled, local characteristics related to human activities, land use, environment and vector ecology influence spatial correlation differently at the different locations. An extension of the methodology that would be of interest, therefore, is to explore the effect of non-stationary covariance structure on non-adaptive and adaptive sampling designs. Another extension would be to construct designs that would be best for prediction of specified non-linear functionals of $S(\cdot)$.

The above-mentioned characteristics (i.e. environmental, climate and human activities) are also known to play an important role in determining how vector distribution and vector-borne disease (e.g. malaria) epidemiology vary over time (Machault et al., 2011). With our current implementation of the AGD technique in Chapter 3, the algorithm identifies locations with high spatial prediction uncertainty and uses this to allocate subsequent sampling at those locations. It does not capture the impact of temporal variation in risk factors on health outcomes. A possible extension of the methodology, therefore, is to take these temporal processes into account, leading to further gains in efficiency of sampling and more accurate risk maps. Malaria is known to be highly seasonal, therefore an important extension would be to allow the design to capture the temporal component in predicting a moving target.

In this thesis, we demonstrate that the application of AGD in continuous malaria prevalence monitoring is feasible for mapping malaria burden in resource-limited settings, albeit on a small study region, namely an area surrounding Majete Wildlife Reserve. An interesting future research question is to focus on national and/or regional scalability of these novel options to conduct fine-scale monitoring of malaria burden and control progress. AGD disease monitoring methodology can offer guidance to accelerate transmission reduction efforts at these levels, through targeting interventions in the identified malaria transmission hotspots.

However, as highlighted earlier on, the lack of a sampling frame at these extended scales limits the application of AGD strategy. This challenge begs for further extension of methodology. An interesting area of research is therefore to explore stratification techniques and apply "multiple phase" adaptive sampling at either or both of two spatial scales namely "stratum" or "within stratum" level. One strategy would be to fix the within strata sample sizes and adaptively select new strata over time. Another strategy would be to fix the number of selected strata and apply adaptivity within the strata over time. The third strategy would be to apply adaptivity at both between and within strata, over time.

Additionally, a study on practicality and cost-effectiveness of the methodology in field applications would be of interest to programme implementers and funders. However, we have demonstrated that adaptive sampling can be easy to use and affordable to implement; see, for example, Chipeta et al. (2016a). Specifically:

- samples are only collected from locations that will deliver useful, additional information in order to understand the heterogeneity of phenomenon of interest (i.e. disease) throughout the study region;
    - ∗ adaptive sampling techniques can reduce both costs and time for carrying out surveys as well as improve the precision of the results for a

given sample size;

- adaptive allocations require a smaller sample size than non-adaptive sampling in achieving a specified level of predictive accuracy.

In the present thesis, we have developed methodology and implemented it in a malaria setting. Another question of interest for future research is to implement these methods to diseases with similar heterogeneity patterns. Neglected tropical diseases such as soil-transmitted helminths (STH) and schistosomiasis exhibit high burden especially in resource-limited areas (Hotez and Kamath, 2009; Schur et al., 2011; Chipeta, Ngwira and Kazembe, 2013; Phiri, Ngwira and Kazembe, 2016). Registries for such diseases in these areas typically do not exist, or where they exist, they are usually incomplete or inaccurate. Adaptive sampling methods could be applied to identify and increase accuracy in hotspots' mapping in order to target them with interventions such as mass drug administration (MDA). Adaptive and non-adaptive geostatistical design methods can and should be used in various other scientific areas and study fields such as crime mapping and environmental studies like pollution level compliance or monitoring.

References

Chipeta, M.G., Ngwira, B. and Kazembe, L.N. (2013) 'Analysis of Schistosomiasis haematobium Infection Prevalence and Intensity in Chikhwawa, Malawi: An Application of a Two-Part Model', *PLoS Neglected Tropical Diseases* 7 (3), e2131.

Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2016a) 'Adaptive geostatistical design and analysis for prevalence surveys', *Spatial Statistics* 15, pp. 70–84.

De Oliveira, V., Kedem, B. and Short, D.A. (1997) 'Bayesian prediction of transformed Gaussian random fields', *Journal of the American Statistical Association* 92 (440), pp. 1422–1433.

Diggle, P.J. and Giorgi, E. (2017) 'Preferential sampling of exposure levels', *unpublished.*

Diggle, P.J., Menezes, R. and Su, T.-L. (2010) 'Geostatistical inference under preferential sampling', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59 (2), pp. 191–232.

Diggle, P.J. and Ribeiro, J.P. (2002) *Bayesian inference in Gaussian model-based geostatistics.* Tech. rep., Lancaster University.

Fanshawe, T.R. and Diggle, P.J. (2012) 'Adaptive Sampling Design for Spatio-Temporal Prediction', *Spatio-temporal design: Advances in efficient data acquisition*, ed. by J. Mateu and W.G. Müller, 1st ed., Chichester, UK: John Wiley & Sons, Ltd, chap. 11, pp. 249–268.

Ferreira, G.d.S. and Gamerman, D. (2015) 'Optimal Design in Geostatistics under Preferential Sampling', *Bayesian Analysis* 10 (3), pp. 711–735.

Gelfand, A.E., Banerjee, S. and Finley, A.O. (2012) 'Spatial Design for Knot Selection in Knot-Based Dimension Reduction Models', *Spatio-temporal design: Advances in efficient data acquisition*, ed. by J. Mateu and W.G. Müller, 1st ed., Chichester, UK: John Wiley & Sons, Ltd, chap. 7, pp. 142–169.

Hotez, P.J. and Kamath, A. (2009) 'Neglected tropical diseases in sub-Saharan Africa: Review of their prevalence, distribution, and disease burden', *PLoS Neglected Tropical Diseases* 3 (8), pp. 2–11.

Journel, A.G. and Huijbregts, C.J. (1978) *Mining Geostatistics*, London: Academic Press.

Machault, V., Vignolles, C., Borchi, F., Vounatsou, P., Pages, F., Briolant, S., Lacaux, J.P. and Rogier, C. (2011) 'The use of remotely sensed environmental data in the study of malaria', *Geospatial Health* 5 (2), pp. 151–168.

Pati, D., Reich, B.J. and Dunson, D.B. (2011) 'Bayesian geostatistical modelling with informative sampling locations', *Biometrika* 98, pp. 35–48.

Phiri, B.B.W., Ngwira, B. and Kazembe, L.N. (2016) 'Analysing risk factors of co-occurrence of schistosomiasis haematobium and hookworm using bivariate regression models: Case study of Chikwawa, Malawi', *Parasite Epidemiology and Control* 1 (2), pp. 149–158.

Schur, N., Hürlimann, E., Garba, A., Traoré, M.S., Ndir, O., Ratard, R.C., Tchuem Tchuenté, L.-A., Kristensen, T.K., Utzinger, J. and Vounatsou, P. (2011) 'Geostatistical Model-Based Estimates of Schistosomiasis Prevalence among Individuals Aged less than 20 Years in West Africa', *PLoS Neglected Tropical Diseases* 5 (6), e1194.

Shaddick, G. and Zidek, J.V. (2014) 'A case study in preferential sampling: Long term monitoring of air pollution in the UK', *Spatial Statistics* 9, pp. 51–65.

Zidek, J.V., Shaddick, G. and Taylor, C.G. (2014) 'Reducing estimation bias in adaptively changing monitoring networks with preferential site selection', *The Annals of Applied Statistics* 8 (3), pp. 1640–1670.

Zimmerman, D.L. (2006) 'Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction', *Environmetrics* 17 (6), pp. 635–652.

Appendix A

Worked examples of inhibitory and adaptive designs.

A.1 Hypothetical data.

In this Appendix, we show a worked example using the functions given in Appendix B. We simulate hypothetical Binomial data to illustrate the usage of the functions to construct an "*adaptive design*". We compare the predictive performance of the adaptive design with a completely random design. For these data, a zero-mean Gaussian process is generated over a 64 by 64 grid covering the unit square $[0,1]^2$, with parameters: $\sigma^2 = 1$, $\phi = 0.1$ and $\kappa = 2$; the nugget effect is not included, hence $\tau^2 = 0$. Binomial observations, with 7 trials at each grid point and probabilities given by the anti-logit of the simulated values of the Gaussian process, constitute the variable Y in the data. We use `geoR` function `grf( )` to generate a simulation of a two-dimensional Gaussian process.

