geosample: an R Package for Geostatistical Sampling Designs

Michael G Chipeta
University of Oxford

Barry Rowlingson
Lancaster University

Peter J Diggle
Lancaster University

Abstract

In this paper we introduce a new R package, **geosample**, for constructing geostatistical sampling designs. The new package implements classes of *adaptive* and *non-adaptive* probability-based sampling designs. Non-adaptive sampling designs choose all sampling locations in a single wave without reference to existing data. Adaptive sampling designs use information from existing data to inform a choice of additional sample locations at each sampling wave. We illustrate the use of the package through the construction of both adaptive and non-adaptive designs, using a simulated data-set and malaria prevalence data from southern Malawi.

Keywords: adaptive sampling designs, inhibitory sampling designs, geostatistics, surveillance sampling, R.

Please cite this manuscript as:

Chipeta, M G, Rowlingson B and Diggle, P J. (2019). geosample: An R package for geostatistical sampling designs. *Under review Journal of Statistical Software.*

1. Introduction

Geostatistics is primarily concerned with the investigation of an unobserved spatial phenomenon $S = \{S(x) : x \in \mathcal{D} \subset \mathbb{R}^2\}$, where $\mathcal{D}$ is a geographical region of interest, using data in the form of measurements y_i at locations $x_i \in \mathcal{D}$. Typically, each y_i can be regarded as a noisy version of $S(x_i)$. We write $\mathcal{X} = \{x_1, \ldots, x_n\}$ and call $\mathcal{X}$ the *sampling design*. This paper introduces a new R (R Core Team 2017) package, **geosample**, for geostatistical sampling designs. The work was motivated by applications to disease prevalence mapping, where the main focus of scientific interest is on deciding which households to sample in each round of sampling so as to optimise the precision of the resulting sequence of area-wide prevalence maps.

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

$\{(y_i, x_i) : i = 1, \ldots, n\}$; and *prediction* of the unobserved realisation of $S(x)$, or particular characteristics of this realisation.

In practice, geostatistical sampling designs that are efficient for parameter estimation are generally inefficient for spatial prediction, and vice-versa (Diggle and Ribeiro 2007; Müller 2007). Additionally, parameter values are usually unknown in practice, hence design for prediction involves a compromise. Furthermore, the diversity of potential predictive targets requires design strategies to be context-specific (Chipeta, Terlouw, Phiri, and Diggle 2016a). 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 (Chipeta *et al.* 2016a; Chipeta, Terlouw, Phiri, and Diggle 2016b).

In this paper, we describe the implementation of geostatistical sampling algorithms for constructing *adaptive* and *non-adaptive* classes of designs as described in Chipeta *et al.* (2016a) and Chipeta *et al.* (2016b), respectively. The **geosample** package includes functionality to determine sampling locations within a set of spatial constraints and information from existing sampling locations. The package makes use of functions from other R packages, including **sp, splancs, rgeos** and **pdist**, which support data manipulation and computation. In order to determine new sampling locations in the case of adaptive sampling, **geosample** requires *predictions* to be made at all unobserved (potential sampling) locations. These can be obtained from existing packages including **PrevMap, geoR, lgcp** and **spatstat** that can carry out predictive inference.

The paper is structured as follows. In Section 2 we give the theoretical background within which adaptive and non-adaptive designs have been developed. Section 2.1 describes a class of non-adaptive designs, including random and inhibitory sampling (with or without *close pairs*). Section 2.2 describes a class of adaptive designs. Section 3 gives an overview of the package by way of a walk-through of a simulated dataset, and an application to malaria prevalence mapping in southern Malawi. Section 4 is a concluding discussion.

2. Methodological framework

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. Methods in the **geosample** package focus on designs for efficient prediction when model parameters are unknown.

2.1. Non-adaptive designs

We first consider non-adaptive geostatistical designs. These offer standard ways of collecting and analysing geostatistical data in which sampling locations are fixed in advance of any data collection. Two standard non-adaptive designs are a *completely random design*, in which the sample locations x_i form an independent random sample from the uniform distribution on $\mathcal{D}$, and a *completely regular design* in which the x_i form a regular square or, less commonly, triangular lattice, ensuring an even coverage over the study region. Diggle and Lophaven (2006) implemented lattice-based designs and Chipeta *et al.* (2016b) implemented *inhibitory geostatistical designs* in which sampled locations exhibit a degree of spatial regularity, intermediate between completely random and lattice designs.

Completely random designs

A completely randomised design has locations x_i; $i = 1, \ldots, n$, 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 methods (Diggle, Menezes, and Su 2010). However, the resulting uneven coverage of $\mathcal{D}$ has a negative impact on spatial prediction. Compeletely randomised design strategies are well established in classical survey sampling (Cochran 1977). An undesirable feature of these designs when the goal is prediction is their tendency to leave large swaths of unsampled areas (Müller 2007). Nevertheless, previous research studies have shown that this class of designs is efficient for estimation of covariance structure. See, for example, Russo (1984); Warrick and Myers (1987); Müller and Zimmerman (1999); Lark (2002).

In the **geosample** package, a completely randomised design $\mathcal{X}$ is implemented by the function `random.sample`. The function takes a sample of the specified number of locations n, either from N potential sampling locations without replacement, or as an independent sample from a designated region $\mathcal{D}$.

Inhibitory designs

In some geostatistical analysis problems, the covariance structure is assumed to be known, and the goal is spatial prediction. Previous research has shown that classes of completely regular designs are then more efficient than completely random designs. See, for example, McBratney, Webster, and Burgess (1981); McBratney and Webster (1981); Yfantis, Flatman, and Behar (1987); Ritter (1996). In practice, the covariance structure is usually unknown and needs to be estimated. Usually, one has to use the same data for estimation of covariance parameters and for spatial prediction, and efficient prediction requires good estimates of the second order characteristics (Müller, Pronzato, Rendas, and Waldl 2015). Designs that offer a compromise between the two contrasting aims are therefore attractive. One such example is the following class of inhibitory designs.

An inhibitory design has n random locations in $\mathcal{D}$ with the constraint that no two locations are separated by a distance of less than a specified value δ. Inhibitory designs adhere to the established principles of random sampling theory while guaranteeing some degree of spatial regularity. This construction has also been suggested as a model for naturally occuring patterns of points that exhibit spatial regularity (Matérn 1986) (originally published in 1960). All design points $\mathcal{X}$ that meet the inhibitory constraint are equally likely to be picked. Chipeta *et al.* (2016b) developed and implemented *simple inhibitory* (**SI**) and *inhibitory with close pairs* (**ICP**) design strategies. In the latter, $n-k$ simple inhibitory sample locations are augmented by k locations each positioned close to one of the randomly selected $n-k$ locations in the simple inhibitory design, uniformly distributed within a disk of radius ζ. Inclusion of close pairs of sampled locations helps to identify a suitable parametric family for the specified correlation structure of a geostatistical dataset.

Inhibitory design construction can be applied whether or not the potential sampling locations are confined to a finite set of points. In the **geosample** package, inhibitory designs for a finite set of points are implemented by the function `discrete.inhibit.sample`, and for points in a continuum, by the function `contin.inhibit.sample`. In each of these implementations, **geosample** package can generate simple inhibitory or inhibitory with close pair samples.

An inhibitory design, $\mathbf{SI}(n,\delta)$, is implemented as follows. Choose a *packing density* for the design, i.e. the proportion of $\mathcal{D}$ covered by n non-overlapping disks of diameter δ, given by $\rho = (n\pi\delta)/(4|\mathcal{D}|)$. An $\mathbf{SI}(n,\delta)$ design on $\mathcal{D}$ is then generated by the following steps.

- Step 1: Draw a sample of locations $x_i : i = 1,\ldots,n$ completely at random in $\mathcal{D}$;
- Step 2: Set $i = 1$;
- Step 3: Calculate the minimum, d_{min}, of the distances from x_i to all other x_j in the current sample;
- Step 4: If $d_{min} \geqslant \delta$, increase i by 1 and return to step 3 if $i \leqslant n$, otherwise stop;
- Step 5: If $d_{min} < \delta$, replace x_i by a new location drawn completely at random in $\mathcal{D}$ and return to step 4.

For efficient parameter estimation, the simple inhibitory sampling scheme can be augmented by pairs of closely spaced points. The algorithm then requires the following additional steps. Let k be the required number of close pairs. Choose a value ζ such that a close pair of points will be separated by a distance of at most ζ. For a total of n points, an $\mathbf{ICP}(n,k,\delta,\zeta)$ design consists of an $\mathbf{SI}(n-k,\delta)$ design with inhibition distance δ augmented by k locations each positioned relative to one of the randomly selected $n-k$ locations in the **SI** design according to the uniform distribution over a disk of radius ζ. The following steps generate the $\mathbf{ICP}(n,k,\delta,\zeta)$ design.

- Step 1: Construct a simple inhibitory design $\mathbf{SI}(n-k, \delta)$;
- Step 2: Sample k from $x_1, \ldots, x_{n-k}$ without replacement and call this set $x_j^*, j = 1, \ldots, k$;
- Step 3: For $j = 1, \ldots, k, x_{n-k+j}$ is uniformly distributed on the disk with center x_j^* and radius ζ.

