Environmetrics 00, 1–22

DOI: 10.1002/env.XXXX

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

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

Summary: 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 designs plus close pairs*. 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.

[a] *Lancaster Medical School, Lancaster University, Lancaster, UK.*
[b] *College of Medicine, University of Malawi, Blantyre, Malawi.*
[c] *Malawi-Liverpool-Wellcome Trust, Blantyre, Malawi.*
[d] *Liverpool School of Tropical Medicine, Liverpool, UK.*
[*] *Correspondence to: M. G. Chipeta, Lancaster University, Lancaster Medical School, Furness College, Lancaster, LA1 4YG, UK. E-mail: m.chipeta@lancaster.ac.uk*

This paper has been submitted for consideration for publication in *Environmetrics*

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 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., 2016), 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 we review the existing literature on non-adaptive geostatistical design strategies. In Section 3 we describe our proposed class of designs. Section 5 reports on a simulation study of the predictive performance of the proposed design class. We also compare the performance of our proposed designs with empirical kriging (EK) optimal designs. Section 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 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 Development Core Team, 2015).

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,

$$MSPE = \int_D \mathrm{E}[\{\hat{S}(x) - S(x)\}^2 dx, \tag{1}$$

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 et al. (1981); McBratney and Webster (1981); Yfantis et al. (1987); Ritter (1996); 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 2.1.

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 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 et al. (2015) and Chipeta et al. (2016).

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.

2.1.1. 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 et al., 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).

2.1.2. 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 et al., 1981; McBratney and Webster, 1981; Olea, 1984; Yfantis et al., 1987). However, square lattices are more common in practice.

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 et al., 2012; Brown et al., 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 et al., 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, 2003, 2004) for details.

Grafström et al. (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 et al. (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).

2.1.4. 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.

3. INHIBITORY GEOSTATISTICAL DESIGNS

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

$$[S|Y] = [S][Y|S] / \int [S][Y|S] \mathrm{d}S \tag{2}$$

A typical spatial prediction problem involves making inferences about a functional $T = \mathcal{T}(\mathcal{S})$ given data (Y_i, X_i), $i = 1, \ldots, n$. 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,

$$APV = \int_{\mathcal{D}} \mathrm{Var}\{S(x)|Y\} \mathrm{d}x \tag{3}$$

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.

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 $j = 1, \ldots, k$, 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

$$\delta_{(k)} = \delta_0 \sqrt{n/(n-k)}, \tag{4}$$

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 $\{X_i,\ i = 1, \ldots, N\}$, 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 (3), using a general-purpose numerical optimiser. Specifically, we use the controlled random search (CRS) procedure for global optimisation (Price, 1976, 1983). The procedure allows for box constraints that we impose on the design parameters of interest above.

4. EMPIRICAL KRIGING OPTIMAL DESIGNS

In our simulation study (Section 5), we compare the performance of inhibitory plus close pairs design with some of the optimal designs we have reviewed in Section 1, such as empirical kriging (EK) designs implemented by Zimmerman (2006) and Müller et al. (2015). These designs minimise the empirical kriging criterion:

$$EK(\mathcal{X}) = \max_{x \in \mathcal{D}} \{\text{Var}[\hat{Y}(x) - Y(x)] + \text{tr}\{M_\theta \, \text{Var}[\partial \hat{Y}(x)/\partial \theta]\}\}. \tag{5}$$

This adds an explicit additive correction term to the normalised classical prediction variance. In Equation (5), $\hat{Y}(x)$ is the posterior mean of $Y(x)$ given data at $\mathcal{X} = \{x_i; i = 1, \ldots, n\}$ 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 $\mathbf{ICP}(n, k, \delta, \zeta)$ designs. In our simulation study in Section 5.3, we follow the SSA algorithm outlined in Müller et al. (2015).

5. SIMULATION STUDY

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 (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 (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,

$$Y|S \sim N(\mu, \tau^2) \tag{6}$$

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

$$Y|S, U \sim Bin(n, p), \tag{7}$$

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.

