Efficient design of geographically-defined clusters with spatial autocorrelation

I. Watson, Samuel

Efficient design of geographically-defined clusters with spatial autocorrelation

Samuel I. Watson

Submit your article to this journal ⤤

View related articles ⤤

View Crossmark data ⤤

Taylor & Francis
Taylor & Francis Group

Efficient design of geographically-defined clusters with spatial autocorrelation

Samuel I. Watson

University of Birmingham, Birmingham, UK

ABSTRACT

Clusters form the basis of a number of research study designs including survey and experimental studies. Cluster-based designs can be less costly but also less efficient than individual-based designs due to correlation between individuals within the same cluster. Their design typically relies on *ad hoc* choices of correlation parameters, and is insensitive to variations in cluster design. This article examines how to efficiently design clusters where they are geographically defined by demarcating areas incorporating individuals and households or other units. Using geostatistical models for spatial autocorrelation, we generate approximations to within cluster average covariance in order to estimate the effective sample size given particular cluster design parameters. We show how the number of enumerated locations, cluster area, proportion sampled, and sampling method affect the efficiency of the design and consider the optimization problem of choosing the most efficient design subject to budgetary constraints. We also consider how the parameters from these approximations can be interpreted simply in terms of 'real-world' quantities and used in design analysis.

1. Introduction

Clusters form the basis of a variety of survey and experimental study designs. In a survey setting, multi-stage cluster sampling can offer a cost-effective alternative to simple random sampling when a study population is large or complete enumeration costly [14]. For a two-stage design, individual sampling units, such as individuals or households, are often grouped into clusters on a geographic basis. These clusters are sometimes referred to as primary sampling units (PSU) or enumeration areas and may be constituted by political divisions like towns, villages, or census tracts, or they may be geographically defined for convenience. Cluster randomised trials (cRCT) are an experimental study design in which interventions are applied to whole clusters, which, as with survey designs, can also include geographically-defined areas [12,31]. cRCTs are useful for evaluating processes applied at a 'higher level' than the individual or where interaction between individuals within a cluster is unavoidable.

**CONTACT** Samuel I. Watson ✉ s.i.watson@bham.ac.uk

In both survey and experimental studies with geographical clusters, individuals from each cluster are typically enumerated to form a sampling frame and then sampled for inclusion and data capture. There is a variety of terminology used for cluster-based designs, often reflecting data collection or enumeration and sampling methodology. Indeed, these methods area also sometimes collectively referred to as 'area sampling'. In this article, we consider a cluster to be a set of individual sampling units grouped together because of their location in a geographically defined area for the purpose of enumeration, sampling, and data collection. We consider a geospatial statistical approach to the design and analysis of these clusters for the purposes of sampling individual units. In particular, for geographical clusters we must select: (i) the boundary of the cluster and hence area of the cluster; (ii) the proportion of the units in the cluster to be sampled; and (iii) the method of sampling (simple random or otherwise). Each of these steps can have an effect on the efficiency of the study design.

The potential benefit of a cluster-based study can be offset by the potential for reduced efficiency and the need for larger sample sizes to achieve the same level of precision. Should there exist correlation between individuals within the same cluster then the total amount of information a sample provides is reduced [21,35]. One means of assessing this loss of information is to calculate the *effective sample size*, which is the sample size of uncorrelated observations that would afford the same precision as the correlated sample. In many applications with clustered data, it is assumed that the covariance between observations in the same cluster is constant, such as in a linear mixed model framework where cluster means are assumed to be realisations of underlying normally-distributed latent variable [21,31,35]. In this case, the effective sample size is a function of the intraclass correlation coefficient (ICC), which is the proportion of the total variance attributable to the cluster-level. Sample sizes can then be inflated appropriately to provide a desired level of precision. However, this requires reliable estimates of the ICC.

There are relatively few studies providing estimates of ICCs from a wide range of settings for different variables. More commonly estimates exist from cRCTs rather than two-stage sampling schemes. For example, Campbell et al. [3] estimate a large number of ICCs using data from 'implementation science' cluster trials and report ICCs between 0.000 and 0.415, although the vast majority were below 0.1 with a median value below 0.05. A similar study of ICCs in cluster trials relating to heart failure reported ICCs between 0.026 and 0.052 [25]. However, these trials do not use geographically-defined clusters, rather organisations or institutions like hospitals and schools. For two-stage sampling designs, Janjua et al. [22] report ICCs of between 0.01 and 0.05 and Gulliford et al. [18] report the same range in a study of the Health Survey for England. The values 0.01 to 0.05 are representative of those used in practice for designing both cluster trials and sampling schemes [19].

The issue with using the ICC in this way for cluster-based designs is that it relies on the assumption of a constant covariance between individuals within a cluster and so it is unaffected by choices regarding the cluster design and sampling methodology. For geographically-defined clusters though, a likely explanation for within-cluster correlation is spatial autocorrelation, which exists for health (e.g. [26,30,36,37]), economic (e.g. [1,29]), and many other outcomes including research activity itself [13]. The archetypal spatially autocorrelated outcome is infectious disease which requires the interaction of individuals for transmission [33]. Given the nature of spatial autocorrelation, one can improve the efficiency of a cluster-based sampling scheme simply by enlarging clusters and sampling

m=100,n=20,R=1
simple random sampling
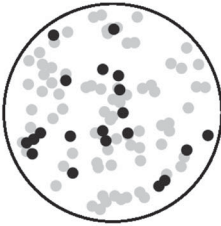
m=100,n=10,R=1
simple random sampling
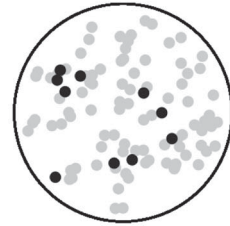
m=225,n=20,R=1.5
simple random sampling
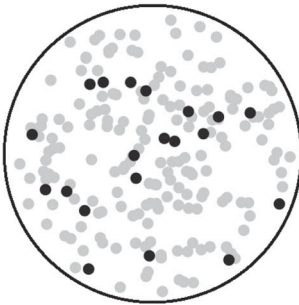
m=100,n=20,R=1
spatially–regular sampling
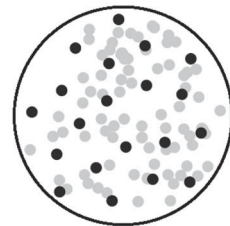
Figure 1. Illustrative example of variations in cluster design and sampling assuming a constant density of fixed locations, where *m* is the number of locations enumerated (grey points), *n* the number sampled (black points), and *R* the radius of the cluster.

them more sparsely so that sampled units are further apart, for example (see Figure 1). However, this may require the generation of prohibitively large sampling frames, a reduction in cluster numbers, or 'dilute' treatment effects from interventional studies. There is thus a trade-off between different cluster designs requiring statistical analysis. However, the reliance on an assumption of constant ICC for design purposes does not allow for this and little information exists on the sensitivity of an ICC to cluster design choices. In the absence of comprehensive estimates of ICCs for different sizes and types of cluster under different sampling schemes, it remains uncertain what effect altering the cluster design would have. Figure 1 illustrates some different design choices. We note that the questions we consider are distinct from the question of how to sample clusters from a set of candidates, which has been addressed elsewhere for example Grafström *et al.* [15], Grafström and Tillé [16], Benedetti *et al.* [2], as we consider how the design of a single geographic cluster affects the efficiency of the overall study design. We also focus on individual level rather than areal analysis.