2.2. Adaptive designs

We now focus on a class of adaptive geostatistical designs, in which sampled locations are defined in batches at a sequence of times, and the locations in any batch use data from earlier batches to optimise data collection towards the analysis objective. The adaptive sampling design criterion ensures that data are collected only from locations that will deliver useful additional information (Chipeta *et al.* 2016a).

An adaptive design strategy takes the following approach.

- Step 1: Specify the finite set, $\mathcal{X}^*$ say, of n^* potential sampling locations $x_i \in \mathcal{D}$. If all points $x \in \mathcal{D}$ are eligible, we approximate this by specifying $\mathcal{X}^*$ as a finely spaced grid to cover $\mathcal{D}$;
- Step 2: Use a non-adaptive design to choose an initial set of sample locations, $\mathcal{X}_0 = \{x_i \in \mathcal{D} : i = 1, \ldots, n_0\}$;
- Step 3: Use the corresponding data Y_0 to estimate the parameters of an assumed geostatistical model;
- Step 4: Specify a selection criterion for the addition of one or more new sample locations to form an enlarged set $\mathcal{X}_0 \cup \mathcal{X}_1$;
- Step 5: Repeat steps 3 and 4 with augmented data Y_1 at the points in $\mathcal{X}_1$;
- Step 6: Continue until the required number of points has been sampled, a required performance criterion has been achieved or no more potential sampling points are available.

In step 2, any initial design can be supplied, but our general recommendation would be to use an *inhibitory plus close pairs* design.

Adaptive sampling is implemented by `adaptive.sample` function. The function implements *singleton adaptive sampling*, in which individual locations are chosen sequentially, allowing x_{k+1} to depend on data obtained at all earlier locations $x_1, \ldots, x_k$, and *batch adaptive sampling*, where sets of $b > 1$ locations are chosen, with each set $(x_{k+1}, \ldots, x_{k+b})$, dependent on data from all earlier locations $x_1, \ldots, x_k$.

2.3. Selection criteria

The `adaptive.sample` function offers a choice of either *predictive variance (PV)* or *exceedance probabilities (EP)* selection criteria in step 4 above. For the predictive target $T = S(x)$ at a particular location x, given an initial set of sampling locations $\mathcal{X}_0 = (x_1, \ldots, x_{n_0})$ the available set of additional sampling locations is $A_0 = \mathcal{X}^* \backslash \mathcal{X}_0$.

In the PV selection criterion, any $x \in A_0$ has the predictive variance, $PV(x) = \text{Var}(T|Y_0)$ (Diggle and Ribeiro 2007). The algorithm then chooses the locations x^* with the largest values of $PV(x)$, either singly or in batches (Chipeta *et al.* 2016a). For the EP selection criterion, each $x \in A_0$ has exceedance probability, $EP(x) = P[\{T(x) > t|y_0\} - 0.5]$ for a given threshold t (Giorgi and Diggle 2017). The algorithm then chooses the locations $x^* = arg\ min_{A_0} EP(x)$, either singly or in batches. When locations are chosen in batches, a *minimum distance* penalty is imposed for both PV and EP criteria. This ensures that no two sampling locations are separated by a distance of less than δ, to avoid sampling from multiple locations x at which the corresponding $S(x)$ are highly correlated.

2.4. Perfomance criteria

For design strategies implemented in **geosample**, we focus on a *predictive target* $T = \mathcal{T}(S)$, where the property of S is of primary interest. We use a generic measure of the predictive accuracy of a design $\mathcal{X}$, the mean square error,

$$MSE(\hat{T}) = E[(T - \hat{T})^2] \tag{1}$$

where $\hat{T} = E[T|Y; \mathcal{X}]$ is the minimum mean square error predictor of T for any given design $\mathcal{X}$ in $\mathcal{D}$.

3. Introduction to the geosample package

In this section, we present an introduction to the **geosample** package functionality by means of a walk-through of some geostatistical sampling examples. The **geosample** package provides compatibility with common spatial packages including **sp** and **sf**. In Section 3.1 we give a unifying workflow for using **geosample** with other R packages, such as **PrevMap, geoR** and other spatial statistics packages, for generating geostatistical samples, estimating parameters and predicting the phenomenon of interest in unobserved locations x^*. Section 3.2 outlines sampling and inference from a simulated dataset using classes of design discussed earlier. Section 3.3 reports an application of the **geosample** package functionality to adaptive sampling for malaria prevalence mapping in Majete, southern Malawi.

3.1. Geostatistical sampling workflow

The **geosample** package focuses on geostatistical sampling designs that compromise between designing for efficient parameter estimation and designing for efficient prediction given the values of relevant model parameters. The workflow relies on functionality and outputs from other R packages as determined by the user, mainly to do with parameter estimation and spatial predictions. Figure 1 is a diagrammatic representation of the workflow.

The first stage involves deciding on and implementing the initial sampling design, dependent on the objective(s) of the geostatistical analysis problem at hand. The initial design is a non-adaptive design, which can be any of the designs outlined in Section 2.1. Once data have been collected from sample locations in the chosen design, the second stage is to analyse the data in order to estimate model parameters, within an assumed geostatistical model. Parameter estimation can take several forms including guess work, also known as curve fitting "by eye", variogram fitting or formal estimation using methods such as maximum likelihood estimation. See Mardia and Marshall (1984); Christensen (2004); Diggle and Ribeiro (2007) for details. In our walk through examples, we assume a linear Gaussian model of the form:

$$Y_i = d(x_i)'\beta + S(x_i) + Z_i, i = 1, \ldots, n \tag{2}$$

where the Z_i are mutually independent $N(0, \tau^2)$ random variables and $S(x)$ is a stationary Gaussian process, with mean μ, variance $\sigma^2 = \text{Var}(S(x))$ and correlation function $\rho(u) = \text{Corr}\{S(x), S(x')\}$, where $u = ||x - x'||$ and $||\cdot||$ denotes Euclidean distance. The $d(x_i)$ are spatially referenced covariates. In all the examples, we work with the Matérn correlation function (Matérn 1986; Diggle and Ribeiro 2007):

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

The third stage is to predict $T^* = (T(x_{(n+1)}), \ldots, T(x_{(n+q)}))^{\top}$ at q additional locations where measurements have not been taken. Estimates of all model parameters are plugged into the prediction equation as if they were the true parameter values, in a process referred to as "*plug-in prediction*". Inferences can be made, depending on the context, for a range of predictive targets, for example: a single value $S(x_0)$; the value of $S(\cdot)$ over an area of interest or subsets thereof; the minimum or maximum value of $S(x)$; or the probability that $S(x)$ is below or above a particular threshold. This requires all relevant explanatory variables to be available at the prediction locations.

The fourth stage is the implementation of adaptive sampling if there is need for additional samples to achieve the required predictive accuracy. Required inputs include predictions at

all unobserved (potential sampling) locations, a sample selection criterion and any spatial constraints. Several sampling rounds can be implemented, allowing for spatial constraints to change at each cycle. This process involves repeated estimation and prediction stages. Adaptive sampling stops when the specified stopping condition(s) have been achieved, see Section 2.2 for details.