5.1. Linear Gaussian Model

For each parameter combination, we generated data at $n = 150$ sampling locations. Figure 1a shows an inhibitory design without close pairs and $\delta = 0.06$, corresponding to packing density $\rho \approx 0.424$, whilst Figure 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.

Figure 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.

[Figure 1 about here.]

[Figure 2 about here.]

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 3 panels A – B. In contrast, when τ^2 is larger, close pairs show substantial benefit, see Figure 3 panels C – E.

[Figure 3 about here.]

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 (3) by a Monte Carlo average over $s = 1000$ independent simulations of Y. Figures 4a to 4b show that inhibitory designs with $\delta = 0.06$ give the best results, agreeing with the findings in Section 5.1, Figure 2. Similarly, Figure 4c again shows that inhibitory designs with an intermediate number of close pairs give the best performance when τ^2 is relatively large.

[Figure 4 about here.]

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 3.3 and Section 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 (6)) for the data-generation process. This includes an unobserved stationary Gaussian process $S(x)$ with mean zero, variance $\sigma^2 = 1$ and a Matérn correlation. We evaluate the integral in Equation (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 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.

[Figure 5 about here.]

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 et al., 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 perimeter is home to a population of approximately 100,000. Figure 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.

[Figure 6 about here.]

The sampling unit is a household. We first fit the Binomial model Equation (7), with three parameters representing the two variance components and the rate of 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 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 (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 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. 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.

[Table 1 about here.]

[Figure 7 about here.]

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 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 (5) diminishes with increasing sample size n. Müller et al. (2015) showed that for a design with $n \geq 10$, $\max_{x \in \mathcal{D}} \text{Var}[\hat{Y}(x) - Y(x)]$ and $EK(\mathcal{X}_n)$ yield similar values, implying that the effect of the correction term in Equation (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 Müller et al. (2015); Zhu and Stein (2006), 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 Development Core Team (2015); see also `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 (3) in the presence of explanatory variables involves no additional principles.

ACKNOWLEDGEMENTS

The MMP study 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.

FUNDING

Michael Chipeta is supported by an 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

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

Bijleveld, A. I., J. A. van Gils, J. van der Meer, A. Dekinga, C. Kraan, H. W. van der Veer, and T. Piersma (2012). Designing a benthic monitoring programme with multiple conflicting objectives. *Methods in Ecology and Evolution 3*(3), 526–536.

Brown, J., B. Robertson, and T. McDonald (2015). Spatially Balanced Sampling: Application to Environmental Surveys. *Procedia Environmental Sciences 27*, 6–9.

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

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

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

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

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

Diggle, P. J., R. Menezes, and T.-L. Su (2010, Mar). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(2), 191–232.

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

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

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

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

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

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

Matérn, B. (1960). *Spatial Variation*. Ph. D. thesis, Stockholm.

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

McBratney, A. B., R. Webster, and T. M. Burgess (1981). The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalised Variables – I: Theory and method. *Computers & Geosciences 7*(4), 331–334.

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

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

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

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

Nychka, D. and N. Saltzman (1998). Design of Air-Quality Monitoring Networks. *Case Studies in Environmental Statistics SE - 4 132*, 51–76.

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

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

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

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

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

R Development Core Team (2015). R: A Language and Environment for Statistical Computing. Technical report, R Foundation for Statistical Computing, Vienna, Austria.