A.2 Design construction.

Using `R` code below, we now show how to construct an initial non-adaptive *simple inhibitory design*, by sampling 70 locations, using `inhibitory.design` function (see the algorithm in Section 2.3.2). We analyse these data and estimate model parameters by fitting a Binomial logistic regression model using `binomial.logistic.MCML` function. Spatial prediction is carried out using `spatial.pred.binomial.MCML` function. Both functions are from `PrevMap`, an `R` package for analysing spatially referenced prevalence data. Figure A.1 shows the initial design with a minimum distance $\delta = 0.1$.

```r
inhibitory.sample <-
  inhibitory.design(dataframe = myDF, coords.col = 1:2, k = 0,
                    data.col = c(3:5), size = 70, delta = 0.1,
                    zeta = 0)
sample.DF <- inhibitory.sample[[1]]
plot(inhibitory.sample[[1]][,1:2], pch = 19, col = "blue",
     cex = 1, xlab = "Xcoord", ylab = "Ycoord", cex.lab = 1.5,
     cex.axis=1)
```

We use data from the simple inhibitory design sample to fit a binomial logistic model (see Equation (1.21)) with no covariates. Results from initial data analysis show an average prediction variance of 0.01049. Figures A.2a and A.2b show prevalence predictions and their standard errors, respectively. Using results from the initial analysis of data from the 70 sample locations, we now show how we use `adaptive.design` function to construct a *batch adaptive design* by sampling a further 30 locations in batches of 5 (see the algorithm in Section 3.3.5). Figure A.3 shows adaptively added sample locations (red dots) and initial locations (blue dots). Note that in order to implement a *singleton adaptive design*, the batch size in the `R` code below needs to be changed to 1.

```r
counter <- 0

while(counter<100){

  my.adapt.sample <-

  adaptive.design(dataframe = myDF,coords.col = 1:2,

                  delta = 0.07, pred.var = pred.v, batch = 5,

                  mysample = my.adapt.sample[,1:5])

  counter <- dim(my.adapt.sample)[1]

  ##' Fit the Binomial logistic model

  adapt.LBGM.Fit <-

  binomial.logistic.MCML(RDT ~ 1, units.m = ~Units.m,

                         coords = ~Longitude + Latitude,

                         data = my.adapt.sample, par0 = par0,

                         control.mcmc = c.mcmc,

                         method = "nlminb", kappa = 2,

                         fixed.rel.nugget = 0,
                         start.cov.pars = par0[3],
                         messages = FALSE,
                         plot.correlogram = FALSE)
  par0 <- c(adapt.LBGM.Fit$estimate[1],
            exp(adapt.LBGM.Fit$estimate[2]),
            exp(adapt.LBGM.Fit$estimate[3]))
  ##' Spatial predictions
  adapt.Spat.Pred <-
  spatial.pred.binomial.MCML(adapt.LBGM.Fit,grid.pred,
                             control.mcmc = c.mcmc,
                             type = "marginal",
                             standard.errors = TRUE,
                             scale.predictions = "prevalence",
                             messages = FALSE)
  pred.v <- adapt.Spat.Pred$prevalence$standard.errors
  adapt.apv <- (mean(pred.v))^2
}
```

```
plot(my.adapt.sample[,1:2], pch = 19, col = "red", cex = 1,
      xlab = "Xcoord", ylab = "Ycoord", cex.lab = 1.5, cex.axis=1)
points(inhibitory.sample[[1]][,1:2], pch = 19, col = "blue",
       cex = 1)
```

Figures A.4a and A.4b show prevalence predictions and their standard errors, respectively, from an adaptive design with 100 sample locations. The adaptive design has an average prediction variance of 0.00793.

```
plot(adapt.Spat.Pred,type = "prevalence",
     summary = "predictions", xlab = "Xcoord",
     ylab = "Ycoord", cex.lab=1.5, cex.axis=1)
contour(adapt.Spat.Pred,type = "prevalence",
        summary = "predictions",nlev = 3, add = TRUE)
```

```
plot(adapt.Spat.Pred,type = "prevalence",
     summary = "standard.errors", maxpixels = 500000,
     alpha = 1, xlab = "Xcoord", ylab = "Ycoord",
     cex.lab = 1.5, cex.axis=1)
```

We compare the adaptive design's predictive performance with a random design, we sample with $n = 100$ using the `R` code below. Figure A.5 shows sample locations for the random design. Figures A.6a and A.6b show prevalence predictions and their standard errors, respectively, from the random design. The average prediction variance for the random design is 0.01933.

```
N <- dim(myDF)[1]
index <- 1:N
index.sample  <- sample(index, 100, replace = FALSE)
```

```r
random.sample  <- myDF[index.sample,]
plot(random.sample[,1:2], pch = 19, col = "blue",
     cex = 1, xlab = "Xcoord", ylab = "Ycoord", cex.lab = 1.5,
     cex.axis=1)
```

```r
plot(rand.Spat.Pred,type = "prevalence",
     summary = "predictions", xlab = "Xcoord",
     ylab = "Ycoord",  cex.lab=1.5, cex.axis=1)
contour(rand.Spat.Pred,type = "prevalence",
       summary = "predictions",
       nlev = 3, add = TRUE)
```

```r
plot(rand.Spat.Pred,type = "prevalence",
     summary = "standard.errors",
     maxpixels = 500000, alpha = 1, xlab = "Xcoord",
     ylab = "Ycoord",  cex.lab = 1.5, cex.axis=1)
```

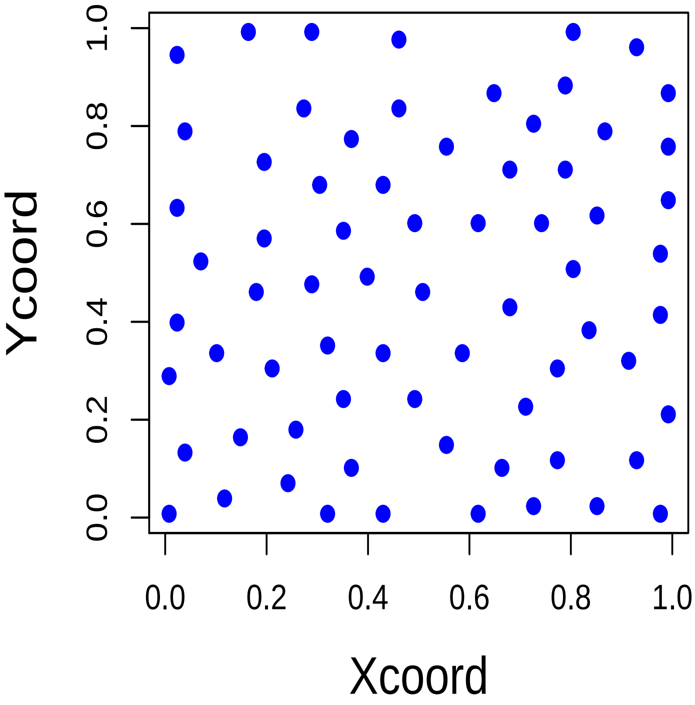
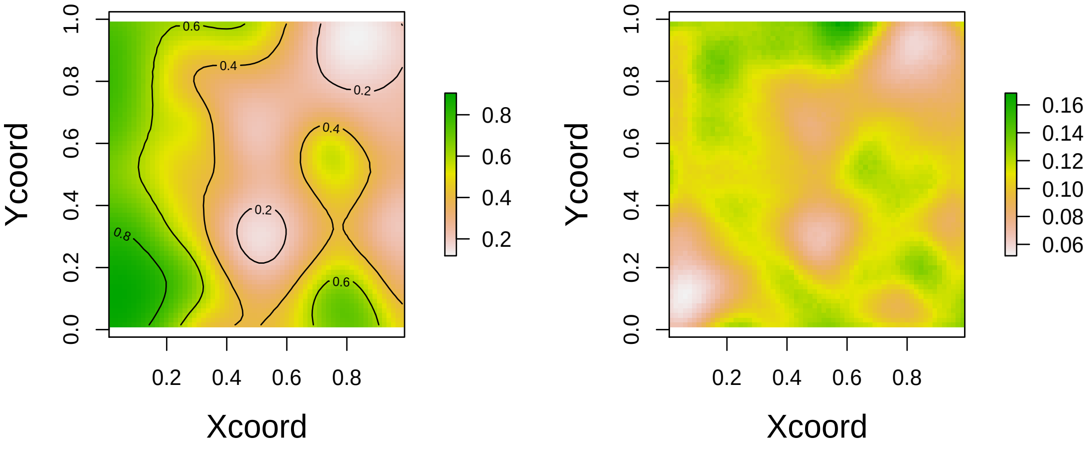


Figure A.1: Initial simple inhibitory design sample of 70 locations, $\delta = 0.1$



(a) (b)

Figure A.2: Panel (a) prevalence prediction using data from initial simple inhibitory design sample locations. Panel (b) standard errors for predictions in panel (a).