If a model for spatial autocorrelation is assumed, then one can evaluate the effective sample size produced by a particular design on the basis of several empirically derived values including the distance between sampled points and information about the nature of

the spatial correlation [17]. Enlarging clusters can increase the average distance between sampled points and improve efficiency, but if the sample size or number of individuals or households that can be enumerated to generate the sample frame is fixed then efficiency is lost by having fewer clusters. The cost to enumerate ever large clusters may also become prohibitive. Efficiency can be considered both in a purely statistical sense (effective sample size) or a cost–benefit sense (effective sample size per unit time or money). It is from this perspective that we examine cluster design in this article.

2. The effective sample size under spatial autocorrelation

For a sampling design for N individuals, the loss of information due to any correlation between observations can be quantified by determining the effective sample size $N^* \leq N$, which is the sample size of uncorrelated observations that would afford the same precision as the correlated sample. This statistic is useful then to calculate the expected precision of a design in 'power' or other types of calculation.

Consider the linear model $Y = \mu + \Sigma^{\frac{1}{2}}\epsilon$ where Y is an $n \times 1$ vector of observations, μ is the population mean, ϵ is an independent and identically distributed $N(0, \sigma_\epsilon^2)$ random variable, and Σ contains information on the covariance structure so that the covariance matrix is given by $\Sigma\sigma_\epsilon^2$. In this context, Griffith [17] shows that the effective sample size is given by

$$N^* = \frac{\text{tr}(\Sigma)}{\mathbf{1}'\Sigma\mathbf{1}}N \tag{1}$$

where $\mathbf{1}$ is an $N \times 1$ vector of ones. For applications with simple clustering, such as a linear mixed model, the covariance matrix is typically assumed to have constant off-diagonal elements, σ_{xx}^2, for units in the same cluster and zero otherwise, and a variance of σ^2 on the diagonal. In this case, Equation (1) reduces to:

$$N^* = \frac{Jn}{1 + (n-1)\sigma_{xx}^2/\sigma^2} \tag{2}$$

where n is the average number of individuals in a cluster, J is the number of clusters, and σ_{xx}^2/σ^2 is the intraclass correlation coefficient (ICC). The quantity $1 + (n-1)\sigma_{xx}^2/\sigma^2$ is sometimes also referred to as the 'design effect' [35]. In the same way, one can derive the 'design effect' for various other covariance matrices with compound symmetry such as designs with repeated measures and different specifications of temporal autocorrelation. Hemming *et al.* [20] provide such a list in the context of cluster randomised trial design.

We can consider the effective sample size when the covariance between observations is instead determined by the spatial autocorrelation. We adopt the standard geostatistical assumptions that each observation with the cluster is a point sample measured with error from a latent continuous process across the area of interest [10]. Under these assumptions, the variogram is a widely used tool for estimating the relationship between distance between sampled locations and the degree of spatial autocorrelation [7]. In particular, it is defined as

$$\gamma(d) = \frac{1}{2}E\left[(Y(x+d) - Y(x))^2\right] \tag{3}$$

Table 1. Covariance functions and approximations to within-cluster mean covariance using the function $g(x) = 1 - 1/(1 + \alpha q_1^\beta)$ for sampling on a disc.

		Approx. mean covariance, $\tilde{s}$	
Model	$f(d, r)$	Simple random	Spatially-inhibited
Gaussian	$\exp(-d_{ij}^2/r^2)$	$\alpha = 0.910, \beta = 2.067$	$\alpha = 0.877, \beta = 2.110$
Exponential	$\exp(-d_{ij}/r)$	$\alpha = 0.768, \beta = 1.194$	$\alpha = 0.744, \beta = 1.207$
K-Bessel	$\dfrac{d_{ij}}{r} K_1\left(\dfrac{d_{ij}}{r}\right)$	$\alpha = 1.847, \beta = 1.577$	$\alpha = 1.801, \beta = 1.604$

where $Y(x)$ is the values of the random variable Y at location vector x and d is a spatial 'lag' or distance. The Euclidean distance is frequently used for these applications Chilès and Delfiner [7], which we assume throughout, although other measures could be used such as travel time, Manhattan distance, or kernal distance (e.g. [34]). The empirical variogram can be defined by three parameters: the nugget, range, and sill. The sill (C_1) is the maximum value of $\gamma(d)$ and is equivalent to the total variance of Y, the range r is the smallest value of d at which $\gamma(d)$ reaches the sill (or 95% of the sill for models where the sill is only asymptotically reached), and the nugget (C_0) is the variance at scales shorter than the smallest observed distance or at the same location, for example the variance we might expect among individuals sampled in the same household or next-door neighbours (i.e. $\gamma(0)$). Then, as Griffith [17] describes, the variogram form of Equation (1) is given by:

$$N^* = \frac{Jn}{1 + \frac{\rho}{n} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} f(d_{ij}, r)} \tag{4}$$

where $f(d_{ij}, r)$ is a given variogram model, d_{ij} denotes the distance separating the locations of observations i and j, and $\rho = (C_1 - C_0)/C_1$.

When the parameter ρ is zero then $C_0 = C_1$ and there is no spatial autocorrelation, and when ρ is one then C_0 is zero, which means there is no variance between observations present at the same location. Three common variogram functions are shown in Table 1 – the same method can be followed for other functions.

The effective sample size in this setup therefore depends only on the distance between sampled points and the parameters r and ρ. It should be clear that as spatial autocorrelation increases $r \to \infty$, then $N^* \to N/(1 - \rho(N - 1))$, and in the presence of no spatial autocorrelation, $r \to 0$ then $N^* \to N$.

2.1. Approximations to the ICC

One can see that there is a correspondence between Equations (2) and (4) in that the equivalent 'spatial' ICC is:

$$\text{ICC}_{\text{sp}} = \frac{\rho}{n(n-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} f(d_{ij}, r) \tag{5}$$

The term $s = \frac{1}{n(n-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} f(d_{ij}, r)$, and hence the locations of sampled units, is required for this calculation. However, prior to enumeration the locations of all units may

be unknown and the precise pattern of sampling also uncertain. For practical applications an approximation to s, $\tilde{s}$, using simpler values that are likely to be known *ex ante* is required (Griffith [17] examine some similar approximations to the average covariance). The approximate effective sample size, for J clusters of (mean) size n would then be:

$$\tilde{N}^* = \frac{Jn}{1 + \rho \tilde{s}(n-1)} \tag{6}$$

Under simple random sampling, if we assume that the possible sampling locations are uniformly distributed across the area of interest, then the average interpoint distance remains the same for all sample sizes for a fixed area. The term s therefore can be well approximated by some function of the radius of the area, R, assuming it is approximately circular. For cases where the area is better approximated by a quadrilateral then we use $\sqrt{A}$ where A is the area of the area of interest. Our approximations to s for simple random sampling are based on functions of the ratio $q_1 = r/R$.