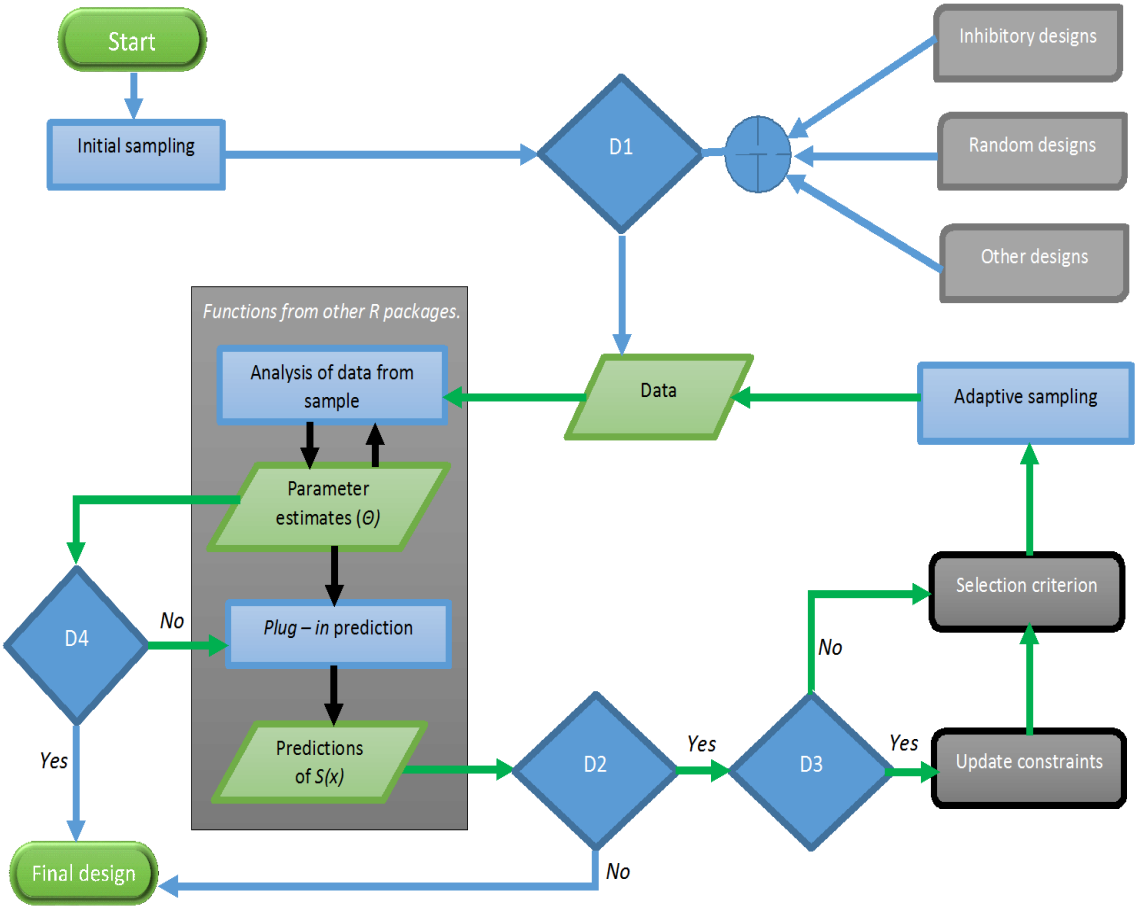

Figure 1: Geostatistical sampling workflow within **geosample** package. **D1**: user decision for *initial design*. **D2**: user decision whether to sample additional samples, in which case adaptive sample will be generated. **D3**: user decision to update sampling constraints. **D4**: user decision to stop further sampling. See text for detailed explanation.

3.2. Simulation example

In this example, we generated a binomial dataset available in the package as `sim.data`. We generated a realisation of Gaussian process $S(x)$ on a 35 by 35 grid covering the unit square, giving a total of $n^* = 1225$ potential sampling locations. We specified $S(x)$ to have expectation $\mu = 0$, variance $\sigma^2 = 1$ and Matérn correlation function (3), with $\phi = 0.15$ and $\kappa = 1.5$, and no measurement error, i.e. $\tau^2 = 0$. Binomial observations, with 8 trials at each grid point and

probabilities given by the anti-logit of the simulated values of the Gaussian process, constitute the response variable y. For the initial sample, we use a simple inhibitory design to sample $n_0 = 30$ locations with $\delta = 0.04$. The results are shown in Figure 2.

```
library("geosample")
library("viridisLite")
data(sim.data)
head(sim.data, n = 6L, addrownums = TRUE)
```

```
## Simple feature collection with 6 features and 3 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: 0 ymin: 0 xmax: 0.1471 ymax: 0
## CRS:           NA
##    data y units.m          geometry
## 1 1.042 4       8       POINT (0 0)
## 2 1.126 5       8 POINT (0.02941 0)
## 3 1.183 6       8 POINT (0.05882 0)
## 4 1.185 7       8 POINT (0.08824 0)
## 5 1.131 5       8  POINT (0.1176 0)
## 6 1.088 5       8  POINT (0.1471 0)
```

```
set.seed(123)
my.sample <- discrete.inhibit.sample(obj = sim.data, size = 30,
                                      delta = 0.04, plotit = TRUE)
```

The first argument in the function `discrete.inhibit.sample` specifies a spatial object i.e. **sf** or **sp** object in which each row contains a spatial location and any associated covariates. Sample size is specified via the argument `size`. Inhibition distance is set via `delta`. Sampled locations are plotted by default, whilst setting the argument `plotit` to `FALSE` turns the plotting off.

For both model parameter estimation and spatial predictions, we use functions from the **PrevMap** package (Giorgi and Diggle 2017). The `binomial.logistic.MCML` function fits a geostatistical binomial logistic model with the following inputs: random variables Y_i of positive counts, binomial denominators m_i, explanatory variables $d_i \in \mathbb{R}^p$ and associated sampling locations $x_i : i = 1, \ldots, n$ in the study region. Conditionally on a zero-mean Gaussian process $S(x)$ and mutually independent zero-mean Gaussian variables Z_i, each Y_i follows a binomial distribution with mean $E\{Y_i|S(x_i), Z_i\} = m_i p_i$ and

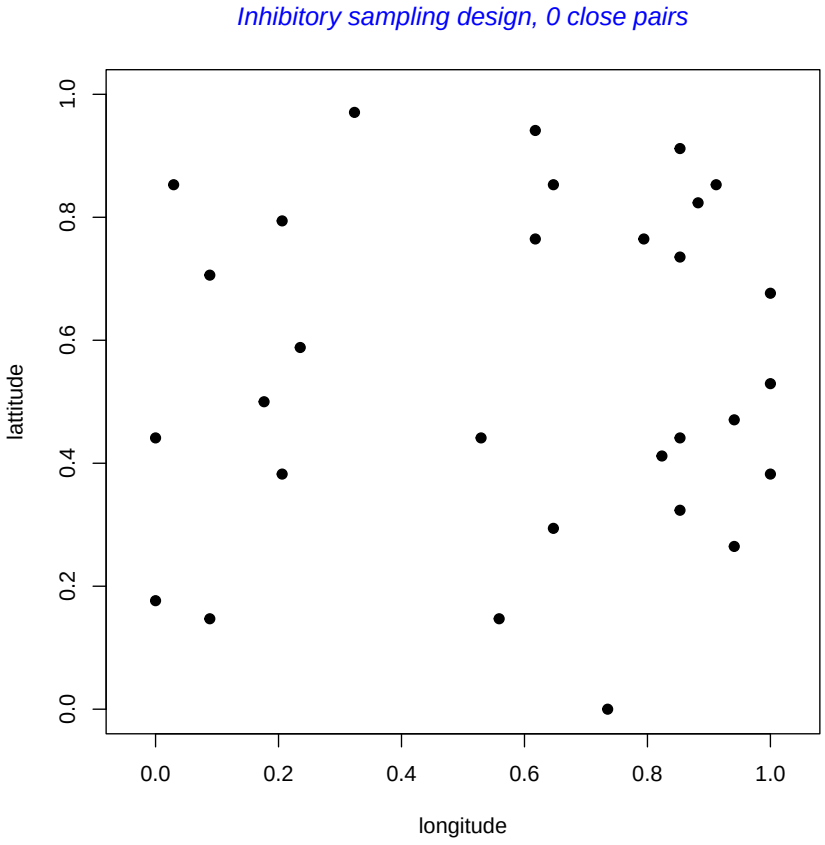

Figure 2: Simple inhibitory (discrete) design with $\delta = 0.04$ and $n_0 = 30$.

$$\log\left\{\frac{p}{1-p}\right\} = d(x_i)'\beta + S(x_i) + Z_i. \tag{4}$$

```
library("PrevMap")
knots <- as.matrix(expand.grid(seq(-0.2, 1.2, length = 15),
                               seq(-0.2, 1.2, length = 15)))
mcmc.ctr <- control.mcmc.MCML(n.sim = 5500, burnin=500, thin = 5)
dat <- my.sample[[4]]
par0 <- c(0.001, 1, 0.4)
model.fit <-
  binomial.logistic.MCML(y ~ 1, units.m = ~units.m, data = dat, par0 = par0,
                         coords=~st_coordinates(dat), fixed.rel.nugget = 0,
                         start.cov.pars = par0[3], control.mcmc = mcmc.ctr,
                         low.rank = TRUE, knots = knots, kappa = 1.5,
                         method = "BFGS", messages = FALSE,
                         plot.correlogram = FALSE)

summary(model.fit, log.cov.pars = FALSE)
```

```
## Geostatistical binomial model
## (low-rank approximation)
## Call:
## binomial.logistic.MCML(formula = y ~ 1, units.m = ~units.m, coords = ~st_coordinates(da
##     data = dat, par0 = par0, control.mcmc = mcmc.ctr, kappa = 1.5,
##     fixed.rel.nugget = 0, start.cov.pars = par0[3], method = "BFGS",
##     low.rank = TRUE, knots = knots, messages = FALSE, plot.correlogram = FALSE)
##
##             Estimate StdErr z.value p.value
## (Intercept)    0.461  0.179    2.57    0.01 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Objective function: 5.111
##
## Matern kernel parameters (kappa=1.5)
## Adjustment factorfor sigma^2: 7.761
##         Estimate StdErr
## sigma^2    0.786   0.13
## phi        0.859   0.28
```

```
##
## Legend:
## sigma^2 = variance of the Gaussian process
## phi = scale of the spatial correlation
```

We use the resulting binomial fit to generate spatial predictions of prevalence at each of the 1225 sampling locations using the `spatial.pred.binomial.MCML` function.

```
model.pred <-
  spatial.pred.binomial.MCML(object = model.fit, type = "joint",
                             control.mcmc = mcmc.ctr,thresholds = 0.45,
                             grid.pred = st_coordinates(sim.data),
                             scale.predictions = "prevalence",
                             scale.thresholds ="prevalence",
                             standard.errors = TRUE, messages = FALSE,
                             plot.correlogram = FALSE)
```