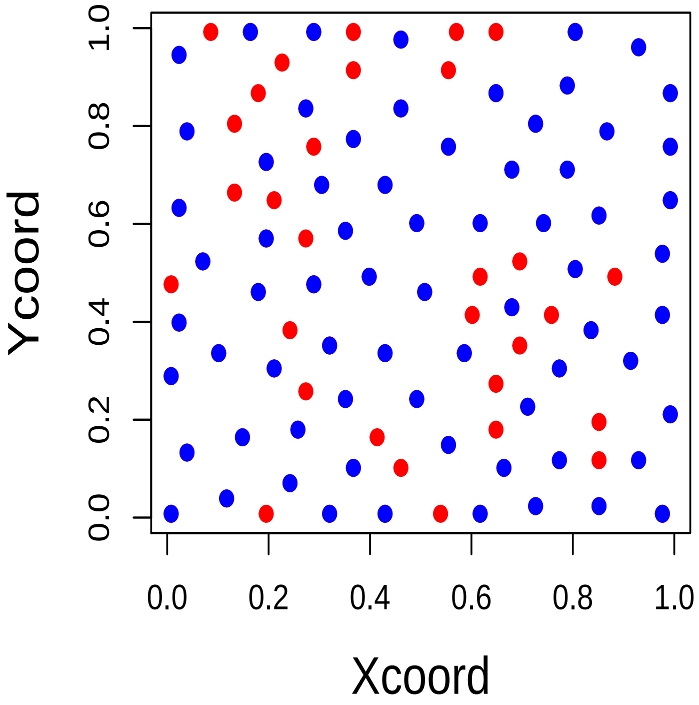


Figure A.3: Initial simple inhibitory samples (blue dots) augmented with adaptive samples (red dots), $n = 100$.

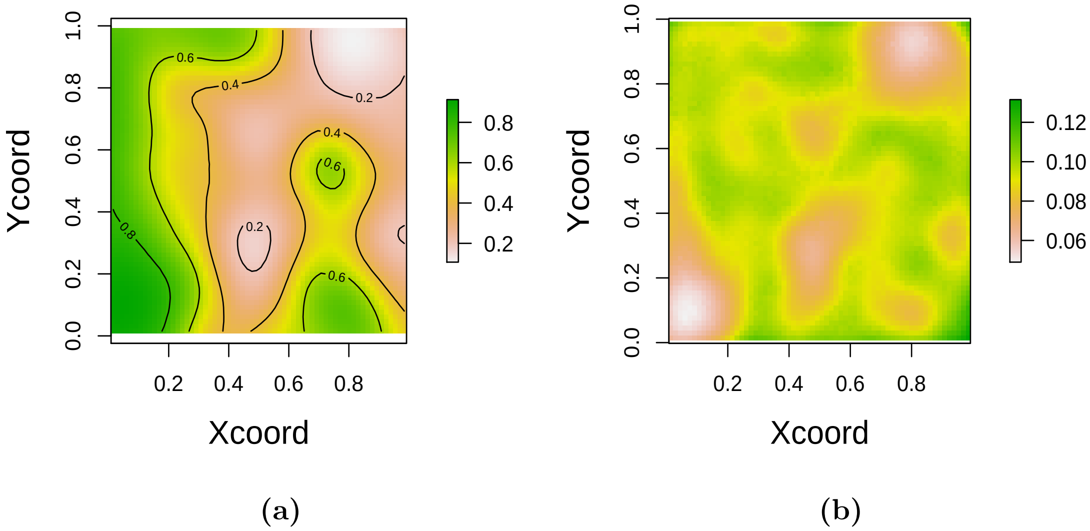


Figure A.4: Panel (a) prevalence prediction using data from adaptive design sample locations. Panel (b) standard errors for predictions in panel (a).

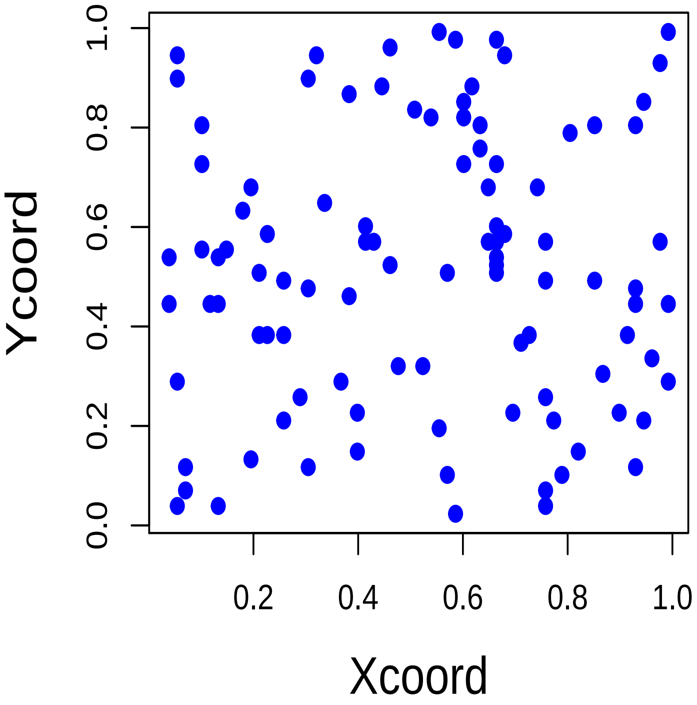
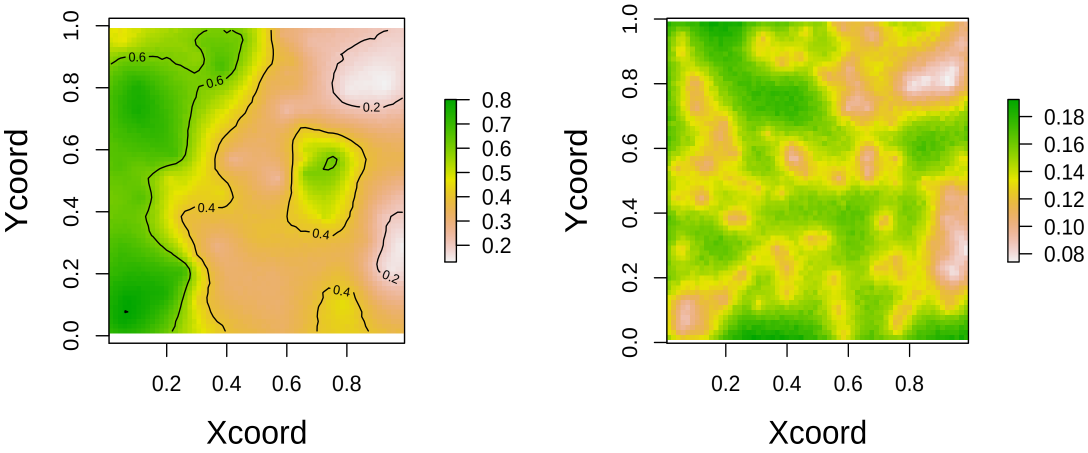


Figure A.5: Random design, $n = 100$ sample locations.



(a)

(b)

Figure A.6: Panel (a) prevalence prediction using data from random design sample locations. Panel (b) standard errors for predictions in panel (a).