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

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

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

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

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

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

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

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

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

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

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

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

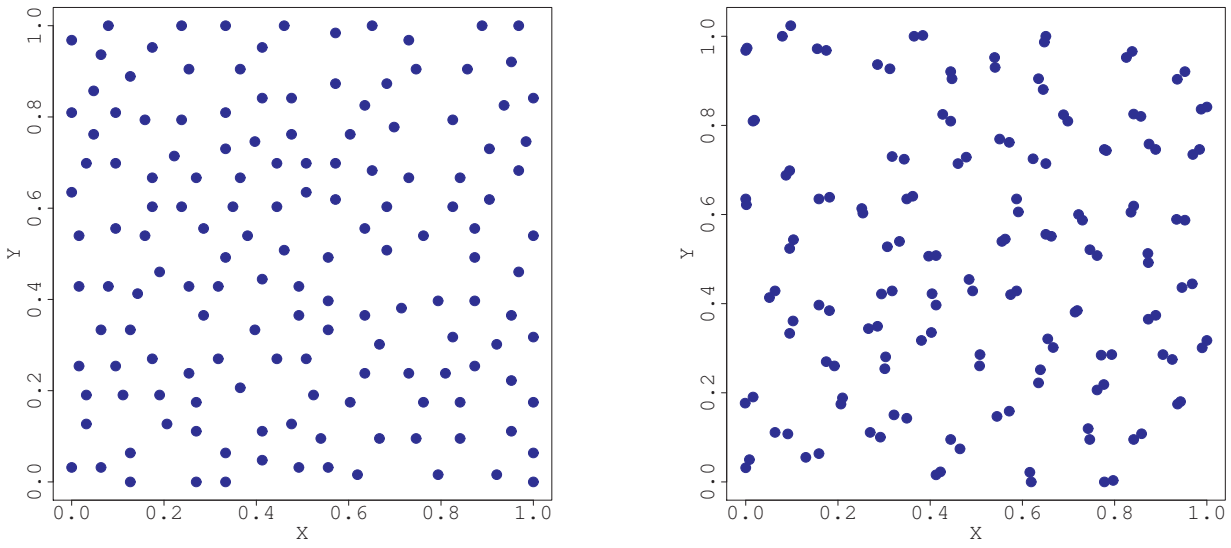
Yfantis, E. A., G. T. Flatman, and J. V. Behar (1987). Efficiency of Kriging Estimation for Square , Triangular , and Hexagonal Grids. *Mathematical Geology 19*(3), 183 – 205.

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

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

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

FIGURES



(a) **(b)**

Figure 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.