As an alternative to simple random sampling, one can use a spatially regular sampling method. Under such a scheme, points are selected at random, but in a way that ensures they are well spread out across the area [9,24]. One particular design is the 'inhibited design with close pairs', which can be used either for sampling from across a continuous area or from a set of fixed locations. A sample of size n is randomly chosen so that no two locations are less than an inhibition distance δ apart. The sample is then supplemented with a set of 'close pairs': k sampled locations are randomly selected and then another location within distance η is randomly sampled [8]. This ensures that sampled locations are well dispersed across the area of interest. The close pairs ensure that information on shorter distances than δ is available (should it be required), although we assume no close pairs are included here. For a fixed area, a reduction in sample size will result in larger distances on average between sampled locations under a spatially-inhibited sampling design. The effect of this is to reduce the average covariance between sampled points. Approximations to s for spatially regular sampling methods are therefore functions of both the area and sample size. In particular, we consider the ratio:

$$q_2 = \frac{r}{\sqrt{R} + R/n}$$

To generate an approximation:

(1) We simulate 10,000 data sets of points uniformly distributed across a unit disc and vary r between 0.001 and 10.000 and n between 10 and 200.
(2) For each set of points we calculate s for each variogram function f from Table 1 and q_1.
(3) We then fit models of the form

$$s = g(q_1) + \epsilon$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$ using the `mle` function in the `stats4` package for R. For g we consider a range of sigmoid functions including the logistic function, the hyperbolic tangent function, a cubic function, and the algebraic function $g(x) = 1 - 1/(1 + \alpha q_1^\beta)$. The best performing model was selected as that with the lowest mean squared error.
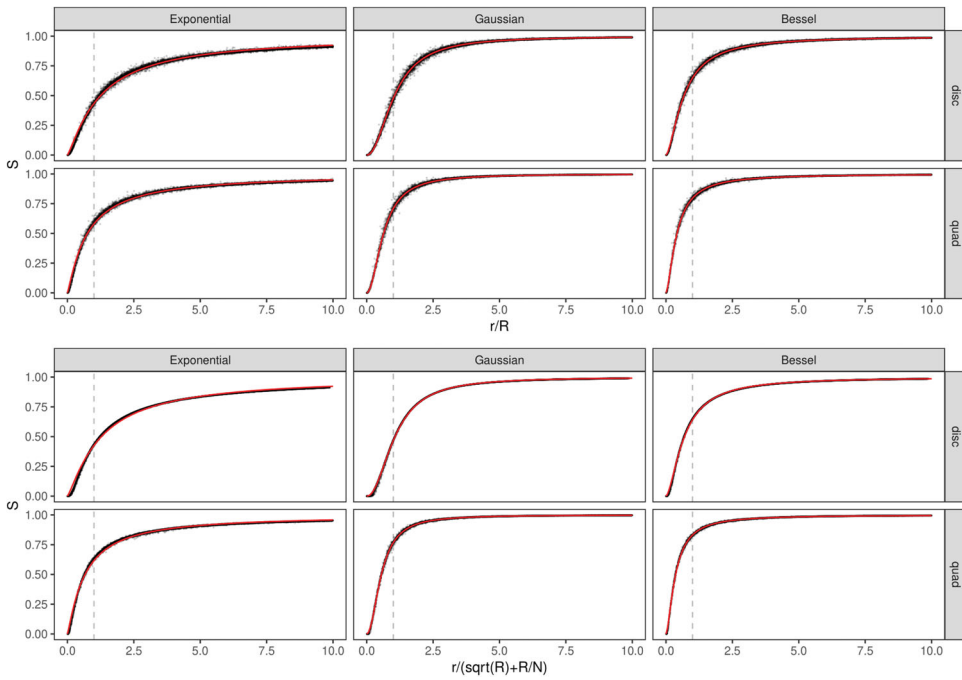
Figure 2. Simulated average covariance values *s* (points) and best fitting approximations (red lines) for different semivariance functions (columns within panels), shape of the area (rows within panels), and uniformly (top panel) or regularly (bottom panel) distributed points.

For the second data set, the simulation parameters remain the same except points were simulated as regularly spaced across the unit disc. Models were specified in terms of q_2. The third and fourth data sets were simulated in the same way as above, except that instead of a unit disc, points were either uniformly or regularly distributed across a quadrilateral with one side of length 1 and the other varying uniformly between 0.1 and 10.0.

Figure 2 shows the simulated data points and best fitting approximation. In all cases the algebraic function $g(x) = 1 - 1/(1 + \alpha q_1^\beta)$ provided the best fit in terms of mean squared error. Table 1 reports the model parameters for the disc (quadrilateral approximations are reported in Table 2, Appendix). Evident from these simulations is that the average covariance can be well approximated by the ratios specified above. Under simple random sampling, depending on the nature of the covariance function, there is little decline in the covariance with increasing R until $R > r$, i.e. the radius or square root of the area is less than the range. When locations are well spread out, such as by using a spatially-regulated sampling method, then the average covariance is smaller. For example, for $r = 0.5$ and $R = 1$ and 20 sampled locations, the value for $\tilde{s}$ with a Gaussian covariance function is 0.150, which reduces to 0.138 with 10 sampled locations. The reductions in average covariance are relatively small, but may translate into useful gains in effective sample size. If $\rho = 0.3$ (so the equivalent ICCs are 0.0500 and 0.0414), then one cluster of size 20 where $s = 0.150$ provides an approximate effective sample size of $\tilde{N}^* = 10.8$, whereas two clusters of size 10 and $s = 0.138$ gives an approximate effective sample size of $\tilde{N}^* = 14.6$, a gain of approximately 20 percentage points in effective sample size per cluster. Figure 3 shows the implied ICC based on Equation (5) – the ICC is highly sensitive to the value of the model parameters.
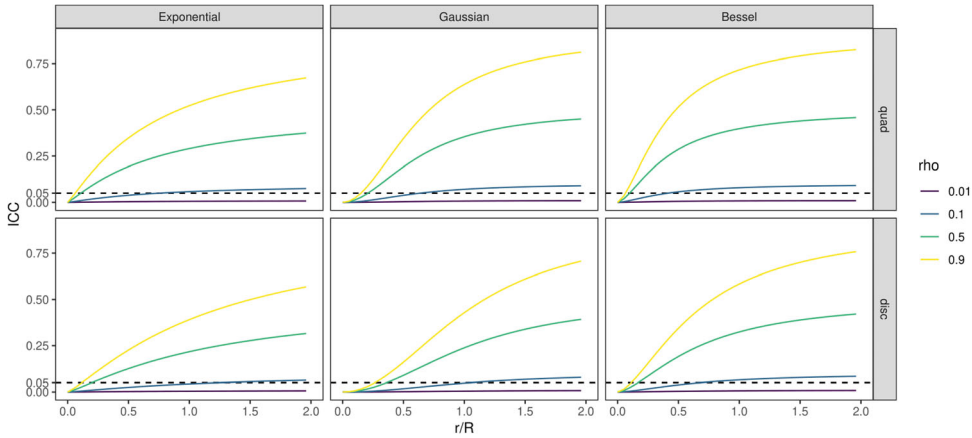
Figure 3. Approximate value of the intraclass correlation coefficient for simple random sampling under different choices of semivariance function and values for the ratio r/R and ρ.

