Adaptive Geostatistical Design and Analysis for Prevalence Surveys

Michael G. Chipeta[a,b,c], Dianne J. Terlouw[a,c,d], Kamija S. Phiri[a] and Peter J. Diggle[b]

[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

December 11, 2015

Abstract
Non-adaptive geostatistical designs (NAGDs) offer standard ways of collecting and analysing exposure and outcome data in which sampling locations are fixed in advance of any data collection. In contrast, adaptive geostatistical designs (AGDs) allow collection of exposure and outcome 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

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 $S = \{S(x) : x \in D \subset \mathbb{R}^2\}$, where D is a geographical region of interest, using data in the form of measurements y_i at locations $x_i \in D$. 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 S and the observed data $\{(y_i, x_i) : i = 1, ..., n\}$; *prediction* of the unobserved realisation of $S(x)$ throughout 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 Jr. (2007); 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, Lalloo, Phiri, and Terlouw, 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 D it becomes appropriate to focus progressively on sub-regions of D where precise prediction is needed to inform public health action, for example to prioritise sub-regions for early intervention.

In Section 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 describes our proposed class of adaptive designs for efficient prediction. Section 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 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 6 is a concluding discussion.

2 Methodological framework

2.1 Geostatistical models for prevalence data

The standard geostatistical model for prevalence data can be formulated 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 $D \subset \mathbb{R}^2$, 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

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

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 approach. Define the *empirical logit transform*,

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

and assume that

$$Y_i^* = d(x_i)'\beta + S(x_i) + Z_i, \tag{2}$$

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 $S(x)$,

$$\rho(u; \phi; \kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^\kappa K_\kappa(u/\phi), \tag{3}$$

where $\phi > 0$ is a scale parameter that controls the rate at which correlation decays with increasing distance, $K_\kappa(\cdot)$ 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 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 5 we use the standard model (1).

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 data-set $(\mathcal{X}_0, Y_0)$, Y_1 the resulting additional measurement data, and so on. After k additions, the complete data-set 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 data-set is

$$[\mathcal{X}, Y] = \int_S [\mathcal{X}, Y, S] dS. \tag{4}$$

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

$$[\mathcal{X}, Y, S] = [S, \mathcal{X}_0, Y_0, \mathcal{X}_1, Y_1] = [S][\mathcal{X}_0|S][Y_0|\mathcal{X}_0, S][\mathcal{X}_1|Y_0, \mathcal{X}_0, S][Y_1|\mathcal{X}_1, Y_0, \mathcal{X}_0, S]. \tag{5}$$

On the right-hand side of (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 (4) and (5) that

$$[\mathcal{X}, Y] = [\mathcal{X}_0][\mathcal{X}_1|\mathcal{X}_0, Y_0] \times \int_S [Y_0|\mathcal{X}_0, S][Y_1|\mathcal{X}_1, Y_0, \mathcal{X}_0, S][S] dS \tag{6}$$

The first term of the right hand side of (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

$$[\mathcal{X}, Y] = [\mathcal{X}|Y_0] \times [Y|\mathcal{X}]. \tag{7}$$

Equation (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 (7) extends to $k > 1$ with essentially only notational changes.

3 An adaptive design strategy

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 $S = \{S(x) : x \in D\}$ denote the realisation of the process $S(x)$ over 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 $T = \mathcal{T}(S)$, called the *predictive target*, represent the property of 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 S and Y, whereas in the expression for $\hat{T}$ it is with respect to S holding Y fixed at its observed value.

One obvious predictive target is $S(x)$ for arbitrary location $x \in D$. 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 c is a policy-relevant threshold; see, for example, Figure 3 of Zouré, Noma, Tekle, Amazigo, Diggle, Giorgi, and Remme (2014). Spatially neutral versions of these targets can be defined by integration over D, hence

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

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.