Several results can be summarised and visualised from the prediction results, including *predictions* and *exceedance probabilities* at each of the prediction locations.

```
par(mfrow = c(1,2))
plot(model.pred, type = "prevalence", col = viridis(256, direction = -1),
     summary = "predictions", zlim = c(0, 1))
contour(model.pred, type="prevalence", summary="predictions", zlim = c(0, 1),
        levels = seq(0.1,0.9, 0.1), add = TRUE)
plot(model.pred,summary="exceedance.prob",zlim=c(0,1),
     col = viridis(256, direction = -1))
contour(model.pred, summary = "exceedance.prob",zlim = c(0, 1),
        levels = seq(0.1,0.3, 0.1), add = TRUE)
```

```
par(mfrow = c(1,1))
```

To implement a *minimum distance batch adaptive sampling* of 10 additional locations, using the *prediction variance* selection criterion, we extract prediction variances at all potential locations. We set the minimum sampling distance to be $\delta = 0.1$.

```
obj.1 <- as.data.frame(cbind(model.pred$grid,
         c(model.pred$prevalence$standard.errors)^2))
colnames(obj.1) <- c("coord1", "coord2", "pred.var")
obj.1 <- sf::st_as_sf(obj.1, coords = c('coord1', 'coord2'))
```

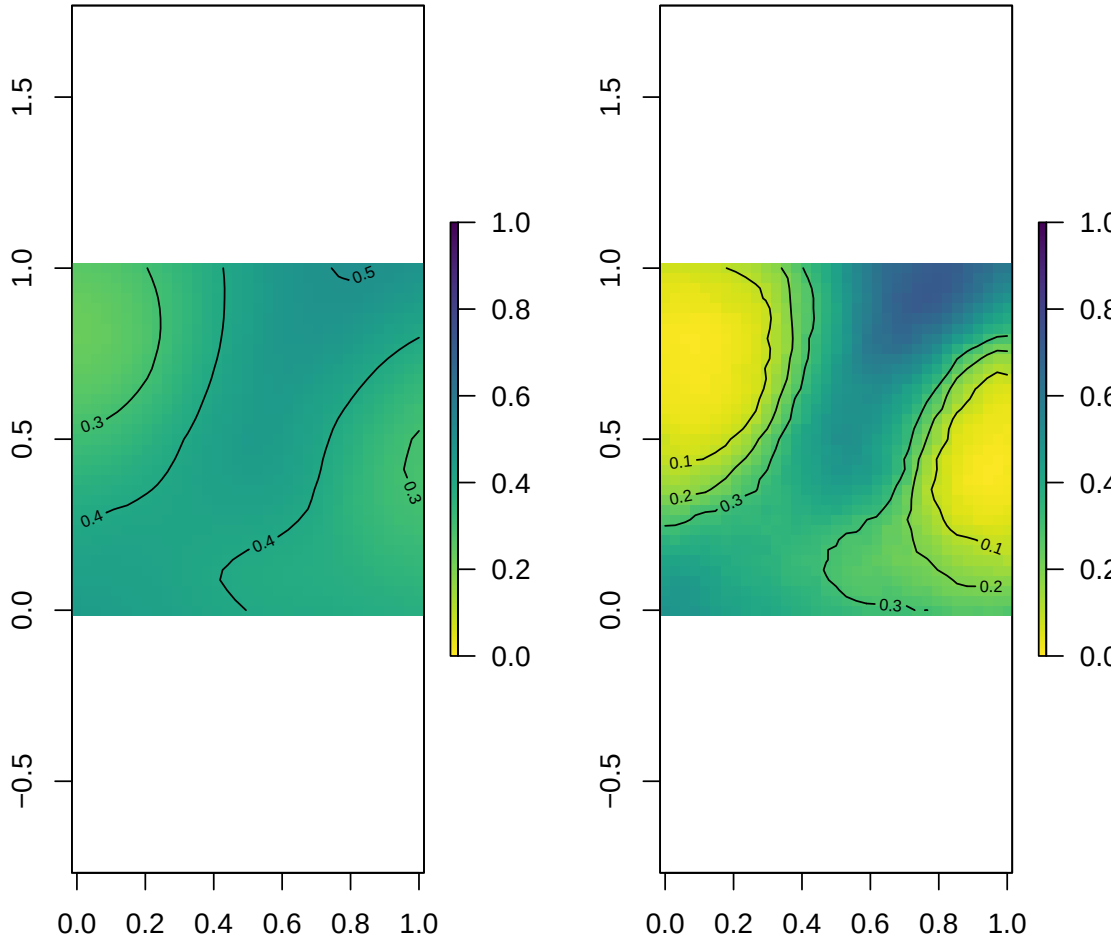

Figure 3: Spatial prediction visualisation. Spatial predictions on the LHS and exceedance probabilities $P(x; 0.45) = P(prev > 0.45 \text{ at location } x)$ on the RHS.

```
adapt.sample.pv <-
  adaptive.sample(obj1 = obj.1, obj2 = dat,
                  pred.var.col = 1,criterion = "predvar",
                  delta = 0.1, batch.size = 10, poly = NULL,
                  plotit = TRUE)
```

The argument `obj1` specifies a spatial object that contains potential sampling locations and their associated *prediction variance* and/or *exceedance probabilities*. Locations from the existing (initial) design are specified via argument `obj2`, which is also a spatial object such as **sf** or **sp** object. The `batch.size` determines the number of additional locations to be sampled per sampling round. A batch size equal to 1 will implement a *singleton adaptive design*. The function has a default behaviour to plot sample locations. These are shown in Figure 4.

Note that similar/comparable *parameter estimation* and *spatial prediction* results can be obtained from several other R packages. The choice of which package to use depends on a number of factors including, methodological implementation in the packages, analysis objective(s) and ease of use by the user. These packages include **geoRglm, geostatsp, geoBayes, spBayes, spGLM, spaMM, spMvGLM** and **geoCount** for count data. See https://cran.r-project.org/web/views/Spatial.html for a comprehensive list.

3.3. Case study: malaria prevalence in Majete, southern Malawi.

We now illustrate the use of the **geosample** package to construct a survey sample for malaria prevalence mapping in an area surrounding Majete Wildlife Reserve (MWR) within Chikwawa district, southern Malawi. The MWR is situated in the lower Shire valley at the edge of the African Rift Valley (15.97°S; 34.76°E). The whole perimeter is home to a population of around 100,000 (at the time of writing). Figure 5 shows the households of the study area. The perimeter is subdivided into 19 community-based organizations (CBOs). In the study, three sets of these CBOs (CBOs - 1 & 2, CBOs -15 & 16, and CBOs - 6, 7 & 8) define *focal areas* A, B, and C, respectively. See Chipeta *et al.* (2016a,b); Kabaghe, Chipeta, McCann, Phiri, van Vugt, Takken, Diggle, and Terlouw (2017); McCann, van den Berg, Diggle, van Vugt, Terlouw, Phiri, Di Pasquale, Maire, Gowelo, Mburu, Kabaghe, Mzilahowa, Chipeta, and Takken (2017) for more details.

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 devices on a Samsung Galaxy Tab 3 running the Android 4.1 Jellybean operating system. We consider *focal area* A of the study area and use a simple inhibitory design to sample 60 households in

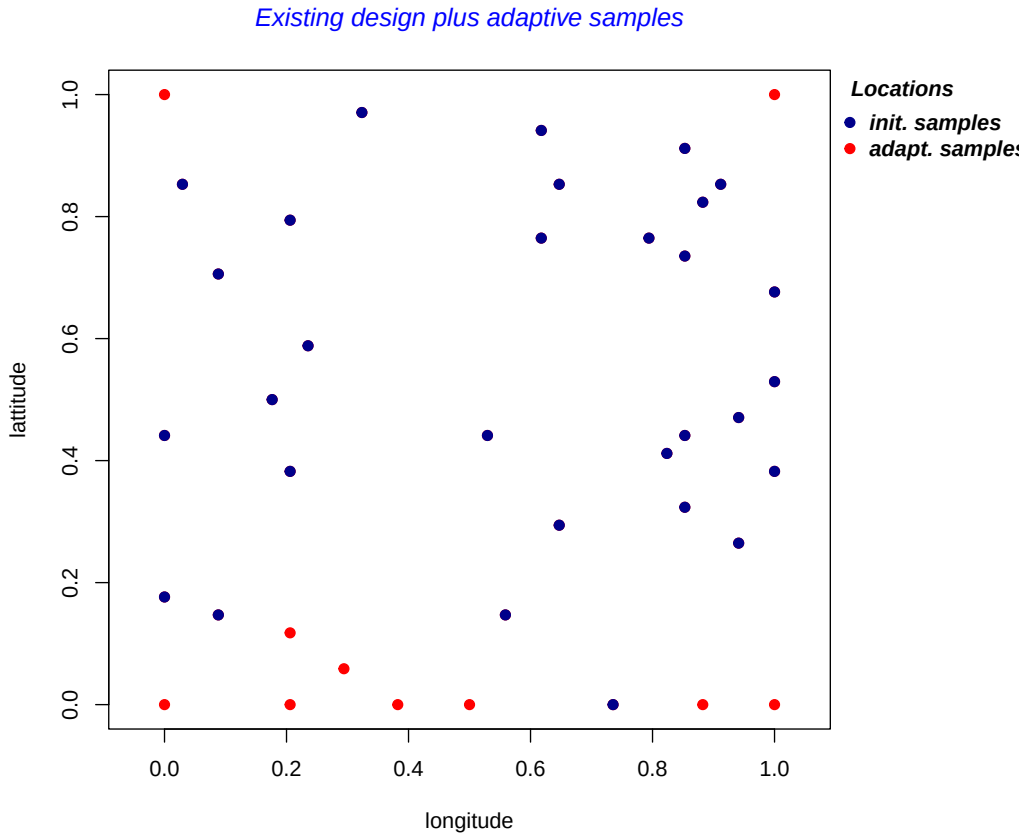

Figure 4: Adaptive sampling design with $\delta = 0.1$ and $b = 10$, Dark blue dots ($n_0 = 30$) are the initial sampling locations. Red dots ($n_a = 10$) are adaptive sampling locations added after analysing data from the initial design.