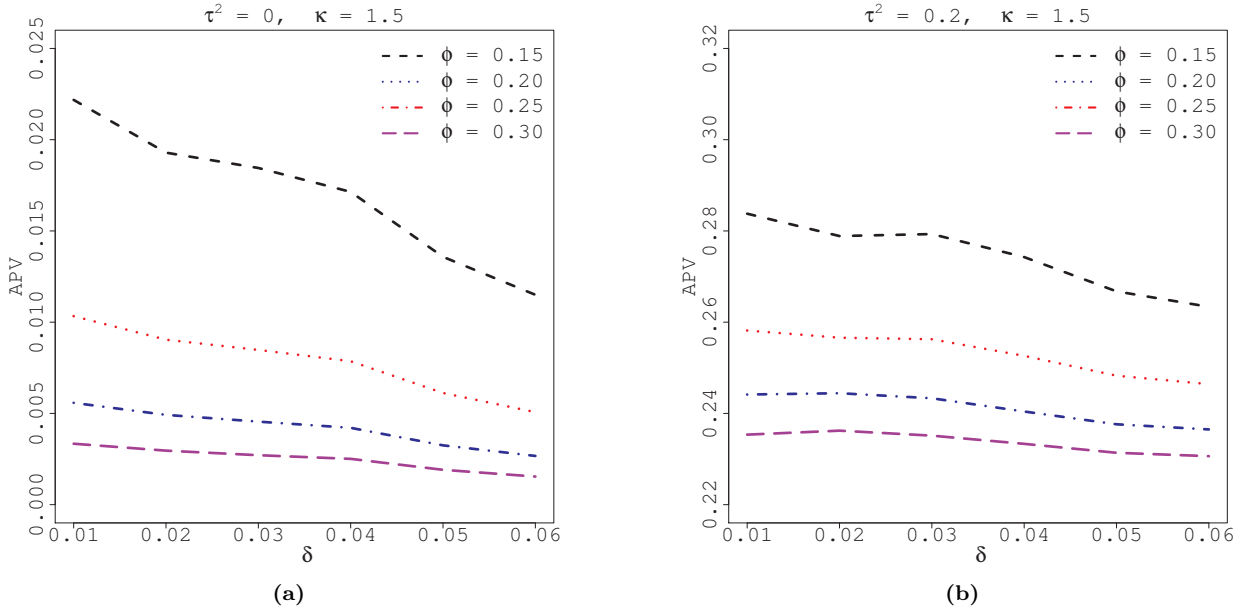


Figure 2. Average prediction variance for varying simple inhibitory designs, $\delta = 0.01$ to 0.06, $\kappa = 1.5, \sigma^2 = 1$ and $n = 150$. Panel (a) $\tau^2 = 0$ and panel (b) $\tau^2 = 0.2$.