Appendix B

R code for inhibitory and adaptive geostatistical designs.

B.1 The `random.design` function.

Description.

This `R` function generates a completely random sample of locations from a population of N locations, forming a grid of X - Y coordinates. The function generates sample locations without replacement. A completely random sample has n locations $x_i,\ i = 1, \ldots, n$ which are independently and uniformly distributed over the region of interest, $\mathcal{D}$. The function can be used retrospectively as well as prospectively.

Usage.

```
my.random.sample <- random.design(xycoords, n)
```

Arguments.

The required inputs for the function are:

- '**xycoords**': A matrix containing X - Y coordinates of N potential sampling locations.
- '**n**': Number of locations to sample.

Value.

The resulting output is a matrix of X - Y coordinates for the sampled locations.

Implementation.

The function `random.design` is given by:

```
random.design <- function(xycoords, n)
{
  res <- xycoords[sample(1:dim(xycoords)[1],
                           size = n, replace = FALSE), ]
  return(res)
  }
```

B.2 The `inhibitory.design` function.

Description.

This `R` function generates *non-adaptive inhibitory* sample locations, with or without close pairs, depending on the arguments (see below). The function can be used retrospectively as well as prospectively.

Usage.

```
my.inhib.sample <-
  inhibitory.design(dataframe, coords.col, data.col, delta, k,
                    size, zeta)
```

Arguments.

The required inputs for the function are:

- '**dataframe**': A data frame containing all potential sampling locations and covariates (if any). If there are no covariates, this will be a matrix of X - Y coordinates for all potential sampling locations.
- '**coords.col**': A vector specifying X - Y coordinates' columns in the data frame.
- '**data.col**': A vector specifying an n x m matrix containing covariates in the data frame.
- '**delta**': Inhibition distance or minimum distance between any two locations in the preliminary sample.
- '**k**': Number of close pairs locations (must be between 0 and $n/2$).
- '**size**': The required total sample size n.
- '**zeta**': Radius of a circle with centre x^*, one of the primary $n-k$ points within which close pairs are placed.

Value.

The function returns a list of two items namely:

- A data frame for sampled locations and their covariates (if any). Otherwise, this will be an `n x 2` matrix of X - Y coordinates for sampled locations.
- $\delta_{(k)}$ value for the $n-k$ simple inhibitory locations.

Implementation.

```
inhibitory.design <- function(dataframe, coords.col = 1:2,
                              data.col = 3, delta, k, size, zeta)
{
  if (!is.matrix(dataframe) & !is.data.frame(dataframe))
    stop("object must be a matrix or data.frame.")
  if (length(data.col) < 2)
    stop("data.names allowed only if there is more than 1 column
         of data.")
  if (any(is.na(dataframe[, coords.col]))) {
    warning("NA's not allowed in the coordinates.")
    dataframe <- dataframe[complete.cases(dataframe), drop = FALSE]
    warning("eliminating rows with NA's.")
  }
  if(any(k>size/2)){
    stop("Close pairs must be between 0 and size/2.")
  }

  ##' Inhibition distance varying with k
  delta <- delta * sqrt(size/(size - k))
  dsq  <- delta*delta
  dif <- size-k
```

```r
if(any(zeta>delta/2)){
  zeta = delta/2
  warning("Zeta > delta/2, zeta=delta/2 will be used.")
}

##' Random sample without replacement.
xy.all <- dataframe[, coords.col]
N   <- dim(xy.all)[1]
index  <- 1:N
index.sample  <- sample(index, dif, replace = FALSE)
xy.sample  <- xy.all[index.sample,]

##' Inhibition process for the n - k design points.
for (i in 2:dif){
  dmin  <- 0
  while (dmin < dsq){
    take <- sample(index, 1)
    dvec  <- (xy.all[take, 1] - xy.sample[, 1])^2 +
      (xy.all[take, 2] - xy.sample[,2])^2;dvec
    dmin  <- min(dvec);dmin
  }
  xy.sample[i,]  <- xy.all[take,]
}
```

```
colnames(xy.sample) <- c("x", "y")


##' Close pairs sampling.
if (k>0) {
  xy.cp <- matrix(NA, nrow = k, ncol = 2)
  cp.mat<-matrix(sample(1:dif,k,replace=FALSE),k,2)
  for (j in 1:k){
    take1<-cp.mat[j,1]; take2<-cp.mat[j,2]
    xy1<-c(xy.sample[take1,]); xy1 <- as.numeric(unlist(xy1))
    angle<-2*pi*runif(1, min = 0, max = 1)
    radius<-zeta*sqrt(runif(1, min = 0, max = 1))
    if(any(radius<delta/4)){
      radius = delta/4
    }
    xy.cp[j,] <-xy1+radius*c(cos(angle),sin(angle))
  }
  colnames(xy.cp) <- c("x", "y")
  xy.sample <- rbind(xy.sample, xy.cp)
}


##' Subset dataframe for sampled locations.
ind.coords <- NULL
for(i in 1:nrow(xy.sample)) {
    ind.sel <- which(xy.sample[i,1]==
                       dataframe[,coords.col[1]] &
                       xy.sample[i,2]==
                       dataframe[,coords.col[2]])
    ind.coords <- c(ind.coords,ind.sel)
  }
  inihib.DF <- dataframe[ind.coords,]

  ##' Return results.
  return(list(inihib.DF = inihib.DF, delta = delta))
}
```

B.3 The `adaptive.design` function.

Description.

The `adaptive.design` function generates *adaptive* sample locations, given the initial or existing sample locations (usually a simple inhibitory design) using the prediction variance criterion. The function can be used retrospectively as well as prospectively.

Usage.

```
my.adaptive.sample <-
  adaptive.design(dataframe, mysample, coords.col, pred.var, batch,
                  delta)
```

Arguments

The required inputs are:

- '**dataframe**': A data frame containing all potential sample locations and covariates at those locations (if any). If there are no covariates, this will be a matrix of X - Y coordinates for all potential sampling locations.
- '**mysample**': A data frame containing previously sampled locations (initial or existing sample) and covariates (if any).
- '**coords.col**': A vector specifying X - Y coordinates' columns in the data frame.
- '**pred.var**': A vector containing prediction variances for all $S(x)$.
- '**batch**': Size of the adaptive sample location(s) to be added to the initial/existing sample points.
- '**delta**': Minimum distance between any two locations in the new batch of sample locations and also from existing sample locations.

Value.

The function returns a data frame for sampled locations(existing and adaptive) and their covariates (if any). Otherwise, it returns a matrix of X - Y coordinates for sampled locations.

Implementation.

```r
adaptive.design <- function(dataframe, mysample, coords.col,
                                  pred.var, batch, delta)
  {
      #Order prediction variance
      pred.v.sort <- order(pred.var, decreasing = T)

      totalbatch = 1
      counter = 1
      rejected <- NULL
      mylocations <- mysample[,c(coords.col)]

      while(totalbatch <= batch){
            #' Calculate distance from high prediction variance
            #  location to existing samples
            distance <-pdist(mylocations,dataframe[,c(coords.col)]
                              [pred.v.sort[counter],])@dist
```

```
          min.dist <- min(distance)


          #' If the location with highest variance is away from
          #'  existing samples, we add it to the sample;
          #' Else, we reject it and check next high prediction
          #'  variance location
          if(min.dist > delta){
                mylocations <- rbind(mylocations,
                                     dataframe[,c(coords.col)]
                                     [pred.v.sort[counter],])
                totalbatch <- totalbatch + 1
                counter <- counter + 1
          }
          else{
                rejected <- rbind(rejected,
                                  dataframe[,c(coords.col)]
                                  [pred.v.sort[counter],])
                counter <- counter + 1
          }
    }
    ind.coords <- NULL


    for(i in 1:nrow(mylocations)) {
            ind.sel <- which(mylocations[i,1]==
                               dataframe[,coords.col[1]] &
                                   mylocations[i,2]==
                               dataframe[,coords.col[2]])
            ind.coords <- c(ind.coords,ind.sel)
      }
      adaptive.sample <- dataframe[ind.coords,]
      return(adaptive.sample)
}
```

Complete Bibliography.