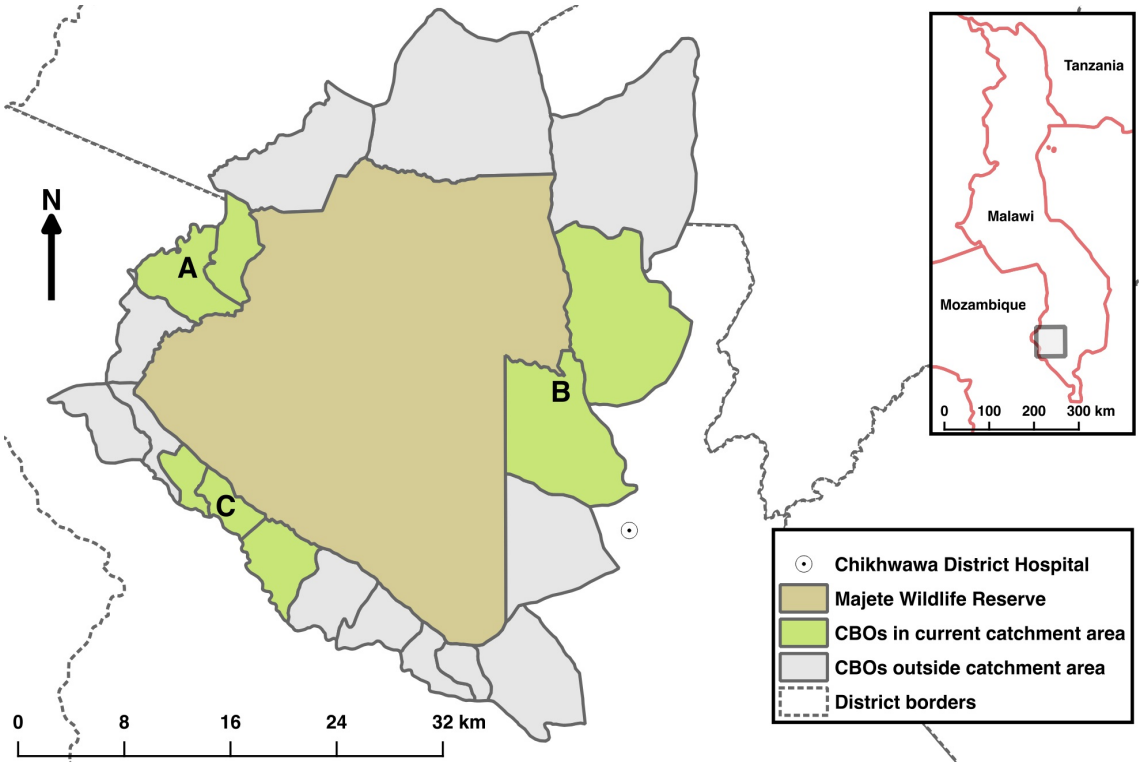

Figure 5: Majete Wildlife Reserve (brown) is surrounded by 19 CBOs (grey and green) comprising the Majete perimeter. Three focal areas (green), labelled as A, B, and C mark the communities selected for malaria indicator surveys. The rest of the CBOs (grey) are outside the project's catchment area. Reprinted from Kabaghe *et al.* (2017).

the initial sample. Data from these households are then analysed using the binomial logistic model (4), and predictive analysis is carried out to map malaria prevalence.

All potential (available) household locations are shown in Figure 6.

```
data("border")
data("majete")
plot(st_geometry(majete), pch = 19, cex = 0.5,
     xlim=range(st_coordinates(border)[,1]),
     ylim=range(st_coordinates(border)[,2]),
     axes = TRUE, xlab="longitude", ylab="latitude")

## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()

plot(border, lwd = 2, add= TRUE)
```

```
## old-style crs object detected; please recreate object with a recent sf::st_crs()
```

The sampled households (black dots) are shown in Figure 7.

```
set.seed(1234)
init.sample <-
  discrete.inhibit.sample(obj = majete, size = 60, delta = 0.4,
                          k = 0,  delta.fix = FALSE,
                          poly = border, plotit = TRUE)
```

```
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
```

We extract data from the sampled households, and fit the binomial logistic model (4) to the data. The binomial logistic estimates are then used for prediction at unobserved households. See Giorgi and Diggle (2017) for details.

```
mrdt <- init.sample[[4]]
glmfit <- glm(rdt~1, data = mrdt,
              family = binomial(link = logit))
ID.coords <- create.ID.coords(data=as.data.frame(mrdt),
                              coords=~st_coordinates(mrdt))
mrdt$units.m <- rep(1,nrow(mrdt))

par0 <- c(coef(glmfit),cov.pars=c(0.93171,3.9549))
control.mcmc <- control.mcmc.MCML(n.sim = 5500,burnin=500,thin=5)
model.fit <-
  binomial.logistic.MCML(rdt~1, units.m=~units.m,par0=par0,
                         coords=~st_coordinates(mrdt),data=mrdt,
                         ID.coords = ID.coords, kappa=0.5,
                         control.mcmc=control.mcmc, method="BFGS",
```

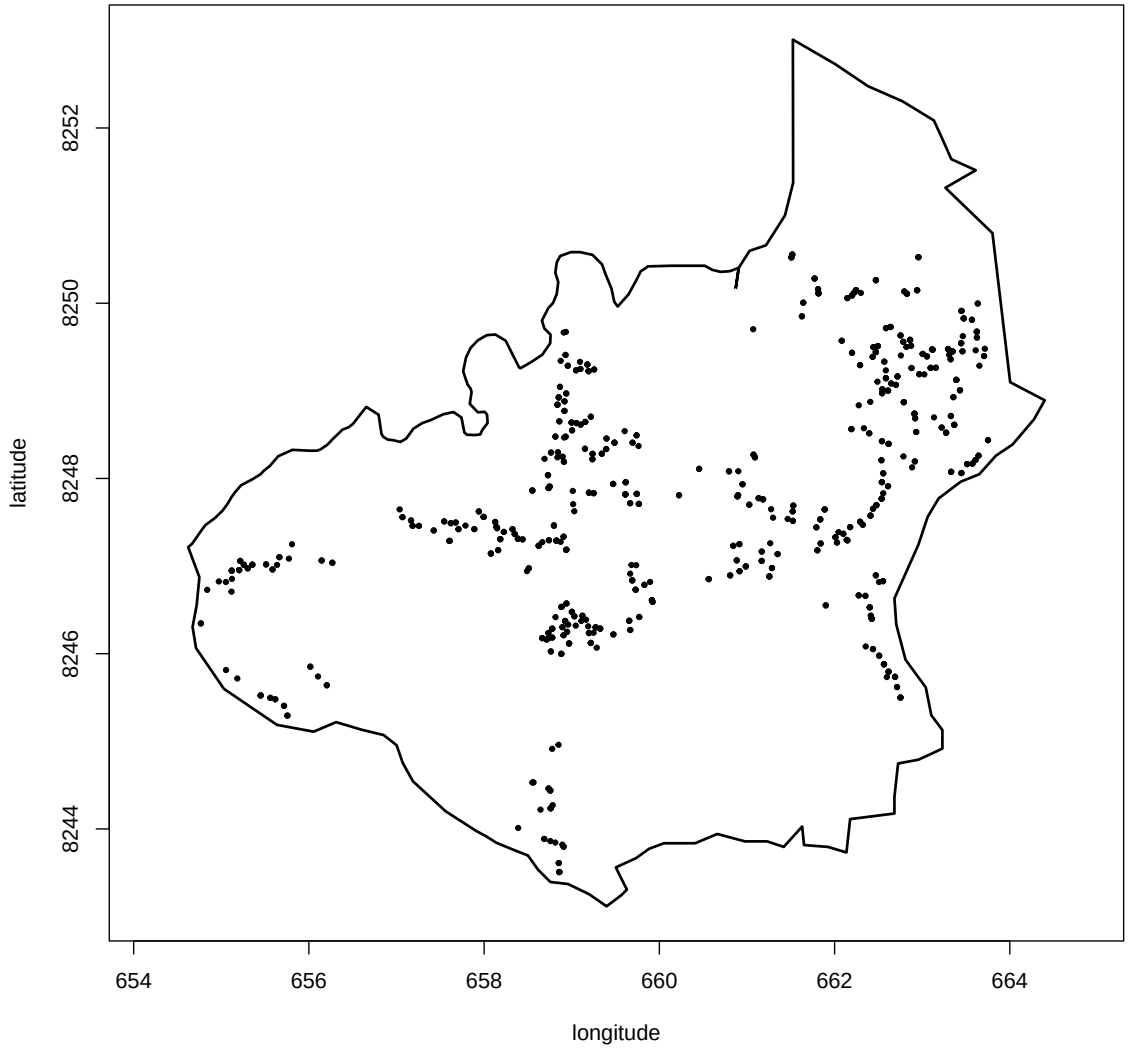

Figure 6: All potential household sampling locations in Majete.