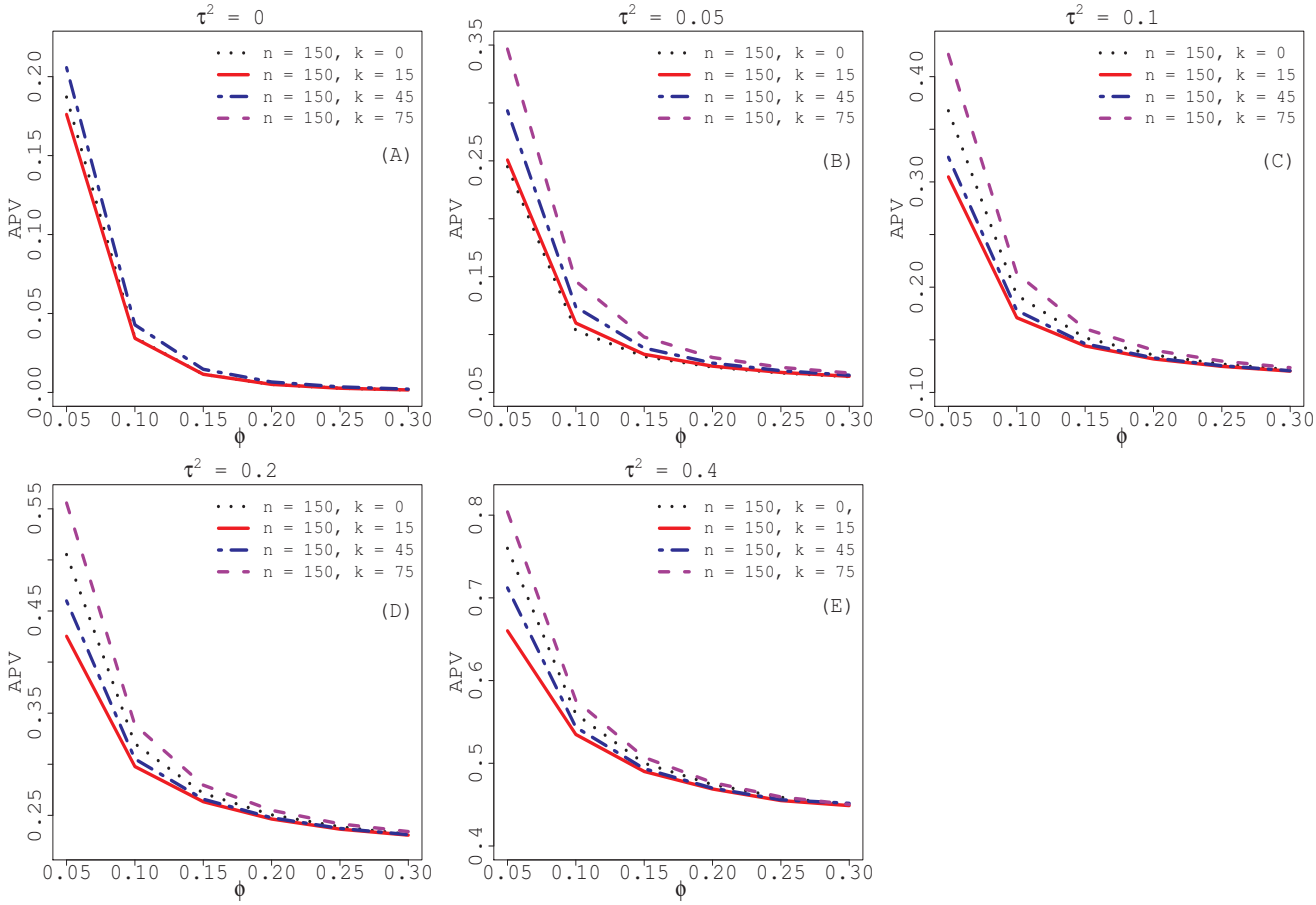


Figure 3. Comparing the efficiencies of inhibitory designs: without close pairs, with 15, 45 and 75 close pairs. The fixed total $n = 150$ for each of the designs.