Aboal, J.R., Real, C., Fernández, J.A. and Carballeira, A. (2006) 'Mapping the results of extensive surveys: The case of atmospheric biomonitoring and terrestrial mosses', *Science of the Total Environment* 356 (1-3), pp. 256–274.

Afrane, Y.A., Zhou, G., Githeko, A.K. and Yan, G. (2013) 'Utility of Health Facility-based Malaria Data for Malaria Surveillance', *PLoS ONE* 8 (2), e54305.

Alemu, K., Worku, A. and Berhane, Y. (2013) 'Malaria infection has spatial, temporal, and spatiotemporal heterogeneity in unstable malaria transmission areas in northwest Ethiopia', *PLoS ONE* 8 (11), e79966.

Amexo, M., Tolhurst, R., Barnish, G. and Bates, I. (2004) 'Malaria misdiagnosis: Effects on the poor and vulnerable', *Lancet* 364 (9448), pp. 1896–1898.

Atkinson, P.M. (1991) 'Optimal ground-based sampling for remote sensing investigations: estimating the regional meant', *International Journal of Remote Sensing* 12 (3), pp. 559–567.

Atkinson, P.M., Webster, R. and Curran, P.J. (1992) 'Cokriging with Ground Based Radiometry', *Remote Sensing Environment* 41, pp. 45–60.

Baidjoe, A.Y., Stevenson, J., Knight, P., Stone, W., Stresman, G., Osoti, V., Makori, E., Owaga, C., Odongo, W., China, P., Shagari, S., Kariuki, S., Drakeley, C., Cox, J. and Bousema, T. (2016) 'Factors associated with high heterogeneity of malaria at fine spatial scale in the Western Kenyan highlands', *Malaria Journal* 15, p. 307.

Banerjee, S., Gelfand, A.E., Finley, A.O. and Sang, H. (2008) 'Gaussian predictive process models for large spatial data sets', *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 70 (4), pp. 825–848.

Bellhouse, D.R. and Herzberg, A.M. (1984) 'Equally spaced design points in polynomial regression: a comparison of systematic sampling methods with the optimal design of experiments.', *The Canadian Journal of Statistics* 12 (2), pp. 77–90.

Benhenni, K. and Cambanis, S. (1992) 'Sampling designs for estimating integrals of stochastic processes', *Annals of Statistics* 20, pp. 161–194.

Beron, K.J. and Vijverberg, W.P.M. (2004) 'Probit in a Spatial Context: A Monte Carlo Analysis', *Advances in Spatial Econometrics: Methodology, Tools and Applications.* Ed. by L. Anselin, S.J. Rey and R.J.G.M. Florax, 1st ed., Berlin Heidelberg: Springer, chap. 8, pp. 169–195.

Berrett, C. and Calder, C.A. (2012) 'Data augmentation strategies for the Bayesian spatial probit regression model', *Computational Statistics and Data Analysis* 56 (3), pp. 478–490.

Bhatt, S., Weiss, D.J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K.E., Moyes, C.L., Henry, A., Eckhoff, P.A., Wenger, E.A., Briet, O., Penny, M.A., Smith, T.A., Bennett, A., Yukich, J., Eisele, T.P., Griffin, J.T., Fergus, C.A., Lynch, M., Lindgren, F., Cohen, J.M., Murray, C.L.J., Smith, D.L., Hay, S.I., Cibulskis, R.E. and Gething, P.W. (2015) 'The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015', *Nature* 526 (7572), pp. 207–211.

Bijleveld, A.I., van Gils, J.A., van der Meer, J., Dekinga, A., Kraan, C., van der Veer, H.W. and Piersma, T. (2012) 'Designing a benthic monitoring programme with multiple conflicting objectives', *Methods in Ecology and Evolution* 3 (3), pp. 526–536.

Bogaert, P. and Russo, D. (1999) 'Optimal spatial sampling design for the estimation of the variogram based on a least squares approach', *Water Resources Research* 35 (4), pp. 1275–1289.

Bousema, T., Griffin, J.T., Sauerwein, R.W., Smith, D.L., Churcher, T.S., Takken, W., Ghani, A., Drakeley, C. and Gosling, R. (2012) 'Hitting Hotspots: Spatial Targeting of Malaria for Control and Elimination', *PLoS Medicine* 9 (1), e1001165.

Bousema, T., Stevenson, J., Baidjoe, A., Stresman, G., Griffin, J.T., Kleinschmidt, I., Remarque, E.J., Vulule, J., Bayoh, N., Laserson, K., Desai, M., Sauerwein, R., Drakeley, C. and Cox, J. (2013) 'The impact of hotspot-targeted interventions on malaria transmission: study protocol for a cluster-randomized controlled trial.', *Trials* 14 (1), p. 36.

Brown, J., Robertson, B.L. and McDonald, T. (2015) ‘Spatially Balanced Sampling: Application to Environmental Surveys’, *Procedia Environmental Sciences* 27, pp. 6–9.

Burton, D.C., Flannery, B., Onyango, B., Larson, C., Alaii, J., Zhang, X., Hamel, M.J., Breiman, R.F. and Feikin, D.R. (2011) ‘Healthcare-seeking behaviour for common infectious disease-related illnesses in rural Kenya: A community-based house-to-house survey’, *Journal of Health, Population and Nutrition* 29 (1), pp. 61–70.

Chilès, J.-P. and Delfiner, P. (2012) *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed., New Jersey: John Wiley & Sons, Inc.

Chilundo, B., Sundby, J. and Aanestad, M. (2004) ‘Analysing the quality of routine malaria data in Mozambique.’, *Malaria Journal* 3, p. 3.

Chipeta, M.G., Ngwira, B. and Kazembe, L.N. (2013) ‘Analysis of Schistosomiasis haematobium Infection Prevalence and Intensity in Chikhwawa, Malawi: An Application of a Two-Part Model’, *PLoS Neglected Tropical Diseases* 7 (3), e2131.

Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2016a) ‘Adaptive geostatistical design and analysis for prevalence surveys’, *Spatial Statistics* 15, pp. 70–84.

Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2016b) ‘Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure’, *Enviromentrics (in press)*, pp. 1–11.

Christensen, O.F. (2004) 'Monte Carlo Maximum Likelihood in Model-Based Geostatistics', *Journal of Computational and Graphical Statistics* 13 (3), pp. 702–718.

Cochran, W.G. (1977) *Sampling Techniques*, 3rd ed., New York: John Wiley & Sons, Ltd.

Cressie, N. (1991) *Statistics for Spatial Data*, New York: Wiley.

Curran, P.J. and Atkinson, P.M. (1998) 'Geostatistics and remote sensing', *Progress in Physical Geography* 22 (1), pp. 61–78.

De Beaudrap, P., Turyakira, E., White, L.J., Nabasumba, C., Tumwebaze, B., Muehlenbachs, A., Guérin, P.J., Boum, Y., McGready, R. and Piola, P. (2013) 'Impact of malaria during pregnancy on pregnancy outcomes in a Ugandan prospective cohort with intensive malaria screening and prompt treatment.', *Malaria Journal* 12, p. 139.

De Oliveira, V., Kedem, B. and Short, D.A. (1997) 'Bayesian prediction of transformed Gaussian random fields', *Journal of the American Statistical Association* 92 (440), pp. 1422–1433.

Diggle, P.J., Thomson, M.C., Christensen, O.F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J.H., Boussinesq, M. and Molyneux, D.H. (2007) 'Spatial modelling and the prediction of Loa loa risk: decision making under uncertainty', *Annals of Tropical Medicine and Parasitology* 101 (6), pp. 499–509.

Diggle, P.J. (2013) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns.* 3rd ed., Boca Raton: CRC Press.

Diggle, P.J. and Giorgi, E. (2015) 'Model-Based Geostatistics for Prevalence Mapping in Low-Resource Settings.', *Journal of the American Statistical Association (in press)*, pp. 1–42.

Diggle, P.J. and Giorgi, E. (2017) 'Preferential sampling of exposure levels', *unpublished.*

Diggle, P.J. and Lophaven, S. (2006) 'Bayesian geostatistical design', *Scandinavian Journal of Statistics* 33 (1), pp. 53–64.