3.2 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 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. $T = \int_D S(x)dx$, and model parameters are known; see, for example, Matérn (1960, Chapter 5); Bellhouse and Herzberg (1984); Fernández et al. (2005); Marchant et al. (2005); Müller (2007); Diggle and Ribeiro Jr. (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); Müller (2005) and Ritter (1996). 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 $\gamma(u) = \frac{1}{2}\mathrm{Var}\{S(x) - S(x')\}$ 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 data-set 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),

$$APV = \int_D \mathrm{Var}\{S(x)|\, Y\}\mathrm{d}x, \tag{8}$$

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

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 $x_i \in D$. 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 $x \in D$ may be a potential sampling location, in which case we take $\mathcal{X}^*$ to be a finely spaced regular lattice to cover D.

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

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.

3.4 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.

3.5 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 (2),

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

where $r = (r_1, \ldots, r_{n_0})$ with $V = \sigma^2 R + \tau^2 I$, R is the n by n matrix with elements $r_{ij} = \rho(||x_i - x_j||)$ and I is the identity matrix (Diggle and Ribeiro, 2007, p137).

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 $A_j = \emptyset$.

4 Simulation study

We conducted a simulation study of our proposed ADG 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 (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 3.5 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.

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 each design by its spatially averaged prediction variance, i.e. APV as defined at (8), in turn averaged over 100 replicate simulations. When the initial sample size is $n_0 = 30$, Figure 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 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.

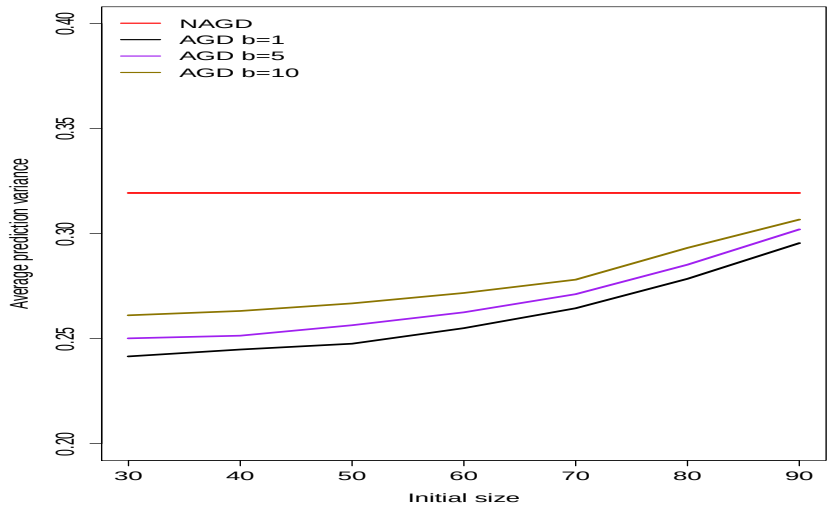


Figure 1: Non-adaptive (NAGD) vs **minimum distance** batch adaptive (AGD) sampling, with $\delta = 0.03$ and AGD batch sizes $b = 1, 5$ and 10; Initial size (n_0) ranges from 30 to 90. See text for details of the simulation model.

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 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 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.

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 2 shows the sampled household locations (red dots) in their respective villages, with black dots indicating all households in each village. Data collected include the outcome of a malaria rapid diagnostic test, age and gender of each individual and socioeconomic 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/`.