3. Simulation study of two-stage sampling schemes

To illustrate the how choices over the size, number, and sampling scheme within clusters can affect the effective sample size we conduct a short simulation-based study. We describe a cluster-based sampling scheme using a number of parameters:

- J: the number of clusters to be sampled;
- m: the average number of individuals/units to be enumerated within each cluster to form the sampling frame;
- $M = Jm$: the total number of enumerated locations;
- p: the proportion of the sampling frame to be sampled;
- $n = mp$: the average number of surveyed or enrolled individuals per cluster;
- $N = Jn = Jmp$: the total number of surveyed individuals.

Our interest lies in determining the effective sample size N^*. We assume that clusters are far enough apart so that there is no correlation between individuals in different clusters. The generic description for two-stage sampling here encompasses a range of methods. For example, using relatively small clusters and setting $p = 1$ is sometimes referred to as 'compact segment sampling' as distinct from 'two-stage cluster sampling', since there is no sampling in the second stage [5]. Various methods of sampling individuals within clusters can also be used. Thus, this framework covers a range of methods, and we explicitly compare different alternatives in a simulation-based experiment.

We examine two scenarios. First, M and p are fixed (and hence n) and only J can be varied. The budget or time may only permit a certain number of locations and individuals to be enumerated and sampled, but the number of clusters can be flexibly varied. Since the number of locations that can be enumerated in a cluster depends on the number of clusters, as J increases we assume that this implies a smaller cluster area. We simulate clusters of a disc shape and assume a density of 100 enumerated points per unit disc of area π. We vary the radius of the disc to give a specified number of points.

The second scenario is a scheme where m is fixed but J can be varied so that as J increases, M increases (as m remains constant) and so the proportion of the enumerated locations in each cluster being sampled p decreases. That is each cluster is the same size and the total number of individuals to be sampled is fixed so adding more clusters requires enumerating more locations and sampling fewer of them.

For both scenarios, we simulate points both uniformly and regularly distributed across the area to simulate random and spatially-regular sampling. Our target sample size is $N = 200$. We conduct 10,000 simulations of each scenario and sampling scheme. For each iteration, we generate n points (either uniformly or regularly) according the parameters m and p and determine n^* using Equation (1) and each of the three semivariance functions, as well as the approximations described in the preceding section. We consider $\rho = 0.01, 0.1, 0.5, 0.9$ and $r = 0.1, 0.5$.

Figure 4 reports the results of the simulations. These results can be interpreted in light of our approximations in the previous section. The top panel shows the effect of varying the number of clusters. Where the range is large ($r = 0.5$), and hence the ratio r/R is large and remains above one, there is an approximately linear relationship between the number of
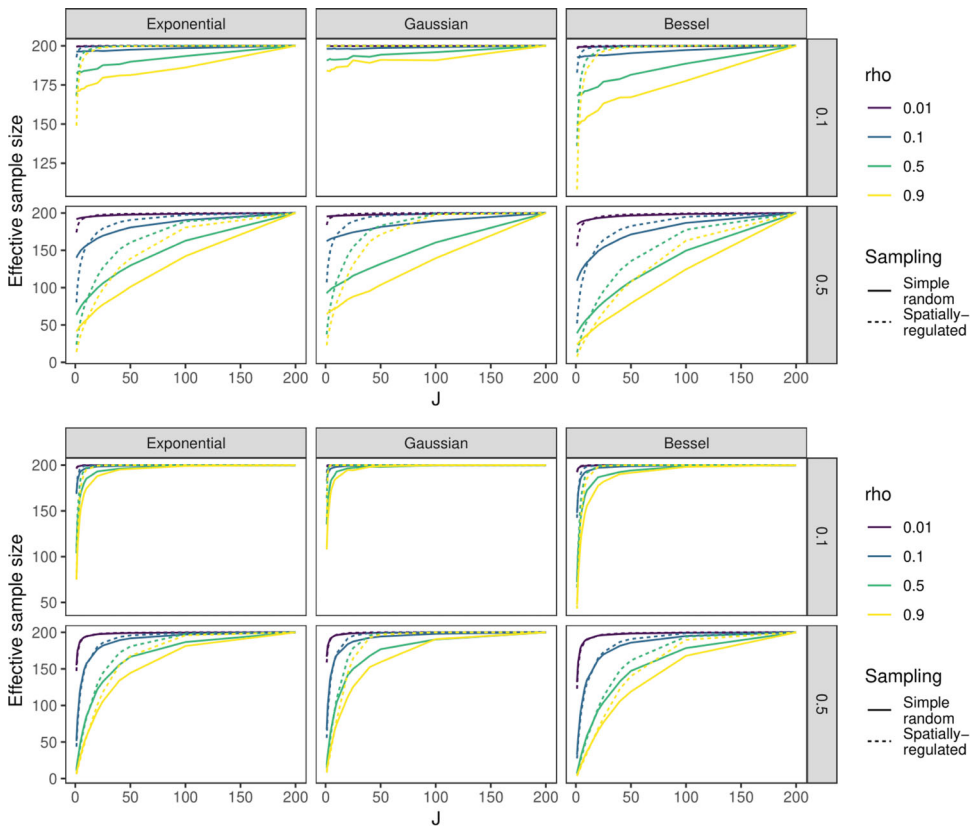

Figure 4. Results from the simulations showing effective sample size where $N = 200$ for values of ρ (color), sampling schemes (line types), values of r (rows within panels), and semivariance functions (columns within panels). Top panel: fixed M and p with varying J. Bottom panel: fixed m and varying p and J.

clusters and the overall effective sample size. However, where a spatially-regulated sampling scheme is used, increasing the number of clusters results in a rapid increase in the effective sample size as ratio q_2 declines below one. Where the range r is low relative to the size of the cluster, there is little covariance between individuals and high effective sample sizes.

The lower panel of the figure reports the results of the simulations where the number of locations to be enumerated (and hence area) per cluster is fixed. An increase in the number of clusters results in fewer locations being sampled. There are exponential increases in the effective sample size under both sampling schemes as the number of clusters increases when J is low indicating that the ratio r/R is likely below one. The effective sample size is generally larger under spatially-regulated sampling, although there is little benefit in scenarios where r or ρ is small.

4. Sampling parameters in practice and expert elicitation

The efficiency of a geographic cluster-based study design depends (approximately) on three parameters: ρ, r, and R (or $\sqrt{A}$). While the radius or area of the cluster area is determined by researchers, the values of the range and correlation parameters r and ρ are unlikely to be known and for most outcomes reliable estimates will not be available. However, both ρ and r (or transformations of them) can have relatively natural interpretations or be framed in terms of 'real-world' variation, so that reasonable values could be specified by those with domain-specific knowledge. Expert elicitation methods could therefore provide useful bounds. This is an advantage over using the ICC, which is also unlikely to be known but does not have a useful 'real-world' interpretation and there is little evidence of how it varies with cluster design. There are well-specified methods for capturing expert beliefs about the distribution of variables or parameters [32]. In particular, we suggest the following interpretations for the cluster-design parameters:

- $\sqrt{1-\rho}$: The ratio of the standard deviation between two very close neighbours to that of the whole population. One can follow procedures for the elicitation of a standard deviation. One might consider questions from which an estimate of standard deviation can be obtained, for example: 'For a given value a_1 of a random variable, what interval $[a_2, a_3]$ would contain with probability b the values of next door neighbours?' and 'What interval $[a_4, a_5]$ would contain with probability b the value of a randomly selected member of the population?'
- r: The range parameter is the minimum distance over which two individuals have negligible correlation. To elicit a value for this parameter, one can follow up the previous questions with: 'What is the minimum plausible distance over which the interval $[a_4, a_5]$ is plausible in the population?'