Diggle, P.J., Menezes, R. and Su, T.-L. (2010) 'Geostatistical inference under preferential sampling', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59 (2), pp. 191–232.

Diggle, P.J. and Ribeiro, J.P. (2002) *Bayesian inference in Gaussian model-based geostatistics.* Tech. rep., Lancaster University.

Diggle, P.J. and Ribeiro, J.P. (2007) *Model-based Geostatistics*, New York: Springer.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998) 'Model-based geostatistics (with discussion)', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47 (3), pp. 299–350.

Fanshawe, T.R. and Diggle, P.J. (2012) 'Adaptive Sampling Design for Spatio-Temporal Prediction', *Spatio-temporal design: Advances in efficient data acquisition*, ed. by J. Mateu and W.G. Müller, 1st ed., Chichester, UK: John Wiley & Sons, Ltd, chap. 11, pp. 249–268.

Fernández, J.A., Real, C., Couto, J.A., Aboal, J.R. and Carballeira, A. (2005) 'The effect of sampling design on extensive bryomonitoring surveys of air pollution', *Science of the Total Environment* 337 (1-3), pp. 11–21.

Fernández, J.A., Rey, A. and Carballeira, A. (2000) 'An extended study of heavy metal disposition in Galicia (NW Spain) based on moss analysis', *The Science of the Total Environment* 254, pp. 31–44.

Ferreira, G.d.S. and Gamerman, D. (2015) 'Optimal Design in Geostatistics under Preferential Sampling', *Bayesian Analysis* 10 (3), pp. 711–735.

Flury, B.D. (1990) 'Acceptance-Rejection Sampling Made Easy.', *Society for Industrial and Applied Mathematics* 32 (3), pp. 474–476.

Gao, B.-B., Wang, J.-F., Fan, H.-M., Xu, K., Hu, M.-G. and Chen, Z.-Y. (2015) 'A stratified optimization method for a multivariate marine environmental monitoring network in the Yangtze River estuary and its adjacent sea', *International Journal of Geographical Information Science* 29 (8), pp. 1332–1349.

Gelfand, A.E., Banerjee, S. and Finley, A.O. (2012) 'Spatial Design for Knot Selection in Knot-Based Dimension Reduction Models', *Spatio-temporal design: Advances in efficient data acquisition*, ed. by J. Mateu and W.G. Müller, 1st ed., Chichester, UK: John Wiley & Sons, Ltd, chap. 7, pp. 142–169.

Gelfand, A.E., Sahu, S.K. and Holland, D.M. (2012) ‘On the Effect of Preferential Sampling in Spatial Prediction.’, *Environmetrics* 23 (7), pp. 565–578.

Gething, P.W., Elyazar, I.R.F., Moyes, C.L., Smith, D.L., Battle, K.E., Guerra, C.A., Patil, A.P., Tatem, A.J., Howes, R.E., Myers, M.F., George, D.B., Horby, P., Wertheim, H.F.L., Price, R.N., Mueller, I., Baird, J.K. and Hay, S.I. (2012) ‘A long neglected world malaria map: Plasmodium vivax endemicity in 2010’, *PLoS Neglected Tropical Diseases* 6 (9), e1814.

Giorgi, E. and Diggle, P.J. (2015) ‘PrevMap : an R Package for Prevalence Mapping’, *Journal of Statistical Software (to appear)*, pp. 1–27.

Giorgi, E., Sesay, S.S.S., Terlouw, D.J. and Diggle, P.J. (2015) ‘Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models’, *Journal of the Royal Statistical Society. Series A: Statistics in Society* 178 (2), pp. 445–464.

Gosoniu, L., Msengwa, A., Lengeler, C. and Vounatsou, P. (2012) ‘Spatially explicit Burden estimates of malaria in Tanzania: Bayesian geostatistical modeling of the malaria indicator survey data’, *PLoS ONE* 7 (5), e23966.

Grafström, A., Lundström, N. and Schelin, L. (2012) ‘Spatially Balanced Sampling through the Pivotal Method’, *Biometrics* 68 (2), pp. 514–520.

Grimes, J.E.T. and Templeton, M.R. (2016) ‘Geostatistical modelling of schistosomiasis prevalence’, *The Lancet Infectious Diseases* 15 (8), pp. 869–870.

Guerra, C.A., Hay, S., Lucioparedes, L.S., Gikandi, P.W., Tatem, A.J., Noor, A.M. and Snow, R.W. (2007) 'Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project', *Malaria Journal* 6, p. 17.

Guo, D., Wu, Y., Shamai, S. and Verdú, S. (2011) 'Estimation in Gaussian noise: Properties of the minimum mean-square error', *IEEE Transactions on Information Theory* 57 (4), pp. 2371–2385.

Guttorp, P. and Sampson, P.D. (1994) 'Methods for estimating heterogeneous spatial covariance functions with environmental applications', *Handbook of Statistics* 12 (236), pp. 661–689.

Guttorp, P. and Sampson, P.D. (2010) 'Discussion of Geostatistical inference under preferential sampling by Diggle, P.J., Menezes, R. and Su, T.', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59 (2), pp. 191–232.

Handcock, M.S. and Stein, M.L. (1993) 'A Bayesian Analysis of Kriging', *Technometrics* 35 (4), pp. 403–410.

Hay, S., Guerra, C.A., Gething, P.W., Patil, A.P., Tatem A.J., Noor, A.M., Kabaria, C.W., Manh, B.H., Elyazar, I.R.F., Brooker, S., Smith, D.L., Moyeed, R.A. and Snow, R.W. (2009) 'A world malaria map: Plasmodium falciparum endemicity in 2007', *PLoS Medicine* 6 (3), e1000048.

Hay, S., Guerra, C.A., Tatem, A.J., Noor, A.M. and Snow, R.W. (2004) 'The global distribution and population at risk of malaria: past, present, and future', *The Lancet Infectious Diseases* 4 (6), pp. 327–336.

Higdon, D. (1998) ‘A process-convolution approach to modeling temperatures in the North Atlantic Ocean (with discussion)’, *Environmental and Ecological Statistics* 5 (2), pp. 173–190.

Higdon, D. (2002) ‘Space and Space-Time Modeling using Process Convolutions’, *Quantitative Methods for Current Environmental Issues*, ed. by C.W. Anderson, V. Barnett, P.C. Chatwin and A.H. EI-Shaarawi, 1st ed., London: Springer, chap. 2, pp. 37–56.

Hotez, P.J. and Kamath, A. (2009) ‘Neglected tropical diseases in sub-Saharan Africa: Review of their prevalence, distribution, and disease burden’, *PLoS Neglected Tropical Diseases* 3 (8), pp. 2–11.

Hu, M.G. and Wang, J.F. (2011) ‘A spatial sampling optimization package using MSN theory’, *Environmental Modelling and Software* 26 (4), pp. 546–548.

Huynh, B.T., Fievet, N., Gbaguidi, G., Dechavanne, S., Borgella, S., Guézo-Mévo, B., Massougbodji, A., Tuikue Ndam, N., Deloron, P. and Cot, M. (2011) ‘Influence of the timing of malaria infection during pregnancy on birth weight and on maternal anemia in Benin’, *American Journal of Tropical Medicine and Hygiene* 85 (2), pp. 214–220.

Isaaks, E.H. and Srivastava, R.M. (1989) *An Introduction to applied geostatistics*, New York: Oxford University Press.

Jardim, E. and Ribeiro, P.J. (2007) ‘Geostatistical assessment of sampling designs for Portuguese bottom trawl surveys’, *Fisheries Research* 85 (3), pp. 239–247.

Journel, A.G. and Huijbregts, C.J. (1978) *Mining Geostatistics*, London: Academic Press.

Kalilani-Phiri, L., Thesing, P.C., Nyirenda, O.M., Mawindo, P., Madanitsa, M., Membe, G., Wylie, B., Masonbrink, A., Makwakwa, K., Kamiza, S., Muehlenbachs, A., Taylor, T.E. and Laufer, M.K. (2013) 'Timing of Malaria Infection during Pregnancy Has Characteristic Maternal, Infant and Placental Outcomes', *PLoS ONE* 8 (9), e74643.