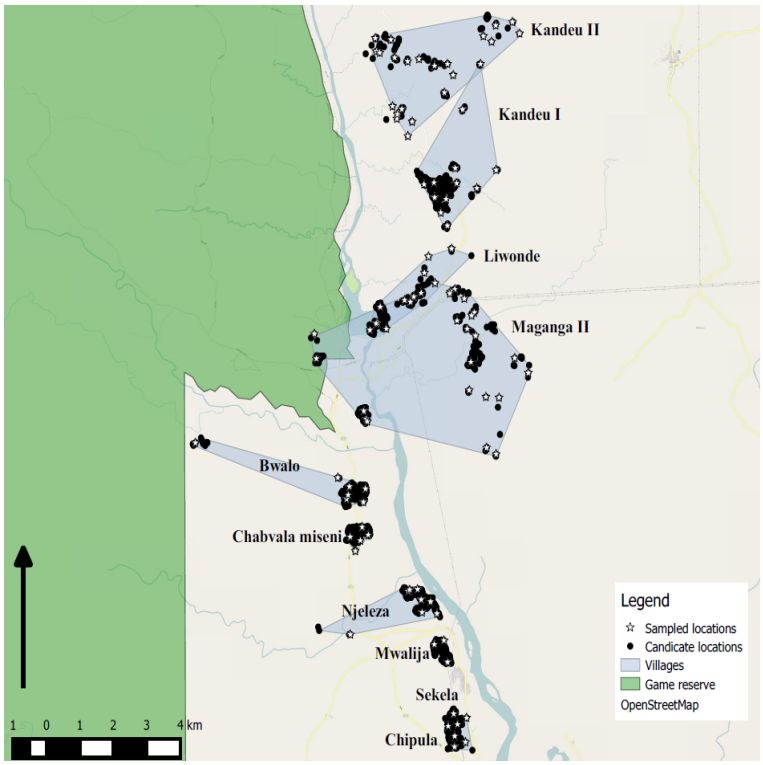


Figure 2: Households within the Majete wildlife reserve perimeter in focal area B (black dots) and sampled household locations (white dots) shown in their respective villages.

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 inclusion criteria (see Section 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 than 0.05. This resulted in the set of covariates shown in Table 1, with terms for elevation, NDVI and the interaction between the two. We then fitted the binomial logistic model (1) to obtain the Monte Carlo maximum likelihood estimates of the parameters and associated 95 % confidence intervals, as also shown in Table 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.

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. $\times$ NDVI	-0.0405	(-0.0588, -0.0223)
σ^2	0.6339	(0.4438, 0.9055)
ϕ	0.2293	(0.1042, 0.5049)

Table 1: Monte Carlo maximum likelihood estimates and 95 % confidence intervals for the model fitted to the Majete malaria data.

From Table 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 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.