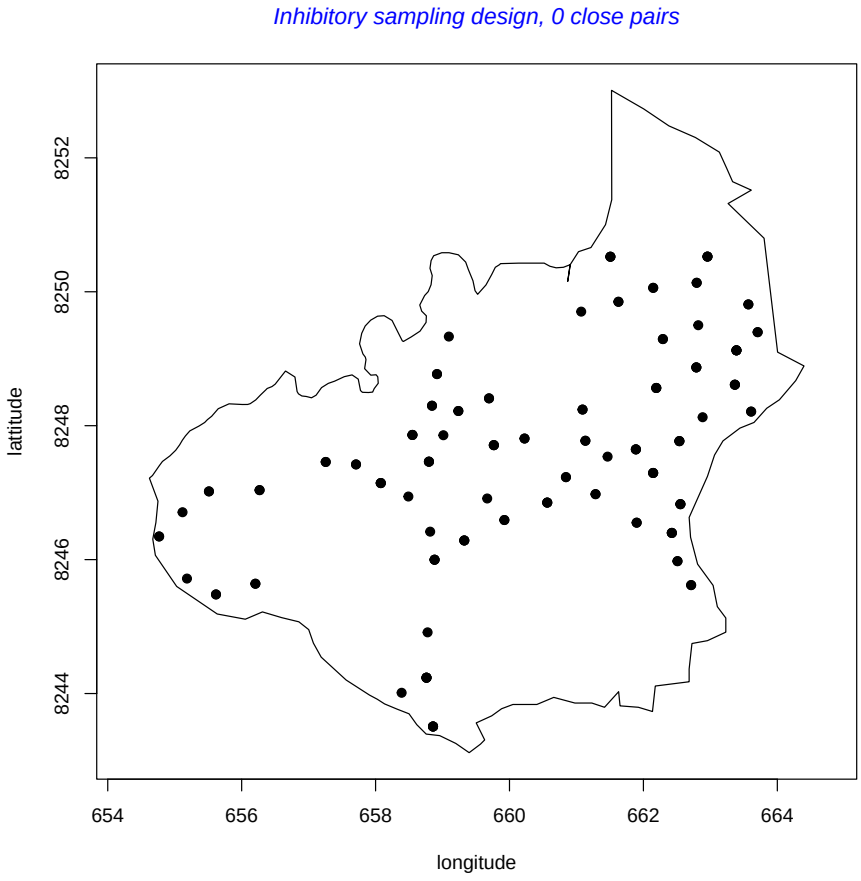

Figure 7: Simple inhibitory (discrete) design with $\delta = 400$ meters and $n_0 = 60$ households (black dots) in Majete.

```
                        fixed.rel.nugget = 0, start.cov.pars=c(par0[3]),
                        messages = FALSE, plot.correlogram = FALSE)
## Fixed relative variance of the nugget effect: 0
summary(model.fit, log.cov.pars = FALSE)
## Geostatistical binomial model
## Call:
## binomial.logistic.MCML(formula = rdt ~ 1, units.m = ~units.m,
##     coords = ~st_coordinates(mrdt), data = mrdt, ID.coords = ID.coords,
##     par0 = par0, control.mcmc = control.mcmc, kappa = 0.5, fixed.rel.nugget = 0,
##     start.cov.pars = c(par0[3]), method = "BFGS", messages = FALSE,
##     plot.correlogram = FALSE)
##
##             Estimate StdErr z.value p.value
## (Intercept)   -1.741  0.276    -6.3   3e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Objective function: 1.898
##
## Covariance parameters Matern function
## (fixed relative variance tau^2/sigma^2= 0)
##         Estimate StdErr
## sigma^2    0.148   2.21
## phi        0.725   0.63
##
## Legend:
## sigma^2 = variance of the Gaussian process
## phi = scale of the spatial correlation
```

We now carry out spatial predictions over all unobserved households, with the model parameters fixed at the MCML estimates, and summarise the predictive distribution of prevalence at each location through its mean, standard deviation and probability that the estimated prevalence is above 15 %. Using these results, an adaptive sample of 40 additional households is taken. The results are shown in Figure 8.

```
avail.locs <- majete[!(majete$geometry) %in% (mrdt$geometry),]
```

```
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
```

```
model.pred <-
  spatial.pred.binomial.MCML(model.fit,
                             grid.pred=unique(st_coordinates(avail.locs)),
                             control.mcmc=control.mcmc, type = "marginal",
                             scale.predictions = "prevalence",
                             standard.errors = TRUE, thresholds = 0.15,
                             scale.thresholds = "prevalence",
                             messages = FALSE, plot.correlogram = FALSE)

pred.vars <- as.data.frame(cbind(model.pred$grid,
                                 c(model.pred$prevalence$standard.errors)^2))

colnames(pred.vars)<- c("coord1", "coord2", "pred.var")
pred.vars <- sf::st_as_sf(pred.vars, coords = c('coord1', 'coord2'))
st_crs(pred.vars) <- st_crs(mrdt)
adapt.sample.pv <-
   adaptive.sample(obj1 = pred.vars, obj2 = mrdt,
                   pred.var.col = 1, criterion = "predvar",
                   delta = 0.15, batch.size = 40,
                   poly = border, plotit = TRUE)
```

```
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
```

Model parameter estimates are updated using the augmented data from the adaptive sampling.

```
mrdt <- majete[(majete$geometry) %in%
                 (adapt.sample.pv$sample.locs$curr.sample$geometry),]
```

```
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
```

```
ID.coords <- create.ID.coords(data=as.data.frame(mrdt),
                              coords=~st_coordinates(mrdt))
mrdt$units.m <- rep(1,nrow(mrdt))
```