One can elicit either point estimates or probability densities. These estimates can be used in two ways. First, as a heuristic in designing the clusters. The clusters should aim to either have $R > r$ or if that is not possible to maximise the number of clusters. Second, the effective sample size can be determined. For example, if we define the approximate effective sample size to be (under an exponential variogram model and simple random sampling):

$$\tilde{N}^*(r, \rho, R, J, n) = \frac{Jn}{1 + \rho(n-1)\left(1 - (1 + 0.77(r/R)^{1.19})^{-1}\right)}$$

then we can plug in point estimates or average over elicited probability densities:

$$\tilde{N}^*(R, J, n) = \int \int \tilde{N}^*(r, \rho, R, J, n) h_r(r) h_\rho(\rho) \, dr \, d\rho$$

5. Dichotomous outcomes

The analysis can be extended to studies that have dichotomous variables as outcomes. The major difference is in the interpretation and specification of ρ. If the outcome now takes values $Y \in \{0, 1\}$ then we can rewrite the variogram in Equation (4) as per an 'indicator variogram':

$$\gamma(d) = \frac{1}{2} E\left[(Pr(Y(x+d) = 1; Y(x) = 0) + Pr(Y(x+d) = 0; Y(x) = 1))^2\right] \quad (7)$$

$$= Pr(Y(x+d) = 1; Y(x) = 0) \quad (8)$$

At the sill the two observations are independent so $\gamma(d) = Pr(Y(x) = 1)$ $(1 - Pr(Y(x) = 1))$, i.e. the variance of Y. At the nugget $\gamma(d)$ is equal to $Pr(Y(x+d) = 0|Y(x) = 1)Pr(Y(x) = 1)$. Plugging these values in gives

$$\rho = \frac{Pr(Y(x) = 1|Y(x+d) = 1) - Pr(Y(x) = 1)}{Pr(Y(x) = 1)(1 - Pr(Y(x) = 1))} \quad (9)$$

where the conditional probability is at the nugget (i.e. small d). Further discussion of indicator variograms can be found elsewhere [4,11].

To elicit a value for ρ we need to ask:

(1) $Pr(Y(x) = 1)$: What the expected probability of the outcome is across the whole population of interest;
(2) $Pr(Y(x) = 1|Y(x+d) = 1)$: If we observe a positive response, what is the probability that someone else living in the same household (or at the same location) would have a positive response.

It is clear that if there exists no spatial autocorrelation then $\rho = 0$, and if the autocorrelation is present then $Pr(Y(x) = 1|Y(x+d) = 1) > Pr(Y(x) = 1)$ and $0 < \rho \leq 1/Pr(Y(x) = 1)$.

6. Optimising cluster-based study design

Efficient cluster-based study design can also be seen as an optimisation problem of maximising effective sample size through the choice of m, p, and J subject to financial or time constraints:

$$\text{Max. } \tilde{N}^*(m, p, J) \quad \text{s.t. } c_m M + c_n N \leq C \quad (10)$$

$$J \leq J_{\max} \quad (11)$$

$$n \leq m \quad (12)$$

where c_m is the cost per enumerated location and c_n is the cost per survey and where and C is the total budget (where 'cost' could be in time or money terms). The function to maximise depends on the choice of correlation function and sampling scheme, but, for example for the exponential correlation function under simple random sampling it is:

$$\tilde{N}^*(m, p, J) = \frac{Jmp}{1 + \rho(mp - 1)\left(1 - (1 + 0.76(r/(R))^{1.37})^{-1}\right)}$$

This optimisation problem can be solved relatively easily using statistical software. R Code to do this is provided in the Supplementary Information.

7. Example

Here we provide an example of the use of the design analysis presented in this article. We adapted a two-stage cluster sampling method for use in a household-based community survey on healthcare use in Kono District, Sierra Leone and Maryland County, Liberia. Census data on population numbers were almost a decade old and were thought to be an unreliable basis for a sampling frame particularly given recent events including the 2015 Ebola crisis. Moreover, population numbers were not captured at a level granular enough to enable stratification at village level or spatial sampling. Ground-truthing revealed that digital maps on the platform OpenStreetMap were out-of-date and between 10 and 20% of structures were not accurately represented (either newly built or had been demolished). There did not exist funding to enumerate and map all the households in the districts, and so a two-stage scheme was adopted. An online application was generated that could be used to draw shapes over satellite images to create spatial data using the ArcGIS platform. Cluster borders could then be drawn around all structures and conurbations using the most recently available satellite images of the areas of interest. Areas were drawn with the aim of containing a pre-specified number of households and be bounded by features easily identifiable on the ground including roads, rivers, and tree lines. The number of households that could be sampled and interviewed was fixed, so the question remained as to how to determine the appropriate size of cluster to ensure reasonable efficiency while not requiring the enumeration of too large a number of households or travel to too many clusters.

While there were multiple outcomes of interest, for the design analysis, we focus here on estimating the proportion of the population who have seen a doctor, nurse, or clinical officer in an outpatient setting in the preceding 12 months. This variable is used as a filter for a number of following questions about the visits. Evidence from similar settings has suggested the outpatient consultation rate is around 0.5 visits per person-year, although the number of visits is not uniformly distributed in the population. $Pr(Y(x) = 1)$ was therefore set at 0.4. For individuals living in the same household they might be more likely to attend if someone else visits a healthcare provider either because of possible shared sources of ill-health or for convenience (e.g. a parent taking a child). $Pr(Y(x) = 1|Y(x + d) = 1)$ was therefore set at 0.6. The range r, which describes the distance over which observations were thought to correlated (see Section 2), was set at 10m as these effects were not thought to extend beyond local neighbours relative to the population of the district outside of major pandemics such as Ebola or Covid-19. We assumed a constant density of households (approximately one per 3 square meters). Budget implications meant the total number of households that could be enumerated and surveyed was 500. The solid line
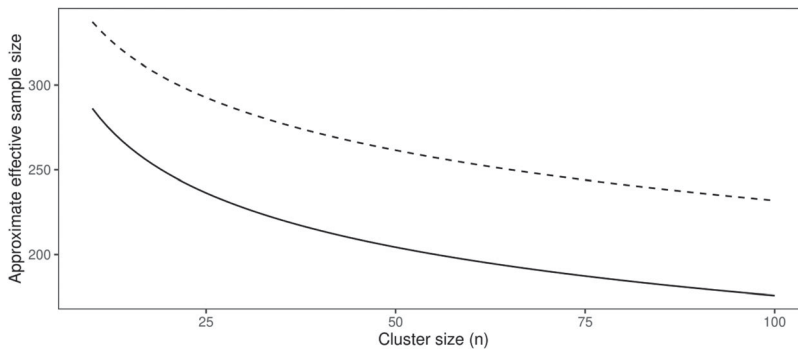
Figure 5. Approximate effective sample size for two strategies: survey all enumerated individuals (solid line) or enumerate double the number of individuals and sample half with simple random sampling (dashed).