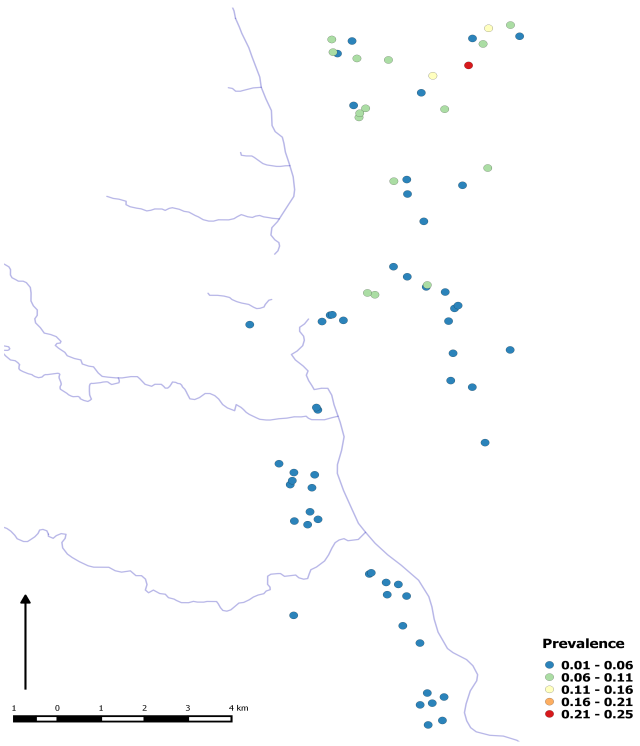


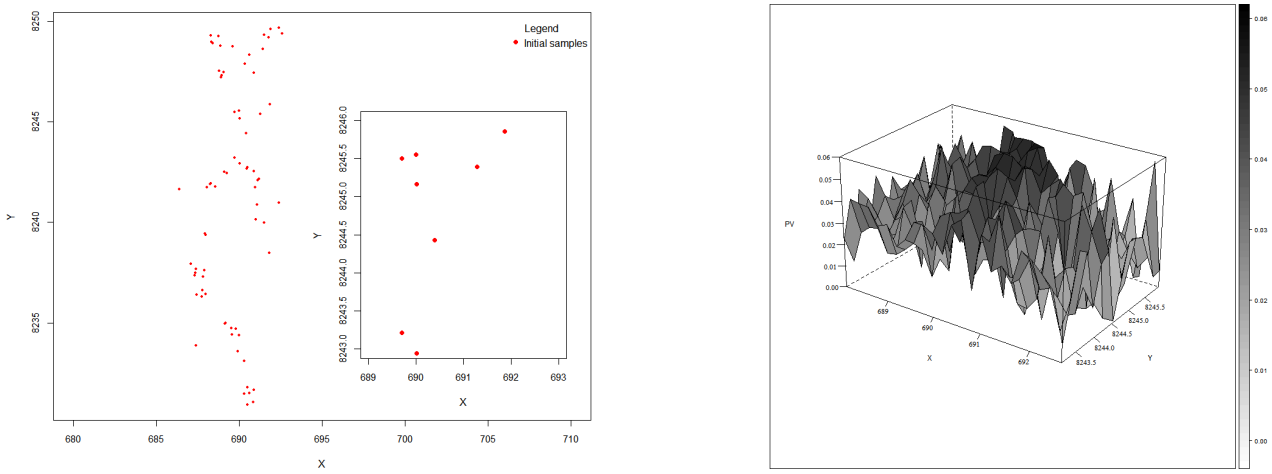
Figure 3: Predictions of $d(x)'\beta + S(x)$ at observed locations in focal area B. The blue lines shows Shire and Matope rivers.

5.3 Adaptive sampling in practice

We now use the *minimum distance batch adaptive sampling* approach explained in Section 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 4a. 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 3.5 above for comparison sake. The black dots in Figure 5a show 90 new locations determined using random sampling. The blue dots in Figure 6a 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

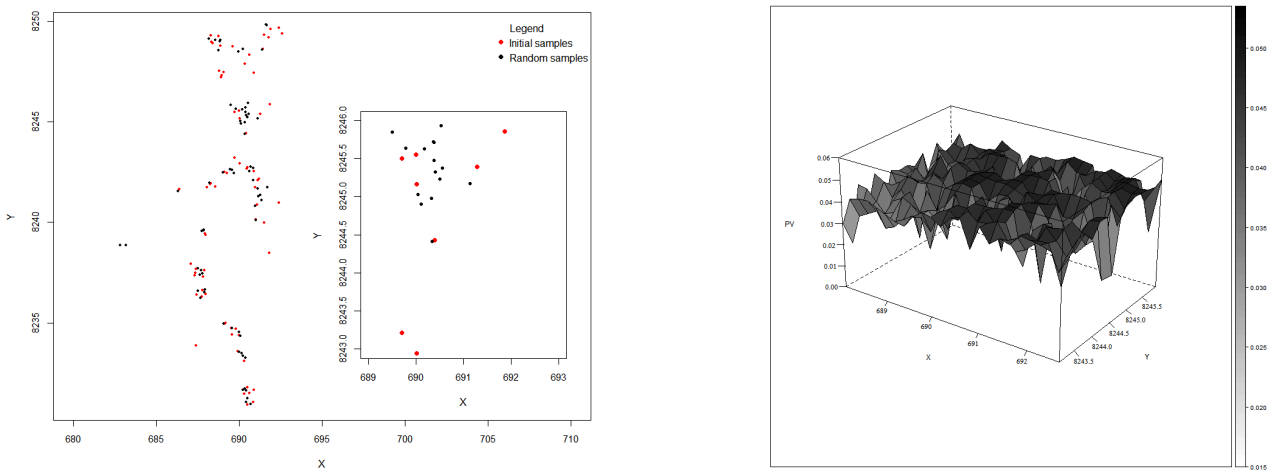
parameters are not known, which is the case in most geostatistical problems.

In Figure 4 we show the initial sample locations and the prediction variance surface for the inset sub-region. Figures 5 and 6 give the same information after addition of 90 randomly and adaptively selected locations, 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.