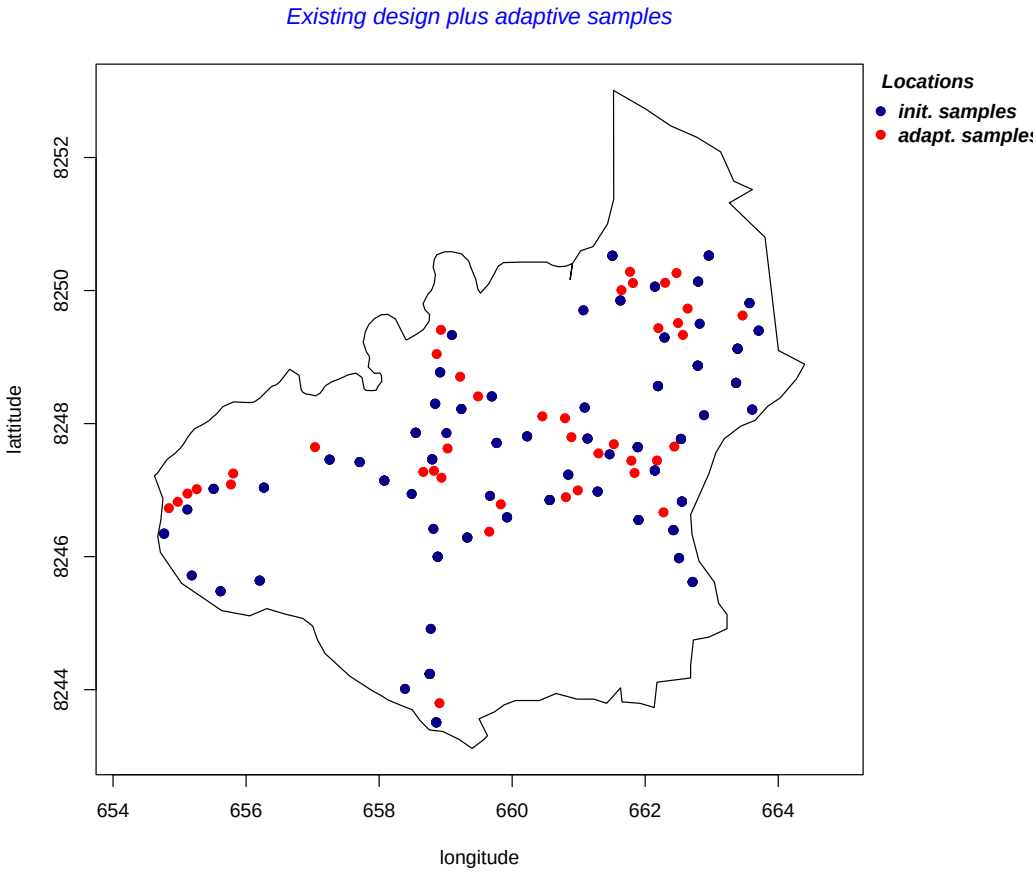

Figure 8: Adaptive sampling design with $\delta = 150$ meters and $b = 40$, Blue dots ($n_0 = 60$) are the initial sampling households. Red dots ($n_a = 40$) are adaptive samples added after analysing data from the initial design.

```
par0 <- c(coef(model.fit))
model.fit <-
  binomial.logistic.MCML(rdt~1, units.m=~units.m,par0=par0,
                         coords=~st_coordinates(mrdt),data=mrdt,
                         ID.coords = ID.coords,
                         control.mcmc=control.mcmc, kappa=0.5,
                         fixed.rel.nugget = 0,
                         start.cov.pars=c(par0[3]),
                         method="BFGS", messages = FALSE,
                         plot.correlogram = FALSE)

## Fixed relative variance of the nugget effect: 0

summary(model.fit, log.cov.pars = FALSE)

## Geostatistical binomial model
## Call:
## binomial.logistic.MCML(formula = rdt ~ 1, units.m = ~units.m,
##     coords = ~st_coordinates(mrdt), data = mrdt, ID.coords = ID.coords,
##     par0 = par0, control.mcmc = control.mcmc, kappa = 0.5, fixed.rel.nugget = 0,
##     start.cov.pars = c(par0[3]), method = "BFGS", messages = FALSE,
##     plot.correlogram = FALSE)
##
##             Estimate StdErr z.value p.value
## (Intercept)   -1.459  0.205   -7.12 1.1e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Objective function: 1.57
##
## Covariance parameters Matern function
## (fixed relative variance tau^2/sigma^2= 0)
##         Estimate StdErr
## sigma^2    0.226   1.78
## phi        0.916   0.55
##
## Legend:
## sigma^2 = variance of the Gaussian process
## phi = scale of the spatial correlation
```

We now carry out spatial predictions over a 5 metre by 5 metre regular grid, with model parameters fixed at the MCML estimates from the accrued data, and summarise the predictive distribution of prevalence in each grid cell through its mean, standard deviation and probability that the estimated prevalence is above 15 %.

```
library(splancs)

##
## Spatial Point Pattern Analysis Code in S-Plus
##
##  Version 2 - Spatial and Space-Time analysis

##
## Attaching package:  'splancs'

## The following object is masked from 'package:raster':
##
##     zoom

pred.poly <- as_Spatial(border)@polygons[[1]]@Polygons[[1]]@coords

## old-style crs object detected; please recreate object with a recent sf::st_crs()

## old-style crs object detected; please recreate object with a recent sf::st_crs()

grid.pred <- gridpts(pred.poly, xs=0.05, ys=0.05)

model.pred <-
  spatial.pred.binomial.MCML(model.fit, grid.pred=grid.pred,
                             control.mcmc=control.mcmc,
                             type = "marginal",
                             scale.predictions = "prevalence",
                             standard.errors = TRUE, thresholds = 0.15,
                             scale.thresholds = "prevalence",
                             messages = FALSE, plot.correlogram = FALSE)
##1. Prevalence predictions
prevpred <-
  rasterFromXYZ(cbind(model.pred$grid[,1],
                      model.pred$grid[,2],
                      model.pred$prevalence$predictions))
prevpred <- raster::disaggregate(prevpred, fact = 10,
                                 method = "bilinear")
```

```
##2. Std error
stderror <-
  rasterFromXYZ(cbind(model.pred$grid[,1],
                      model.pred$grid[,2],
                      model.pred$prevalence$standard.errors))
stderror <- raster::disaggregate(stderror, fact = 10,
                                 method = "bilinear")

##3. Exceedance probablities
exceed <-
  rasterFromXYZ(cbind(model.pred$grid[,1],
                      model.pred$grid[,2],
                      model.pred$exceedance.prob))
exceed <- raster::disaggregate(exceed, fact = 10,
                                 method = "bilinear")
par(mfrow = c(2,2))
plot(prevpred, main = "(a)", col = viridis(256, direction = -1))
plot(exceed, main="(b)", zlim = c(0,1), col = viridis(256, direction = -1))
plot(stderror, main = "(c)", col = viridis(256, direction = -1))
par(mfrow = c(1,1))
```

4. Conclusions and future developments

We have demonstrated the use of the **geosample** package for geostatistical sampling of spatially referenced data. The package is compatible with existing R packages for parameter estimation and predictive inference. It uses novel and computationally efficient algorithms for constructing adaptive and non-adaptive geostatistical designs, including traditional random sampling. The package also provides automatic visualisation of the results by plotting the sampled locations as illustrated in Figures 2 and 4. When sampling is only possible at a pre-determined set of locations, for example households within a community or communities within a region, the package requires that all such potential sampling locations are available in georeferenced form.

In the adaptive case, the package offers the user a choice between two design selecton criteria: *prediction variance* and *exceedance probability*. We plan to add flexibility to this aspect of the package by allowing the user to define their own criterion.

We also plan to incorporate costs associated with travelling between any two potential sampling locations. Given a cost matrix, *least-cost path* (LCP) selection criterion would identify the

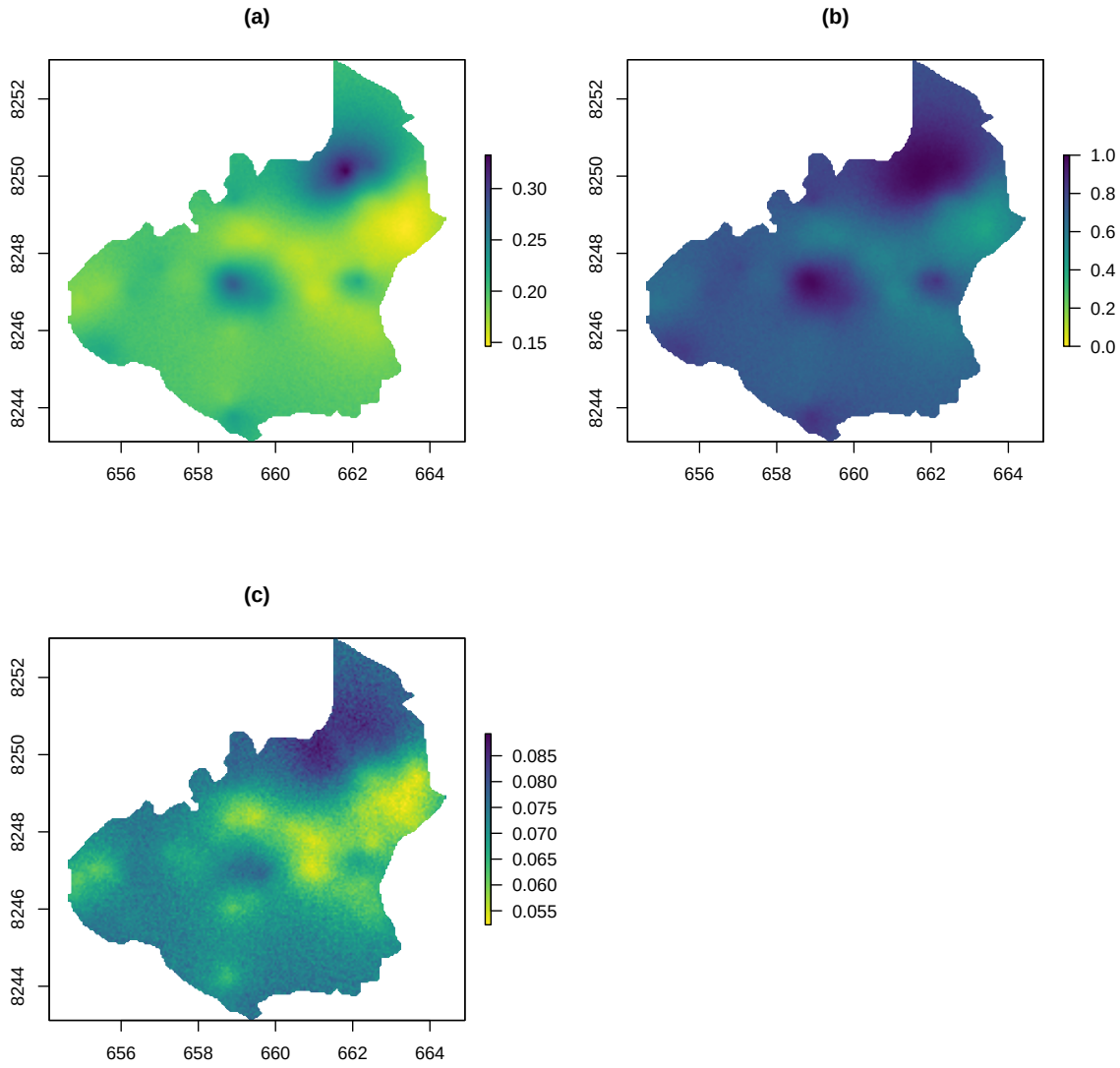

Figure 9: (a) Malaria prevalence in Majete. (b) Exceedance probabilities $P(x; 0.15)$ for the predictions. $P(x; 0.15) = P(prev > 0.15$ at location $x)$. (c) Standard errors of predictions.

most economical path of travel (Adriaensen, Chardon, De Blust, Swinnen, Villalba, Gulinck, and Matthysen 2003), which could be balanced against statistical efficiency so as to give an optimal design for fixed total cost, rather than for fixed total sample size. In contexts like our example of malaria prevalence mapping, an appropriate cost matrix might need to take account of distance, terrain and predicted travel times/speeds (Driezen, Adriaensen, Rondinini, Doncaster, and Matthysen 2007; Houben, Van Boeckel, Mwinuka, Mzumara, Branson, Linard, Chimbwandira, French, Glynn, and Crampin 2012; Li, Li, Li, Qiao, Yang, and Zhang 2010).