in Figure 5 shows the approximate effective sample size assuming an exponential correlation function where all enumerated households are surveyed ($p = 1$). The most efficient strategy was to have the largest number of clusters possible which was around 30, giving 15 households per cluster and an approximate sample size of 260. As an alternative we examined doubling the cluster size and enumerating more households but taking a smaller sample to achieve the same approximate effective sample size (Figure 5 dashed line) – this would reduce the number of households needing to be surveyed per cluster by around 20% or 3 households in a cluster of size 15 but double the number being enumerated. However, this was less cost-effective. The optimum strategy identified by solving the optimisation problem in the previous section, assuming enumeration would take 10 minutes, surveying 50 minutes, and with a total 'budget' of 20,000 minutes, provided $J = 35$, $m = 15$, and $p = 0.72$, which gives an effective sample size of 274.

8. Discussion

The design of cluster-based studies is not an exact science. The evidence required to make precise *ex ante* calculations of the precision afforded by a sample does not generally exist. Yet it is still important to ensure research designs are as efficient as possible. For studies based on geographically-defined clusters, the standard method is to use the intraclass correlation coefficient (ICC) to estimate power, expected confidence interval width, or other measure of the information in a sample [14,19,35]. However, the values used for the ICC are often *ad hoc*, based on commonly used values or occasionally evidence from studies of a similar design. While these values are representative of those estimated in empirical studies, it is not known how they should be varied with respect to clusters size, density, or method of sampling. Moreover, the ICC does not have a simply interpretable 'real world' meaning, making expert elicitation-type exercises difficult. In this article, we have explicated the efficiency of different geographical cluster design when the autocorrelation can be described in geostatistical terms. In particular, we show how effective sample size can be reasonably approximated with information on the range and parameter ρ. Small changes in cluster design can translate into relatively large changes in the effective sample size. Generally speaking the most efficient strategy is to maximise the number of clusters if the clusters

are far enough apart to be independent. At the extreme there would be well-spaced out individuals in clusters of one. However, there is likely to be an upper limit on clusters for practical reasons including travelling and the number of locations that are far enough apart in the area of interest in which case methods to determine optimal designs are required.

One can extend the analysis here to examine correlated clusters according to the distances between sampled units in each cluster. But *inter*-cluster correlation would introduce further complications for interventional studies due to 'contamination': intervening in one cluster has effects in another. In the presence of spatial autocorrelation, this could be due to the underlying correlation or direct effects of the intervention over space. With appropriate geostatistical analysis, these effects could be distinguished in experimental studies. The use of geostatistical methods in an experimental setting may provide useful avenues for research in this regard, however, analyses that take into space into account are exceedingly rare in cluster randomised trials. A systematic review of studies in 2017 only identified ten cluster trials that incorporated any kind of spatial analysis, and this was generally limited to allowing the intervention effect to vary linearly by distance from the intervention location [23]. Model mis-specification could result in inefficient or biased estimators of treatment effects from these trials. Indeed, the interventional context provides an additional factor to consider for design in that larger clusters are more likely to have more 'dilute' treatment effects if the intervention's effects decay by distance from a point source. This phenomenon also raises questions from cluster trials with geographically defined clusters as average treatment effects are clearly dependent on cluster design. There has been little development of the use of spatial statistical methods for experimental and interventional study designs, and further research is required.

For household surveys, there has been little direct comparison of the most efficient sampling methods in settings where no sampling frame exists, particularly in terms of maximising the effective sample size for a given budget. Our generic framework here encompasses a range of cluster-based area sampling methods, however other alternatives have also been proposed such as 'EPI sampling', which randomly generates point locations in an area from which to sample an individual. Chao *et al.* [5] compared two-stage methods with simple random sampling from a complete sample frame and EPI sampling using a dataset of small businesses in South Africa. They suggested that a two-stage method was likely to be most cost-effective, however only considered a 'compact segment sampling' approach with small clusters of a particular size in which all individuals were enumerated and surveyed. As this article demonstrates though, it is likely further improvements could be made by varying some of the design parameters. Milligan *et al.* [28] compared EPI sampling and 'compact segment sampling' in a field study and concluded that the latter method was preferred for its scientific validity but it was not strictly more efficient or cost-effective. The methods in this article permit a more precise approach to the decision over the design of such clusters and sampling strategies. We did not include EPI sampling as a comparator as it is not a two-stage method and has been shown to perform poorly in previous work.

The analysis in this article reflects a more standard geostatistical sampling problem discussed widely elsewhere, that is how to efficiently sample locations across an area of interest in the presence of spatial autocorrelation [17,27]. The target for inference could be a population mean, the predicted prevalence of a disease and its spatial distribution, or the parameters of a particular statistical model. A number of sampling schemes exist, which aim to evenly spread out a sample across an area. For example, a lattice-based design

might overlay a hexagonal lattice over the area of interest and sample individuals at the vertices, which may be supplemented with simple random sampling in some of the cells to capture information on shorter-range correlation [38]. We opted to consider a spatially-inhibited sampling design, which is more efficient than simple random sampling in the presence of spatial autocorrelation, but also is simple to implement for a set of fixed locations. Nevertheless, alternative sampling schemes may provide differing levels of efficiency for cluster-based designs. We also note that under spatially-regular sampling not all locations have an equal probability of being included in the sample as less densely packed areas will be more likely to be sampled, so appropriate probability weights should be calculated.

There are of course a number of weaknesses to the approach proposed here. We rely on an assumption that sampling locations are uniformly distributed across the area of interest. This may not be a strong assumption for small, local neighbourhoods in an urban area in general, however, clusters that cross neighbourhood boundaries may have highly variable population density. The variogram functions used here may be relatively simple and not accurately describe the nature of spatial autocorrelation, however they are widely used for this purpose as they provide a good approximation to real-world variation. Indeed, the practical use of these methods is likely to be more complex than the scenarios presented here, for example, most clusters will not be regular discs or quadrilaterals. Nevertheless, these methods provide a more scientifically grounded approach to cluster design than currently used methods and as such greater validity. This article has not considered areal analysis, that is where the unit of observation is the aggregated cluster-level. In an areal analysis, the choice of the size of the cluster would further introduce scale effects, which has been discussed elsewhere [6]. Areal analyses, though, are not common for the types of study we have discussed in this article.

The design of cluster-based studies commonly relies on ICCs, and sample size calculations are relatively straightforward under this approach. However, using ICCs in this way for geographically-defined clusters may lack scientific validity without reliable evidence on cluster-level variance. There also exists little evidence regarding how ICCs should be varied with respect to different design parameters like sampling method and cluster size. We have provided tools for more reliable cluster-design for surveys and experimental studies. One recommendation that arises from this study in particular is to use spatially regular sampling methods with geographically defined clusters as opposed to simple random sampling to improve efficiency.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] S. Basu and T. G. Thibodeau, *Analysis of spatial autocorrelation in house prices*, J. Real Estate Financ. Econom. 17 (1998), pp. 61–85. DOI:10.1023/A:1007703229507.