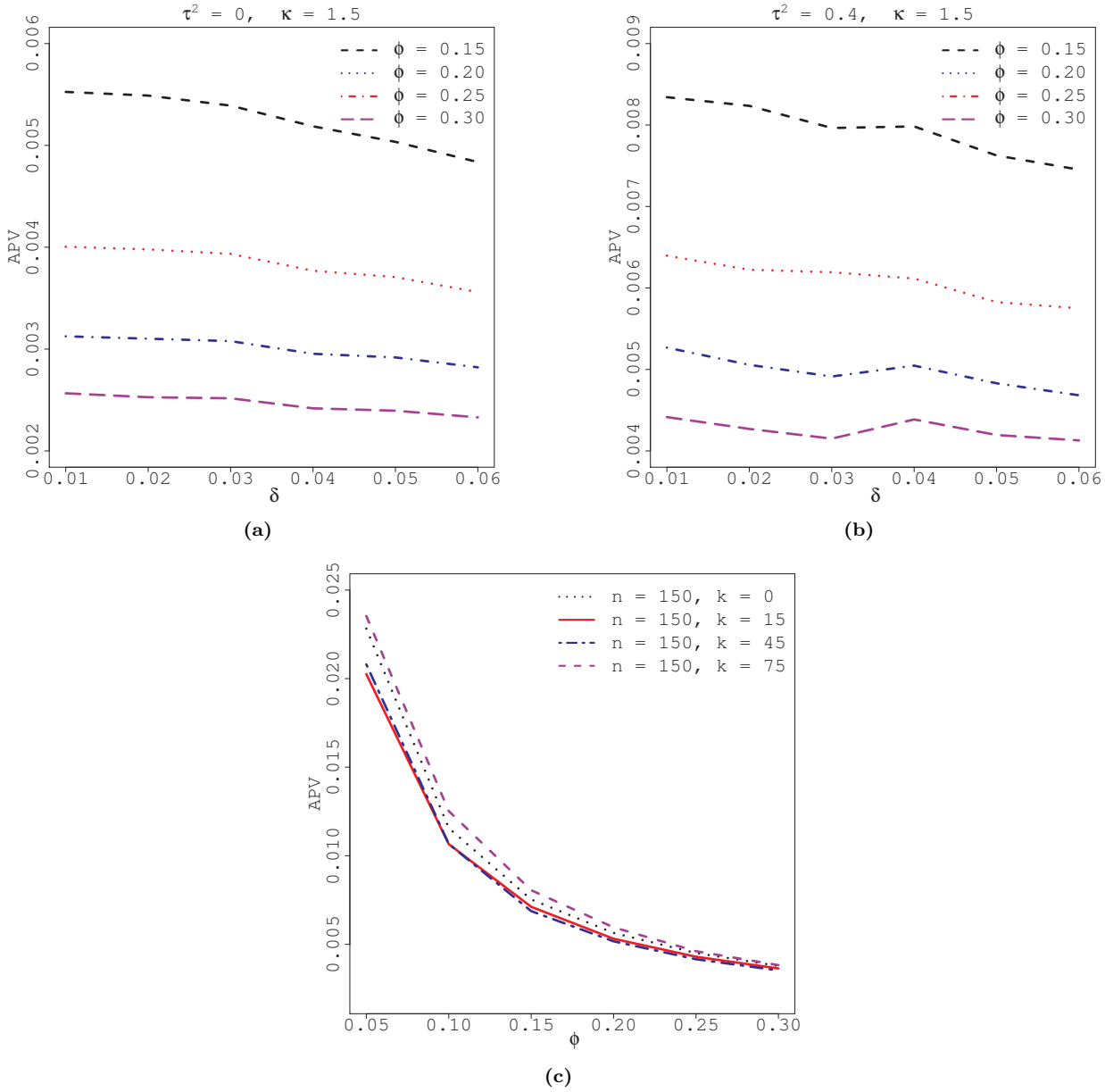


Figure 4. Average prediction variance for varying simple inhibitory designs - Binomial model, $\delta = 0.01$ to 0.06, $\kappa = 1.5, \sigma^2 = 1$ and $n = 150$. Panel (a) $\tau^2 = 0$ and panel (b) $\tau^2 = 0.4$. Panel (c) compares the efficiencies of inhibitory designs with 15, 45 and 75 close pairs. The fixed total $n = 150$ for each of the designs.

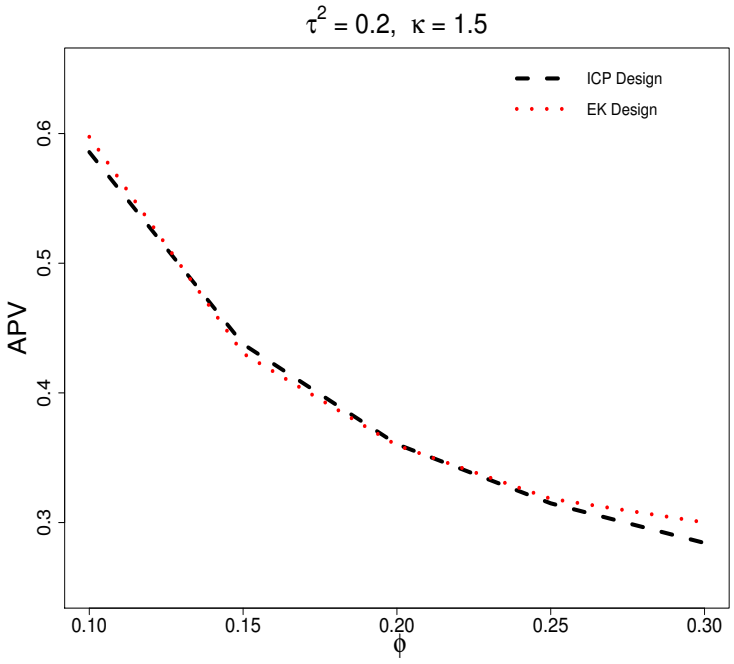


Figure 5. Inhibitory plus close pairs design vs Empirical kriging optimal design.