A third extension is to relax the requirement for all potential sampling locations to be georeferenced beforehand. In our example of malaria prevalence mapping for the Majete study this involved substantial effort in the field. For prevalence mapping at larger geographical scales, the corresponding effort would have been prohibitive. One approach that we plan to investigate is to use a two-stage stratified sampling procedure, in which the study area is divided into a large number of strata, for example administrative units. A suitable design strategy might then be first to sample strata using a convenient reference location for each stratum, for example its centroid, then to georeference all potential sampling units within each sampled stratum.

We will report these extensions separately in due course.

Acknowledgements

We thank Majete Malaria Project (MMP) for allowing us to use part of the project data in illustrating the **geosample** package functionality. M G Chipeta was supported by the Malawi-Liverpool Wellcome Trust/Lancaster University post-doctoral training fellowship. P J Diggle is supported by the MMP grant funded by the Dioraphte foundation, Netherlands. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders. We thank all the contributors and reviewers. This work utilises a number of independent R extensions, including **splancs** (Rowlingson and Diggle 2017), **pdist** (Wong 2013), **dplyr** (Wickham, Francois, Henry, and Muller 2018), **PrevMap** (Giorgi and Diggle 2017), **geoR** (Ribeiro Jr. and Diggle 2016), **sp** (Pebesma and Bivand 2005) and **sf** (Pebesma 2018).

References

Adriaensen F, Chardon JP, De Blust G, Swinnen E, Villalba S, Gulinck H, Matthysen E (2003). "The application of 'least-cost' modelling as a functional landscape model." *Landscape and Urban Planning*, **64**(4), 233–247. ISSN 01692046. doi:10.1016/S0169-2046(02)00242-6. 1132.

Chipeta MG, Terlouw DJ, Phiri KS, Diggle PJ (2016a). "Adaptive geostatistical design and analysis for prevalence surveys." *Spatial Statistics*, **15**, 70–84. ISSN 22116753. doi:10.1016/j.spasta.2015.12.004. URL http://linkinghub.elsevier.com/retrieve/pii/S2211675315001153.

Chipeta MG, Terlouw DJ, Phiri KS, Diggle PJ (2016b). "Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure." *Enviromentrics*, **28**(1), 1–11. ISSN 1099-095X. doi:DOI10.1002/env.2425. URL http://dx.doi.org/10.1002/env.2425http://arxiv.org/abs/1605.00104.

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

Cochran WG (1977). *Sampling Techniques*. 3 edition. John Wiley & Sons, Ltd., New York.

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

Diggle PJ, Menezes R, Su TL (2010). "Geostatistical inference under preferential sampling." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**(2), 191–232.

Diggle PJ, Ribeiro JP (2007). *Model-based Geostatistics*. Springer, New York.

Driezen K, Adriaensen F, Rondinini C, Doncaster CP, Matthysen E (2007). "Evaluating least-cost model predictions with empirical dispersal data: A case-study using radiotracking data of hedgehogs (Erinaceus europaeus)." *Ecological Modelling*, **209**(2-4), 314–322. ISSN 03043800. doi:10.1016/j.ecolmodel.2007.07.002.

Giorgi E, Diggle PJ (2017). "PrevMap : an R Package for Prevalence Mapping." *Journal of Statistical Software*, **78**, 1–29.

Houben RMGJ, Van Boeckel TP, Mwinuka V, Mzumara P, Branson K, Linard C, Chimbwandira F, French N, Glynn JR, Crampin AC (2012). "Monitoring the impact of decentralised chronic care services on patient travel time in rural Africa - methods and results in Northern Malawi." *International Journal of Health Geographics*, **11**(1), 1. ISSN 1476072X. doi:10.1186/1476-072X-11-49. URL InternationalJournalofHealthGeographicsInternationalJournalofHealthGeographics.

Kabaghe AN, Chipeta MG, McCann RS, Phiri KS, van Vugt M, Takken W, Diggle PJ, Terlouw AD (2017). "Adaptive geostatistical sampling enables efficient identification of malaria hotspots in repeated cross-sectional surveys in rural Malawi." *Plos One*, **12**(2), e0172266. ISSN 1932-6203. doi:10.1371/journal.pone.0172266. URL http://dx.plos.org/10.1371/journal.pone.0172266.

Lark RM (2002). "Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood." *Geoderma*, **105**(1-2), 49–80. ISSN 00167061. doi:10.1016/S0016-7061(01)00092-1. URL http://linkinghub.elsevier.com/retrieve/pii/S0016706101000921.

Li H, Li D, Li T, Qiao Q, Yang J, Zhang H (2010). "Application of least-cost path model to identify a giant panda dispersal corridor network after the Wenchuan earthquake-Case study of Wolong Nature Reserve in China." *Ecological Modelling*, **221**(6), 944–952. ISSN 03043800. doi:10.1016/j.ecolmodel.2009.12.006. URL http://dx.doi.org/10.1016/j.ecolmodel.2009.12.006.

Mardia KV, Marshall RJ (1984). "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression." *Biometrika*, **71**(1), 135–146.

Matérn B (1986). *Spatial Variation*. 2 edition. Springer, Berlin.

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

McBratney AB, Webster R, Burgess TM (1981). "The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalized Variables–I: Theory and method." *Computers & Geosciences*, **7**(4), 331–334.

McCann R, van den Berg H, Diggle P, van Vugt M, Terlouw D, Phiri K, Di Pasquale A, Maire N, Gowelo S, Mburu M, Kabaghe A, Mzilahowa T, Chipeta M, Takken W (2017). "Assessment of the effect of larval source management and house improvement on malaria transmission when added to standard malaria control strategies in southern Malawi: Study protocol for a cluster-randomised controlled trial." *BMC Infectious Diseases*, **17**(1). ISSN 14712334. doi:10.1186/s12879-017-2749-2.

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

Müller WG, Pronzato L, Rendas J, Waldl H (2015). "Efficient prediction designs for random fields." *Applied Stochastic Models in Business and Industry*, **31**(2), 178–194. ISSN 15264025. doi:10.1002/asmb.2084.

Müller WG, Zimmerman DL (1999). "Optimal designs for Variogram estimation." *Environmentrics*, **10**, 23–37.

Pebesma E (2018). "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal*. URL https://journal.r-project.org/archive/2018/RJ-2018-009/index.html.

Pebesma EJ, Bivand RS (2005). "Classes and methods for spatial data in R." URL https://cran.r-project.org/doc/Rnews/.

R Core Team (2017). "R: A Language and Environment for Statistical Computing." https://www.R-project.org/. URL https://www.r-project.org/.

Ribeiro Jr PJ, Diggle PJ (2016). "geoR: Analysis of Geostatistical Data." URL https://cran.r-project.org/package=geoR.

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

Rowlingson BS, Diggle PJ (2017). "splancs: Spatial and Space-Time Point Pattern Analysis." URL https://cran.r-project.org/package=splancs.

Russo D (1984). "Design of an Optimal Sampling Network for Estimating the Variogram." *Soil Science Society of America Journal*, **48**(4), 708–716. ISSN 0361-5995. doi:10.2136/sssaj1984.03615995004800040003x.

Warrick AW, Myers DE (1987). "Optimization of sampling locations for variogram calculations." *Water Resources Research*, **23**(3), 496–500. ISSN 0043-1397. doi:10.1029/WR023i003p00496.

Wickham H, Francois R, Henry L, Muller K (2018). "dplyr: A Grammar of Data Manipulation." URL https://cran.r-project.org/package=dplyr.

Wong J (2013). "pdist: Partitioned Distance Function." URL https://cran.r-project.org/package=pdist.

Yfantis EA, Flatman GT, Behar JV (1987). "Efficiency of Kriging Estimation for Square , Triangular , and Hexagonal Grids." *Mathematical Geology*, **19**(3), 183–205.

Affiliation:

Michael G Chipeta, Big Data Institute, University of Oxford, Oxford, OX3 7LF, UK
E-mail: mchipeta@mlw.mw