[2] R. Benedetti, F. Piersimoni, and P. Postiglione, *Spatially balanced sampling: A review and A reappraisal*, Int. Stat. Rev. 85 (Dec 2017), pp. 439–454. DOI:10.1111/insr.12216.

[3] M. K. Campbell, P. M. Fayers, and J. M. Grimshaw, *Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research*, Clinical Trials 2 (2005), pp. 99–107. DOI:10.1191/1740774505cn071oa.

[4] J. R. Carr, R. E. Bailey, and E. D. Deng. *Use of indicator variograms for an enhanced spatial analysis*. J. Int. Assoc. Math. Geol, 17 (8) (Nov 1985), pp. 797–881. DOI:10.1007/BF01034062

[5] L.-W. Chao, H. Szrek, K. Peltzer, S. Ramlagan, P. Fleming, R. Leite, J. Magerman, G. B. Ngwenya, N. Sousa Pereira, and J. Behrman, *A comparison of EPI sampling, probability sampling, and compact segment sampling methods for micro and small enterprises*, J. Dev. Econ. 98 (May 2012), pp. 94–107. DOI:10.1016/j.jdeveco.2011.08.007.

[6] Lei Chen, Yong Gao, Di Zhu, Yihong Yuan, and Yu Liu. *Quantifying the scale effect in geospatial big data using semi-variograms*. PLOS ONE, 14 (11) (Nov 2019) pp. e0225139. DOI:10.1371/journal.pone.0225139.

[7] J.P. Chilès and P. Delfiner, *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed., John Wiley & Sons, Inc., 2012. DOI:10.1002/9781118136188.

[8] M. Chipeta, D. Terlouw, K. Phiri, and P. Diggle, *Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure*, Environmetrics 28 (Feb 2017), pp. e2425. DOI:10.1002/env.2425.

[9] E. Delmelle. *Spatial Sampling*. The SAGE Handbook of Spatial Analysis, 2009, pp. 183–206.

[10] P. J. Diggle and E. Giorgi, *Model-Based geostatistics for prevalence mapping in low-Resource settings*, J. Am. Stat. Assoc. 111 (Jul 2016), pp. 1096–1120. DOI:10.1080/01621459.2015.1123158.

[11] O. Dubrule. *Indicator variogram models: do we have much choice?* Math. Geosci., 49 (4) (May 2017) pp. 441–465. DOI: 10.1007/s11004-017-9678-x

[12] S. Eldridge and S. Kerry, *A Practical Guide to Cluster Randomised Trials in Health Services Research*, Wiley, Chichester, UK, 2012.

[13] J. P. Elhorst and K. Zigova, *Competition in research activity among economic departments: evidence by negative spatial autocorrelation*, Geogr. Anal. 46 (Apr 2014), pp. 104–125. DOI:10.1111/gean.12031.

[14] F. Fowler*Survey Research Methods*, 4th ed., 2012. DOI: 10.4135/9781452230184.

[15] A. Grafström, N. L. P. Lundström, and L. Schelin, *Spatially balanced sampling through the Pivotal method*, Biometrics 68 (Jun 2012), pp. 514–520. DOI:10.1111/j.1541-0420.2011.01699.x.

[16] A. Grafström and Y. Tillé, *Doubly balanced spatial sampling with spreading and restitution of auxiliary totals*, Environmetrics 17 (2013), pp. 61–85. DOI:10.1002/env.2194.

[17] D. A. Griffith, *Effective geographic sample size in the presence of spatial autocorrelation*, Ann. Assoc. Amer. Geogr. 95 (Dec 2005), pp. 740–760. DOI:10.1111/j.1467-8306.2005.00484.x.

[18] M. C. Gulliford, O. C. Ukoumunne, and S. Chinn, *Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: data from the health survey for England 1994*, Am. J. Epidemiol. 149 (May 1999), pp. 876–883. DOI:10.1093/oxfordjournals.aje.a009904.

[19] K. Hemming, S. Eldridge, G. Forbes, C. Weijer, and M. Taljaard, *How to design efficient cluster randomised trials*, BMJ 358 (Jul 2017), pp. j3064. DOI:10.1136/bmj.j3064.

[20] K. Hemming, J. Kasza, R. Hooper, A. Forbes, and M. Taljaard, *A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT calculator*, Int. J. Epidemiol. 49 (Feb 2020), pp. 979–995. DOI:10.1093/ije/dyz237.

[21] R. Hooper, S. Teerenstra, E. de Hoop, and S. Eldridge, *Sample size calculation for stepped wedge and other longitudinal cluster randomised trials*, Stat. Med. 35 (Nov 2016), pp. 4718–4728. DOI:10.1002/sim.7028.

[22] N. Z. Janjua, M. I. Khan, and J. D. Clemens, *Estimates of intraclass correlation coefficient and design effect for surveys and cluster randomized trials on injection use in Pakistan and developing countries*, Tropical Medicine and International Health 11 (2006), pp. 1832–1840. DOI:10.1111/j.1365-3156.2006.01736.x

[23] C. Jarvis, G. L. Di Tanna, D. Lewis, N. Alexander, and W. J. Edmunds, *Spatial analysis of cluster randomised trials: a systematic review of analysis methods*, Emerg. Themes. Epidemiol. 14 (Dec 2017), pp. 12. DOI:10.1186/s12982-017-0066-2.

[24] Jin-Feng Wang, A. Stein, *Bin-Bo Gao, and Yong Ge*. A review of spatial sampling. Spatial Statistics, 2 (Dec 2012) , pp. 1–14. DOI:10.1016/j.spasta.2012.08.001.

[25] S. Kul, K. Vanhaecht, and M. Panella, *Intraclass correlation coefficients for cluster randomized trials in care pathways and usual care: hospital treatment for heart failure*, BMC. Health. Serv. Res. 14 (2014), pp. 376. DOI:10.1186/1472-6963-14-84

[26] V. Lorant, I. Thomas, D. Deliege, and R. Tonglet, *Deprivation and mortality: the implications of spatial autocorrelation for health resources allocation*, Soc. Sci. Med. 53 (Dec 2001), pp. 1711–1719. DOI:10.1016/S0277-9536(00)00456-1.

[27] M. Manton, *Collecting Spatial Data: Optimal Design of Experiments for Random Fields*, 3rd ed. Eos, Transactions American Geophysical Union, 2008. DOI:10.1029/2008eo160009

[28] P. Milligan, A. Njie, and S. Bennett, *Comparison of two cluster sampling methods for health surveys in developing countries*, Int. J. Epidemiol. 33 (2004), pp. 469–476. DOI:10.1093/ije/dyh096.

[29] I. Molho, *Spatial autocorrelation in British unemployment*, J. Reg. Sci. 35 (Nov 1995), pp. 641–658. DOI:10.1111/j.1467-9787.1995.tb01297.x.

[30] R. L. Munasinghe and R. D. Morris, *Localization of disease clusters using regional measures of spatial autocorrelation*, Stat. Med. 15 (Apr 1996), pp. 893–905. DOI:10.1002/(SICI)1097-0258(19960415)15:7/9¡893::AID-SIM258¿3.0.CO;2-M.