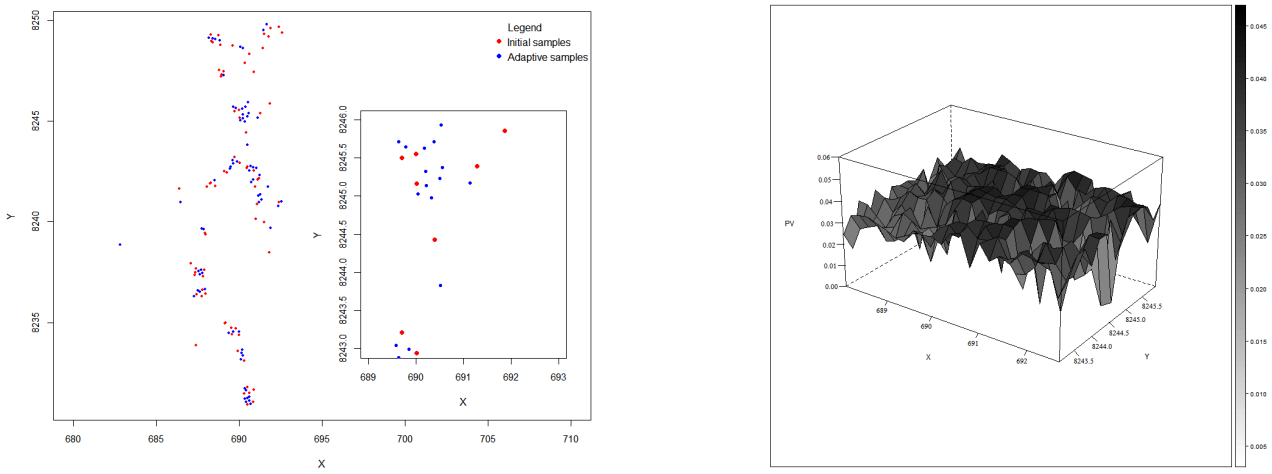
(a) (b)

Figure 4: (a) Initial inhibitory sampling design locations (red dots) in focal area B. Inset shows a subset of locations. (b) Prediction variance surface for the inset sub-region from 4a.



(a) (b)

Figure 5: (a) Initial inhibitory sampling design locations (red dots) and random sampling design locations (black dots) in focal area B. Inset shows a subset of locations. (b) Prediction variance surface for the inset sub-region from 5a.



(a) (b)

Figure 6: (a) Initial inhibitory sampling design locations (red dots) and adaptive sampling design locations (blue dots) in focal area B. Inset shows a subset of locations. (b) Prediction variance surface for the inset sub-region from 6a.

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 framework for survey-based disease mapping in poor resource settings. The particular design strategy described in Section 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 relative 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 (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.

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 policymaking 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-stratifcation of the study region into sub-regions can bring substantial gains in efficiency (Wang, Haining, and Cao, 2010; Hu and Wang, 2011; Gao, Wang, Fan, Xu, Hu, and Chen, 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 et al. (2004), Guerra et al. (2007), Hay et al. (2009) and Gething et al. (2012).

Acknowledgement

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, Dr 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. Eqaully spaced design points in polynomial regression: a comparison of systematic sampling methods with the optimal design of experiments. *Can. J. Stat.*, 12(2):77–90, 1984.

Benhenni, K. and Cambanis, S. Sampling designs for estimating integrals of stochastic processes. *Ann. Stat.*, 20:161–194, 1992.

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

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

Diggle, P. J. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. C&H/CRC Monographs on Statistics & Applied Probability. CRC Press, Boca Raton, 3 edition, jul 2013. ISBN 978 - 1 - 4665 - 6023 - 9. doi: doi:10.1201/b15326-1. URL `http://dx.doi.org/10.1201/b15326-1`.

Diggle, P. J. and Lophaven, S. Bayesian geostatistical design. *Scand. J. Stat.*, 33(1):53–64, 2006.

Diggle, P. J. and Ribeiro Jr., P. *Model-based Geostatistics*. Springer, New York, 2007. ISBN 9780387329079.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. Model-based geostatistics (with discussion). *J. R. Stat. Soc. Ser. C (Applied Stat.*, 47(3):299–350, jan 1998. ISSN 00359254. doi: 10.1111/1467-9876.00113.