Kazembe, L.N., Kleinschmidt, I. and Sharp, B.L. (2006) 'Patterns of malaria-related hospital admissions and mortality among Malawian children: an example of spatial modelling of hospital register data', *Malaria Journal* 5, p. 93.

Kazembe, L., Kleinschmidt, I., Holtz, T. and Sharp, B. (2006) 'Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data', *International Journal of Health Geographics* 5, p. 41.

Kitanidis, P.K. (1987) 'Parametric estimation of covariances of regionalized variables', *Journal of the American Water Resources Association* 23 (4), pp. 557–567.

Kleinschmidt, I. (2001) 'Spatial statistical analysis, modelling and mapping of malaria in Africa', PhD thesis, University of Basel.

Kondo, M.C., Bream, K.D.W., Barg, F.K. and Branas, C.C. (2014) 'A random spatial sampling method in a rural developing nation.', *BMC Public Health* 14, p. 338.

Krige, D.G. (1951) 'A Statistical Approach to Some Mine Valuation and Allied Problems on the Witwatersrand.', *Journal of the Chemical, Metallurgical and Mining Society of South Africa* 52, pp. 119–139.

Lark, R.M. (2002) 'Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood', *Geoderma* 105 (1-2), pp. 49–80.

Machault, V., Vignolles, C., Borchi, F., Vounatsou, P., Pages, F., Briolant, S., Lacaux, J.P. and Rogier, C. (2011) 'The use of remotely sensed environmental data in the study of malaria', *Geospatial Health* 5 (2), pp. 151–168.

Marchant, B.P., Lark, R.M. and Wheeler, H.C. (2005) *Developing methods to improve sampling efficiency for automated soil mapping*, tech. rep., Home-Grown Cereals Authority.

Mardia, K.V. and Marshall, R.J. (1984) 'Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression', *Biometrika* 71 (1), pp. 135–146.

Matérn, B. (1960) *Spatial Variation*, tech. rep., Stockholm: Statens Skogsforsningsinstitut.

Matérn, B. (1986) *Spatial Variation*, 2nd ed., Berlin: Springer.

McBratney, A.B., Webster, R. and Burgess, T.M. (1981) 'The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalized Variables–I: Theory and method.', *Computers & Geosciences* 7 (4), pp. 331–334.

McBratney, A.B. and Webster, R. (1981) 'The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalised Variables–II: Program and Examples.', *Computers & Geosciences* 7 (4), pp. 335–365.

MERIT Initiative (2012) *Meningitis Environmental Risk Information Technologies*, http://merit.hc-foundation.org/ProgramActivities2012.html.

Müller, W.G. (2005) 'A comparison of spatial design methods for correlated observations', *Environmetrics* 16, pp. 495–505.

Müller, W.G. (2007) *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*, 3rd ed., Berlin: Springer-Verlag.

Müller, W.G., Pronzato, L., Rendas, J. and Waldl, H. (2015) 'Efficient prediction designs for random fields', *Applied Stochastic Models in Business and Industry* 31 (2), pp. 178–194.

Müller, W.G. and Zimmerman, D.L. (1999) 'Optimal designs for Variogram estimation', *Enviromentrics* 10, pp. 23–37.

Mzilahowa, T., Hastings, I.M., Molyneux, M.E. and McCall, P.J. (2012) 'Entomological indices of malaria transmission in Chikhwawa district, Southern Malawi.', *Malaria Journal* 11, p. 380.

Nansseu; J.R.N., Noubiap; J.J.N., Ndoula; S.T., Zeh; A.F.M. and Monamele, C.G. (2013) 'What Is the Best Strategy for the Prevention of Transfusion-Transmitted Malaria in Sub-Saharan African Countries Where Malaria Is Endemic?', *Malaria Journal* 12, p. 465.

Noor, A.M., Kinyoki, D.K., Mundia, C.W., Kabaria, C.W., Mutua, J.W., Alegana, V.A., Fall, I.S. and Snow, R.W. (2014) 'The changing risk of Plasmodium falciparum malaria infection in Africa: 2000-10: A spatial and temporal analysis of transmission intensity', *The Lancet* 383 (9930), pp. 1739–1747.

Nowak, W. (2010) 'Measures of parameter uncertainty in geostatistical estimation and geostatistical optimal design', *Mathematical Geosciences* 42 (2), pp. 199–221.

Nychka, D. and Saltzman, N. (1998) 'Design of Air-Quality Monitoring Networks', *Case Studies in Environmental Statistics SE - 4*, Lecture Notes in Statistics 132, ed. by D. Nychka, W. Piegorsch and L. Cox, pp. 51–76.

Olea, R.A. (1984) 'Sampling design optimization for spatial functions', *Journal of the International Association for Mathematical Geology* 16 (4), pp. 369–392.

Ouédraogo, A., Tiono, A.B., Diarra, A., Sanon, S., Yaro, J.B., Ouedraogo, E., Bougouma, E.C., Soulama, I., Gansané, A., Ouedraogo, A., Konate, A.T., Nebie, I., Watson, N.L., Sanza, M., Dube, T.J.T. and Sirima, S.B. (2013) 'Malaria Morbidity in High and Seasonal Malaria Transmission Area of Burkina Faso', *PLoS ONE* 8 (1), e50036.

Paciorek, C.J. and Schervish, M.J. (2006) 'Spatial Modelling Using a New Class of Nonstationary Covariance Functions.', *Environmetrics.* 17 (5), pp. 483–506.

Paciorek, C.J. (2003) 'Nonstationary Gaussian Processes for Regression and Spatial Modelling', PhD thesis, Carnegie Mellon University.

Pati, D., Reich, B.J. and Dunson, D.B. (2011) ‘Bayesian geostatistical modelling with informative sampling locations’, *Biometrika* 98, pp. 35–48.

Patil, A.P., Gething, P.W., Piel, F.B. and Hay, S.I. (2011) ‘Bayesian geostatistics in health cartography: the perspective of malaria’, *Trends in Parasitology* 27 (6), pp. 246–253.

Pettitt, A.N. and McBratney, A.B. (1993) ‘Sampling Designs for Estimating Spatial Variance Components’, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 42 (1), pp. 185–209.

Phiri, B.B.W., Ngwira, B. and Kazembe, L.N. (2016) ‘Analysing risk factors of co-occurrence of schistosomiasis haematobium and hookworm using bivariate regression models: Case study of Chikwawa, Malawi’, *Parasite Epidemiology and Control* 1 (2), pp. 149–158.

Pilz, J. and Spöck, G. (2006) ‘Spatial sampling design for prediction taking account of uncertain covariance structure’, *7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.* Lisbon, Portugal, pp. 109–118.

Plagemann, C., Kersting, K. and Burgard, W. (2008) ‘Nonstationary Gaussian Process Regression using Point Estimates of Local Smoothness’, *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II*, ed. by W. Daelemans, B. Goethals and K. Morik, 1st ed., Berlin Heidelberg: Springer-Verlag, chap. 13, pp. 204–219.

Price, W.L. (1976) ‘A controlled random search procedure for global optimisation’, *The Computer Journal* 20 (4), pp. 367–370.

Price, W.L. (1983) ‘Global optimization by controlled random search’, *Journal of Optimization Theory and Applications* 40 (3), pp. 333–348.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*, https://www.R-project.org/, Vienna, Austria.

Reid, H., Haque, U., Clements, A.C.A., Tatem, A.J., Vallely, A., Ahmed, S.M., Islam, A. and Haque, R. (2010) ‘Mapping malaria risk in Bangladesh using Bayesian geostatistical models’, *American Journal of Tropical Medicine and Hygiene* 83 (4), pp. 861–867.

Ritter, K. (1996) ‘Asymptotic optimality of regular sequence designs’, *The Annals of Statistics* 24 (5), pp. 2081–2096.

Roca-Feltrer, A., Lalloo, D.G., Phiri, K. and Terlouw, D.J. (2012) ‘Short Report : Rolling Malaria Indicator Surveys (rMIS): a potential district-level malaria monitoring and evaluation (M & E) tool for program managers.’, *American Journal of Tropical Medicine and Hygiene* 86 (1), pp. 96–98.

Roll Back Malaria Partnership (2016) *MIS Toolkit*, http://www.malariasurveys.org/.