[31] D. M. Murray, *Design and Analysis of Group Randomised Trials*, Oxford University Press Inc., New York NY, 1998.

[32] A. O'Hagan, *Eliciting expert beliefs in substantial practical applications [Read before the royal statistical society at ameeting on 'Elicitation' on Wednesday, April 16th, 1997, the president, professor A. F. M. Smithin the chair]*, J. R. Stat. Soc.: Ser. D (The Stat.) 47 (Mar 1998), pp. 21–35. DOI:10.1111/1467-9884.00114.

[33] S. Riley, *Large-Scale spatial-Transmission models of infectious disease*, Science 316 (Jun 2007), pp. 1298–1301. DOI:10.1126/science.1134695.

[34] R. Shahid, S. Bertazzon, M. L. Knudtson, and W. A. Ghali, *Comparison of distance measures in spatial analytical modeling for health service planning*, BMC. Health. Serv. Res. 9 (Dec 2009), pp. 200. DOI:10.1186/1472-6963-9-200..

[35] C. J. Skinner, *Design effects of two-Stage sampling*, J. R. Stat. Soc., Ser. B (Methodological) 48 (1986), pp. 89–99.

[36] S. E. Spielman and E.-H. Yoo, *The spatial dimensions of neighborhood effects*, Soc. Sci. Med. 68 (Mar 2009), pp. 1098–1105. DOI:10.1016/j.socscimed.2008.12.048.

[37] P.-J. Tsai, M.-L. Lin, C.-M. Chu, and C.-H. Perng, *Spatial autocorrelation analysis of health care hotspots in Taiwan in 2006*, BMC. Public. Health. 9 (Dec 2009), pp. 464. DOI:10.1186/1471-2458-9-464

[38] E. A. Yfantis, G. T. Flatman, and J. V. Behar, *Efficiency of kriging estimation for square, triangular, and hexagonal grids*, Math. Geol. 19 (1987), pp. 183–205. DOI:10.1007/BF00897746.

Appendix 1. Approximations

Table A1. Covariance functions and approximations to within-cluster mean covariance using the function $g(x) = 1 - 1/(1 + \alpha q_1^\beta)$ for sampling on a quadrilateral.

Model	$f(d,r)$	Approx. mean covariance, $\bar{s}$	
		Simple random	Spatially-inhibited
Gaussian	$\exp(-d_{ij}^2/r^2)$	$\alpha = 2.432, \beta = 1.967$	$\alpha = 3.301, \beta = 2.023$
Exponential	$\exp(-d_{ij}/r)$	$\alpha = 1.385, \beta = 1.136$	$\alpha = 1.614, \beta = 1.154$
K-Bessel	$\frac{d_{ij}}{r} K_1 \left(\frac{d_{ij}}{r} \right)$	$\alpha = 3.904, \beta = 1.540$	$\alpha = 4.937, \beta = 1.570$

```
# Find the optimum parameters for a sampling scheme. Inputs are:
# r = range
```

```
# R0 = average area occupied by one location
# rho = parameter rho
# cm = cost per enumerated location
# cn = cost per survey
# C = total budget
# Jlims = c(lower, upper) = the lower and upper limits to the
                            number of clusters
# func = one of exp, gaus, or bess corresponding to exponential,
          Gaussian, or k-Bessel
# samp = one of simple or spatial for simple random sampling or
          spatially-regulated sampling
optimSamp <- function(r,R0,rho,cm,cn,C,Jlims,func,samp){
  if(!func%in%c('exp','gaus','bess')){
    stop("Function has to be one of 'exp', 'gaus', or 'bess' ")
  }
  if(!samp%in%c('simple','spatial')){
    stop("Sampling has to be either 'simple' or 'spatial' ")
  }

  a <- c(r,R0,rho)
  b <- c(cm,cn,C)

  if(func=="exp"){
    if(samp=="simple"){
      eval_f0 <- function(x,a,b){
        s <- -1*x[1]*x[2]*x[3]/(1+(x[2]*x[3]-1)*a[3]*
        (1 - 1/(1 + 0.77*(a[1]/(a[2]*sqrt(x[2])))^1.19)))
        return(s)
      }
    }
    if(samp=="spatial"){
      eval_f0 <- function(x,a,b){
        s <- -1*x[1]*x[2]*x[3]/(1+(x[2]*x[3]-1)*a[3]*
          (1 - 1/(1 + 0.88*(a[1]/
              (sqrt(a[2]*sqrt(x[2])) +
                a[2]*sqrt(x[2])/x[2]*x[3] ))^2.11)))
        return(s)
      }
    }
  }
  if(func=="gaus"){
    if(samp=="simple"){
      eval_f0 <- function(x,a,b){
        s <- -1*x[1]*x[2]*x[3]/(1+(x[2]*x[3]-1)*a[3]*
          (1 - 1/(1 + 0.91*(a[1]/(a[2]*sqrt(x[2])))^2.07)))
        return(s)
      }
    }
    if(samp=="spatial"){
      eval_f0 <- function(x,a,b){
        s <- -1*x[1]*x[2]*x[3]/(1+(x[2]*x[3]-1)*a[3]*
          (1 - 1/(1 + 0.88*(a[1]/(a[2]*sqrt(x[2])))^2.11)))
        return(s)
      }
```

```
      }
    }
    if (func=="bess"){
      if (samp=="simple"){
        eval_f0 <- function(x,a,b){
          s <- -1*x[1]*x[2]*x[3]/(1+(x[2]*x[3]-1)*a[3]*
            (1 - 1/(1 + 1.85*(a[1]/(a[2]* sqrt(x[2])))^1.58)))
          return(s)
        }
      }
      if (samp=="spatial"){
        eval_f0 <- function(x,a,b){
          s <- -1*x[1]*x[2]*x[3]/(1+(x[2]*x[3]-1)*a[3]*
            (1 - 1/(1 + 1.80*(a[1]/
              ( sqrt(a[2]* sqrt(x[2])) +
                [2]* sqrt(x[2])/x[2]*x[3] ))^1.60)))
          return(s)
        }
      }
    }

    eval_g0 <- function(x,a,b){
      return(x[1]*x[2]*b[1] + x[1]*x[2]*x[3]*b[2] - b[3])
    }

    res1 <- nloptr(x0 = c(floor(Jlims[2]/2),20,1),
            eval_f = eval_f0,
            lb = c(Jlims[1],1,0),
            ub = c(Jlims[2],Inf,1),
            eval_g_ineq = eval_g0,
            opts = list("algorithm"="NLOPT_LN_COBYLA",
                        xtol_rel=1e-4),
            a=a,
            b=b)

    return(c(
      J = round(res1$solution[1],0),
      m = round(res1$solution[2],0),
      p = round(res1$solution[3],2),
      N = round(round(res1$solution[1],0)*round(res1$solution[2],0)*
                round(res1$solution[3],2),0),
      ess = -1*round(res1$objective,0)
    ))
}

# example call to function:
> optimSamp(10,15,0.5,30,50,25000,c(1,20),"exp","simple")
  J   m   p   N ess
 20  16   1 320 215
```