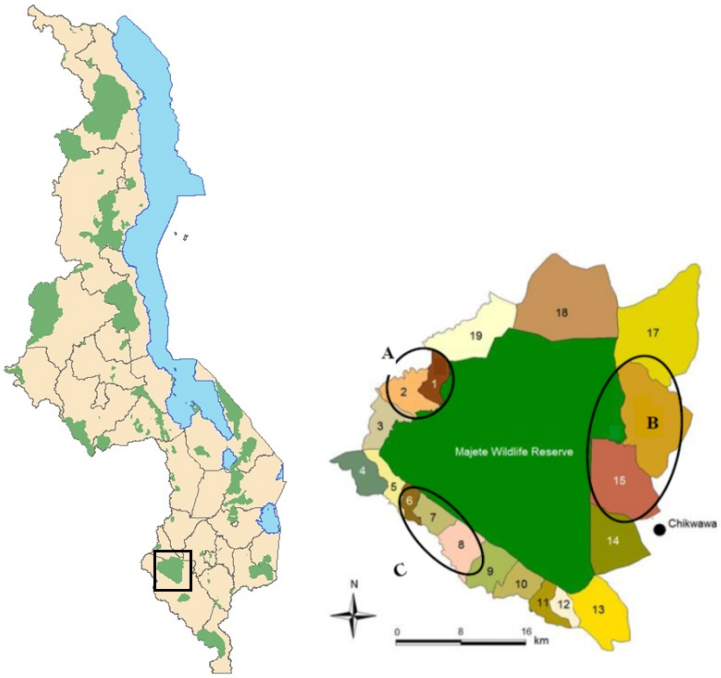


Figure 6. The map of Malawi, showing Majete Wildlife Reserve highlighted (left) and its perimeter with focal areas A, B and C highlighted (right).

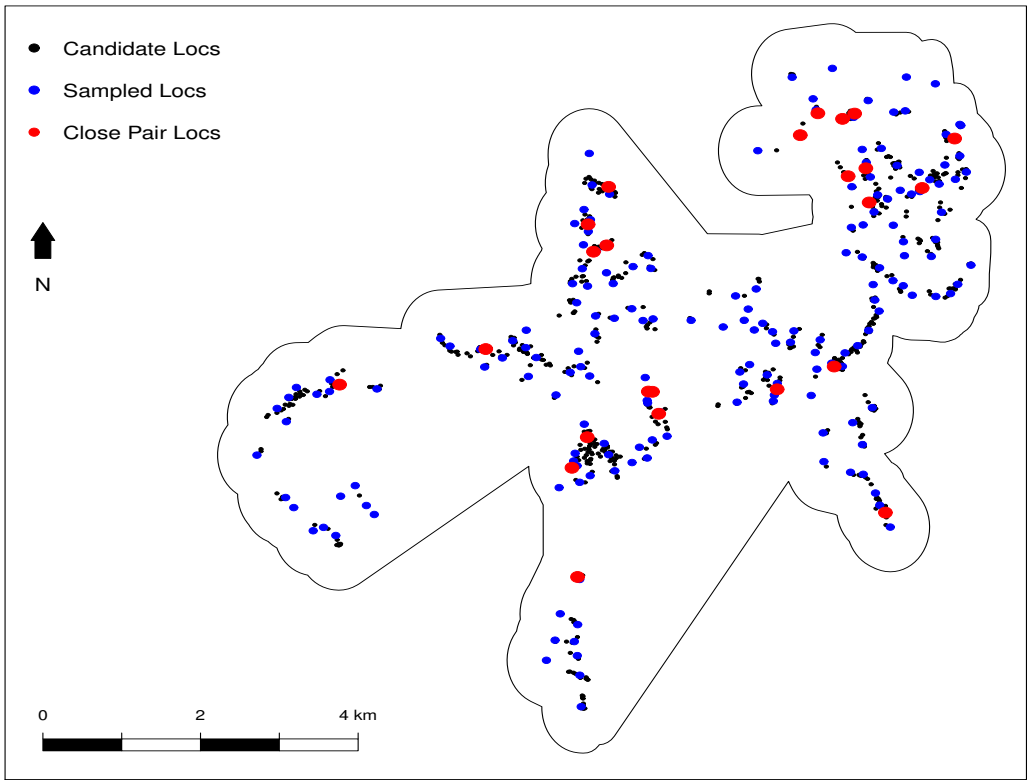


Figure 7. Inhibitory (blue dots) plus close pairs design locations (red dots) and all potential sampling locations (black dots), in focal area A

TABLES

Table 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)