Diggle, P. J., Menezes, R., and Su, T.-l. Geostatistical inference under preferential sampling. *J. R. Stat. Soc. Ser. C (Applied Stat.*, 59(2):191–232, mar 2010. ISSN 00359254. doi: 10.1111/j.1467-9876.2009.00701.x.

Fernández, J., Real, C., Couto, J., Aboal, J., and Carballeira, A. The effect of sampling design on extensive bryomonitoring surveys of air pollution. *Sci. Total Environ.*, 337(1-3):11–21, 2005. ISSN 00489697. doi: 10.1016/j.scitotenv.2004.07.011.

Gao, B.-B., Wang, J.-F., Fan, H.-M., Xu, K., Hu, M.-G., and Chen, Z.-Y. A stratified optimization method for a multivariate marine environmental monitoring network in the Yangtze River estuary and its adjacent sea. *Int. J. Geogr. Inf. Sci.*, 29(8):1332–1349, 2015. ISSN 1365-8816. doi: 10.1080/13658816.2015.1024254.

Gelfand, A. E., Sahu, S. K., and Holland, D. M. On the Effect of Preferential Sampling in Spatial Prediction. *Environmetrics*, 23(7):565–578, nov 2012. ISSN 1180-4009. doi: 10.1002/env.2169.

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. A long neglected world malaria map: {\it Plasmodium vivax} endemicity in 2010. *PLoS Negl. Trop. Dis.*, 2012. doi: 10.1371/journal.pntd.0001814.

Giorgi, E. and Diggle, P. J. PrevMap : an R Package for Prevalence Mapping, 2015.

Guerra, C. a., Hay, S. I., Lucioparedes, L. S., Gikandi, P. W., Tatem, A. J., Noor, A. M., and Snow, R. W. Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project. *Malar. J.*, 6(1):1–13, 2007. ISSN 1475-2875. doi: 10.1186/1475-2875-6-17. URL `http://dx.doi.org/10.1186/1475-2875-6-17`.

Hay, S. I., Guerra, C. A., Gething, P. W., Patil, A. P., Tatem A.J., Noor, A. M., Kabaria1, C. W., Manh, B. H., Elyazar, I. R. F., Brooker, S., Smith, D. L., Moyeed, R. A., and Snow, R. W. A world malaria map: Plasmodium falciparum endemicity in 2007. *PLoS Med.*, 6, 2009. doi: 10.1371/journal.pmed.1000048.

Hay, S. I., Guerra, C. A., Tatem, A. J., Noor, A. M., and Snow, R. W. The global distribution and population at risk of malaria: past, present, and future. *Lancet Infect. Dis.*, 4(6): 327–336, jun 2004. ISSN 1473-3099. doi: 10.1016/S1473-3099(04)01043-6. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3145123/`.

Hu, M. G. and Wang, J. F. A spatial sampling optimization package using MSN theory. *Environ. Model. Softw.*, 26(4):546–548, 2011. ISSN 13648152. doi: 10.1016/j.envsoft.2010.10.006.

Krige, D. G. A Statistical Approach to Some Mine Valuation and Allied Problems on the Witwatersrand. *J. Chem. Metall. Min. Soc. South Africa*, 52:119–139, 1951.

Marchant, B. P., Lark, R. M., and Wheeler, H. C. Developing methods to improve sampling efficiency for automated soil mapping. Technical Report 364, Home-Grown Cereals Authority, 2005.

Matérn, B. *Spatial Variation*. PhD thesis, Stockholm, 1960.

McBratney, A. B., Webster, R., and Burgess, T. M. The design of optimal sampling schemes for local estimation and mapping of of regionalized variables - I: Theory and method. *Comput. Geosci.*, 7(4):331–334, 1981.

Müller, W. G. A comparison of spatial design methods for correlated observations. *Environmetrics*, 16:495–505, 2005. ISSN 11804009. doi: 10.1002/env.717.

Müller, W. G. *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*. Springer-Verlag, Berlin, 3 edition, 2007. ISBN 9783540311744.