Rowe, A.K. (2009) ‘Potential of integrated continuous surveys and quality management to support monitoring, evaluation, and the scale-up of health interventions in developing countries’, *American Journal of Tropical Medicine and Hygiene* 80 (6), pp. 971–979.

Rowe, A.K., Kachur, S.P., Yoon, S.S., Lynch, M., Slutsker, L. and Steketee, R.W. (2009) 'Caution is required when using health facility-based data to evaluate the health impact of malaria control efforts in Africa', *Malaria Journal* 8, p. 209.

Royle, J. and Nychka, D. (1998) 'An algorithm for the construction of spatial coverage designs with implementation in SPLUS', *Computers & Geosciences* 24 (5), pp. 479–488.

Rubin, D.B. (1976) 'Inference and Missing Data', *Biometrika* 63 (3), pp. 581–592.

Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications.* London: Chapman & Hall/CRC.

Russo, D. (1984) 'Design of an Optimal Sampling Network for Estimating the Variogram', *Soil Science Society of America Journal* 48 (4), pp. 708–716.

Schabenberger, O. and Gotaway, C.A. (2005) *Statistical Methods for Spatial Data Analysis*, Boca Raton: Chapman & Hall/CRC.

Schur, N., Hürlimann, E., Garba, A., Traoré, M.S., Ndir, O., Ratard, R.C., Tchuem Tchuenté, L.-A., Kristensen, T.K., Utzinger, J. and Vounatsou, P. (2011) 'Geostatistical Model-Based Estimates of Schistosomiasis Prevalence among Individuals Aged less than 20 Years in West Africa', *PLoS Neglected Tropical Diseases* 5 (6), e1194.

Shaddick, G. and Zidek, J.V. (2014) 'A case study in preferential sampling: Long term monitoring of air pollution in the UK', *Spatial Statistics* 9, pp. 51–65.

Snow, R.W., Craig, M., Deichmann, U. and Marsh, K. (1999) ‘Estimating mortality, morbidity and disability due to malaria among Africa’s non-pregnant population’, *Bulletin of the World Health Organization* 77 (8), pp. 624–640.

Stanton, M.C., Agier, L., Taylor, B.M. and Diggle, P.J. (2014) ‘Towards real-time spatiotemporal prediction of district level meningitis incidence in sub-Saharan Africa’, *Journal of the Royal Statistical Society. Series A: Statistics in Society* 177 (3), pp. 661–678.

Stanton, M.C. and Diggle, P.J. (2013) ‘Geostatistical analysis of binomial data: generalised linear or transformed Gaussian modelling?’, *Environmetrics* 24 (3), pp. 158–171.

Staub, C.G., Binford, M.W. and Stevens, F.R. (2013) ‘Elephant herbivory in Majete Wildlife Reserve, Malawi’, *African Journal of Ecology* 51, pp. 536–543.

Stein, M.L. (1999) *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer.

Steketee, R.W., Nahlen, B.L., Parise, M.E. and Menendez, C. (2001) ‘The burden of malaria in pregnancy in malaria-endemic areas’, *American Journal of Tropical Medicine and Hygiene* 64 ((1, 2) Supplementary), pp. 28–35.

Stevens, D.L. and Olsen, A.R. (1999) ‘Spatially Restricted Surveys over Time for Aquatic Resources.’, *International Biometric Society* 4 (4), pp. 415–428.

Stevens, D.L. and Olsen, A.R. (2003) ‘Variance estimation for spatially balanced samples of environmental resources’, *Environmetrics* 14 (6), pp. 593–610.

Stevens, D.L. and Olsen, A.R. (2004) ‘Spatially Balanced Sampling of Natural Resources’, *Journal of the American Statistical Association* 99 (465), pp. 262–278.

Su, Y.S.Y. and Cambanis, S. (1993) ‘Sampling Designs for Estimation of a Random Process’, *Stochastic Processes and their Applications* 46, pp. 47–89.

Thompson, S.K. and Collins, L.M. (2002) ‘Adaptive sampling in research on risk-related behaviors.’, *Drug and Alcohol Dependence* 68, S57–S67.

Van der Hoek, W., Konradsen, F., Amerasinghe, P.H., Perera, D., Piyaratne, M.K. and Amerasinghe, F.P. (2003) ‘Towards a risk map of malaria for Sri Lanka: The importance of house location relative to vector breeding sites’, *International Journal of Epidemiology* 32 (2), pp. 280–285.

Van Groenigen, J.W., Siderius, W. and Stein, A. (1999) ‘Constrained optimisation of soil sampling for minimisation of the kriging variance’, *Geoderma* 87 (3-4), pp. 239–259.

Van Groenigen, J.W. and Stein, A. (1998) ‘Constrained Optimization of Spatial Sampling using Continuous Simulated Annealing’, *Journal of Environmental Quality* 27 (5), pp. 1078–1086.

Vyas, S. and Kumaranayake, L. (2006) ‘Constructing socio-economic status indices: How to use principal components analysis’, *Health Policy and Planning* 21 (6), pp. 459–468.

Walker, P.G.T., Griffin, J.T., Ferguson, N.M. and Ghani, A.C. (2016) 'Estimating the most efficient allocation of interventions to achieve reductions in Plasmodium falciparum malaria burden and transmission in Africa: a modelling study', *The Lancet Global Health* 4 (7), e474–e484.

Wang, J., Haining, R. and Cao, Z. (2010) 'Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning', *International Journal of Geographical Information Science* 24 (4), pp. 523–543.

Warrick, A.W. and Myers, D.E. (1987) 'Optimization of sampling locations for variogram calculations', *Water Resources Research* 23 (3), pp. 496–500.

Wienand, J. (2013) 'Woody vegetation change and elephant water point use in Majete Wildlife Reserve: implications for water management strategies', PhD thesis, Stellenbosch University.

Woolhouse, M.E., Dye, C., Etard, J.F., Smith, T., Charlwood, J.D., Garnett, G.P., Hagan, P., Hii, J.L., Ndhlovu, P.D., Quinnell, R.J., Watts, C.H., Chandiwana, S.K. and Anderson, R.M. (1997) 'Heterogeneities in the transmission of infectious agents: implications for the design of control programs.', *Proceedings of the National Academy of Sciences of the United States of America* 94 (1), pp. 338–342.

World Health Organisation (2012) *World Malaria Report 2012*, tech. rep., Geneva: World Health Organisation.

World Health Organisation (2013) *World Malaria Report 2013*, tech. rep., Geneva: World Health Organisation.

World Health Organisation (2015a) *Global technical strategy for malaria 2016-2030*, tech. rep., Geneva: World Health Organisation.

World Health Organisation (2015b) *World Malaria Report 2015*, tech. rep., Geneva: World Health Organisation.

World Health Organisation (2016) *World Malaria Report 2016*, tech. rep., Geneva: World Health Organisation.

Yfantis, E.A., Flatman, G.T. and Behar, J.V. (1987) 'Efficiency of Kriging Estimation for Square , Triangular , and Hexagonal Grids', *Mathematical Geology* 19 (3), pp. 183–205.

Zhu, Z. (2002) 'Optimal Sampling Design and Parameter Estimation of Gaussian Random Fields', PhD thesis, University of Chicago.

Zhu, Z. and Stein, M.L. (2006) 'Spatial sampling design for prediction with estimated parameters', *Journal of Agricultural, Biological, and Environmental Statistics* 11 (1), pp. 24–44.

Zidek, J.V., Shaddick, G. and Taylor, C.G. (2014) 'Reducing estimation bias in adaptively changing monitoring networks with preferential site selection', *The Annals of Applied Statistics* 8 (3), pp. 1640–1670.

Zimmerman, D.L. (2006) 'Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction', *Environmetrics* 17 (6), pp. 635–652.

Zouré, H.G.M., Noma, M., Tekle, A.H., Amazigo, U.V., Diggle, P.J., Giorgi, E. and Remme, J.H.F. (2014) 'The geographic distribution of onchocerciasis in the

20 participating countries of the African Programme for Onchocerciasis Control: (2) pre-control endemicity levels and estimated number infected', *Parasites & Vectors* 7, p. 326.