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

Nychka, D. and Saltzman, N. Design of Air-Quality Monitoring Networks. *Case Stud. Environ. Stat. SE - 4*, 132:51–76, 1998. doi: 10.1007/978-1-4612-2226-2{_}4.

Pati, D., Reich, B. J., and Dunson, D. B. Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, 98:35–48, 2011. ISSN 00063444. doi: 10.1093/biomet/asq067.

Pilz, J. and Spöck, G. Spatial sampling design for prediction taking account of uncertain covariance structure. In *7th Int. Symp. Spat. Accuracy Assess. Nat. Resour. Environ. Sci.*, 2006.

Ritter, K. Asymptotic optimality of regular sequence designs. *Ann. Stat.*, 24(5):2081–2096, 1996. ISSN 00905364. doi: 10.1214/aos/1069362311.

Roca-Feltrer, A., Lalloo, D. G., Phiri, K., and Terlouw, D. J. Short Report : Rolling Malaria Indicator Surveys (rMIS): a potential district-level malaria monitoring and evaluation (M&E) tool for program managers. *Am. J. Trop. Med. Hyg.*, 86(1):96–98, jan 2012. ISSN 1476-1645. doi: 10.4269/ajtmh.2012.11-0397.

Royle, J. and Nychka, D. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Comput. Geosci.*, 24(5):479–488, jun 1998. ISSN 00983004. doi: 10.1016/S0098-3004(98)00020-X.

Russo, D. Design of an Optimal Sampling Network for Estimating the Variogram. *Soil Sci. Soc. Am. J.*, 48(4):708–716, feb 1984. ISSN 0361-5995. doi: 10.2136/sssaj1984.03615995004800040003x.

Shaddick, G. and Zidek, J. V. A case study in preferential sampling: Long term monitoring of air pollution in the UK. *Spat. Stat.*, 9:51–65, 2014. ISSN 22116753. doi: 10.1016/j.spasta.2014.03.008. URL `http://www.sciencedirect.com/science/article/pii/S2211675314000219`.

Stanton, M. C. and Diggle, P. J. Geostatistical analysis of binomial data: generalised linear or transformed Gaussian modelling? *Environmetrics*, 24(3):158–171, may 2013. ISSN 1099-095X. doi: 10.1002/env.2205.

Su, Y. S. Y. and Cambanis, S. Sampling Designs for Estimation of a Random Process. *Stoch. Process. their Appl.*, 46:47–89, 1993. doi: 10.1109/ISIT.1993.748644.

Thompson, S. K. and Collins, L. M. Adaptive sampling in research on risk-related behaviors. *Drug Alcohol Depend.*, 68 Suppl 1:S57———67, nov 2002. ISSN 0376-8716. URL `http://www.ncbi.nlm.nih.gov/pubmed/12324175`.

Wang, J., Haining, R., and Cao, Z. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. *Int. J. Geogr. Inf. Sci.*, 24(4):523–543, 2010. ISSN 1365-8816. doi: 10.1080/13658810902873512.

Yfantis, E. A., Flatman, G. T., and Behar, J. V. Efficiency of Kriging Estimation for Square , Triangular , and Hexagonal Grids. *Math. Geol.*, 19(3), 1987.

Zhu, Z. and Stein, M. L. Spatial sampling design for prediction with estimated parameters. *J. Agric. Biol. Environ. Stat.*, 11(1):24–44, mar 2006. ISSN 1085-7117. doi: 10.1198/108571106X99751.

Zidek, J. V., Shaddick, G., and Taylor, C. G. Reducing estimation bias in adaptively changing monitoring networks with preferential site selection. *Ann. Appl. Stat.*, 8(3):1640–1670, 2014. ISSN 1932-6157. doi: 10.1214/14-AOAS745.

Zouré, H. G. M., Noma, M., Tekle, A. H., Amazigo, U. V., Diggle, P. J., Giorgi, E., and Remme, J. H. F. 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. *Parasit. Vectors*, 7:326, 2014. ISSN 1756-3305. doi: 10.1186/1756-3305-7-326.