A spatial sampling scheme for a road network

by

Hayley Reynolds

Submitted in partial fulfillment of the requirements for the degree

Magister Scientiae

In the Department of Statistics

In the Faculty of Natural and Agricultural Sciences

University of Pretoria

October 2017

I, *Hayley Reynolds*, declare that this mini-dissertation (100 credits), which I hereby submit for the degree Magister Scientiae in Mathematical Statistics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:

Date:

Summary

Rabies transmission is a concern, specifically in developing countries. The most common cause of transmission of rabies is through a bite from an infected host. For this reason, it is necessary to vaccinate animal populations of villages so as to reduce the spread of this disease. The most common approach sees state veterinarians set up vaccination stations for a few days in the centre of villages with the idea that the villagers will take the initiative to bring their animals for vaccination. This can be unreliable.

This mini-dissertation compares traditional and spatial random sampling in an attempt to develop an efficient and cost-effective sampling strategy to achieve herd immunity in villages. The houses of the village of Buchanchari in Tanzania are sampled according to traditional as well as spatial strategies which, combined with graph theory, aim to optimise the cost of the sampling procedures. Each sampling technique is bootstrapped 1000 times in order to obtain the distribution of the cost functions in each case. This is necessary since the sampling techniques used in this mini-dissertation are design-based and, therefore, random for each iteration. The theory of the sampling methods used is also discussed in detail.

The results of the bootstrapped samples lead to the conclusion that the spatial nature of data is an important component to consider when sampling geographical locations. These sampling procedures can be applied to various types of spatial data and the implementation and interpretation of the procedures are uncomplicated.

Acknowledgments

I would like to formally express my gratitude to STATOMET, the Centre for Artificial Intelligence Research (CAIR) and the National Research Foundation of South Africa (NRF CSUR grant number 90315) for financial support in the form of a postgraduate bursary and sponsorship. This afforded me the unforgettable opportunity of attending the 2016 and 2017 SASA conferences. I would also like to thank the DST/NRF SARChI Chair in Mathematical Models and Methods in Bioengineering and Biosciences for sponsoring my attendance at the 2017 Biomath conference.

To my supervisors, Dr Inger Fabris-Rotelli, Dr Theodor Loots and Prof Alfred Stein, your patience, motivation, and immense knowledge was invaluable and I am forever grateful for this experience.

To my friends in the Statistics Department at the University of Pretoria, you made every day enjoyable and filled with laughter. Thank you to my incredible family, without your love and support none of this would have been possible and finally to Jason le Roux...you know.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Waldo Tobler, a professor of Geography and Statistics at the University of California, suggested the First Law of Geography: "Everything is related to everything else, but near things are more related to each other" [69]. This is the driving force behind spatial statistics.

Spatial statistics is a stochastic process with a specific methodology which was developed through application in areas such as mining and engineering [33]. This is due to the desire to model phenomena whose location is of interest or directly contributes to a stochastic model [33]. Georges Matheron established the theory and methodology behind this work and is known as the founder of geostatistics [33]. Geostatistics uses adapted methods of regression to describe the spatial continuity of natural phenomena [39], therefore the methods developed by Matheron are also known more generally as regionalised variable theory. Data used in geostatistics is measured in a space where the region R is restricted and is most likely to exhibit dependence. To be specific, spatial dependence implies that data found nearer to one another will have more attributes in common than data found further apart. The data found will then be modelled and used to predict possible observations at locations for which no data have been recorded.

Other characteristics of the data which are also important are stationarity, ergodicity and isotropy. Stationarity is a set of assumptions regarding data distribution which allows for parameter estimation based on a standard set of properties [65]. Strict stationarity is the strongest form of stationarity and is defined as a process whose probability distribution does not change with a shift in time. Ergodicity and stationarity are closely related [75]. If $\bar{X}$ is the mean of a sample and $\mathbf{X}$ a random variable, then $\bar{X} = E[\mathbf{X}]$. This property allows spatial averages to be used for the entire space of data [21]. In an isotropic field, the variation of $\mathbf{X}$ is the same in every direction. In other words, the covariance function $C(\mathbf{h})$ depends only on the length of the distance vector $\mathbf{h}$ [2]. If the variation of $\mathbf{X}$ does not exhibit the same behaviour in every direction, then the field is anisotropic [41].

Spatial statistics is an ever-changing field due to its vast application. An increased availability of technology used to capture spatial data, such as satellites, allows for analysis of very large data sets [33]. When researchers attempt to obtain information regarding events which exhibit spatial variation, it is important to find the optimal sample locations within the region of interest R. Obtaining these samples is known as spatial sampling and is applicable in numerous fields, such as mining, environmental monitoring, geography and telecommunications [24]. Spatial sampling aims to collect samples from higher dimensions, that is 1-, 2- or 3-dimensions [74], these samples can then be used to estimate statistics about a parameter in an area and extend these estimates to unsampled locations [74].

There are two major approaches to sampling, they are design- and model-based sampling. The distinguishing factor between design- and model-based sampling is that design-based sampling considers the population values to be unknown but fixed, whereas with model-based sampling, the values are unfixed and any set of values observed is a single realisation of a stochastic model [74]. While design- and model-based sampling are both discussed in more detail in Chapter 4, design-based sampling is the main focus of this mini-dissertation. This is due to the fact that design-based sampling will allow for a first, uncomplicated comparison between traditional and spatial sampling techniques, future work will explore design-based comparison.

1.1 Motivation for study

The theory of sampling strategies will be applied to a census dataset regarding villages in Tanzania, courtesy of the Wellcome Trust[1]. This extensive data set is valuable in that it assists with the illustration of the effectiveness of the sampling techniques be presented hereafter since it provides census data of each household location and the number of animals in each household. Currently, state veterinarians vaccinate animals against rabies in the Tanzanian villages by setting up a vaccination point in the middle of the village and waiting for the villagers to bring the animals for vaccination. This is, of course, unreliable and ineffective. Therefore, a sampling strategy is proposed that not only covers a minimum of 70% of the animals in the village for sufficient herd immunity but also accounts for the spatial nature of the data and provides a more optimal approach for the veterinarian in terms of time. The village of Buchanchari will be used in all the applications for illustrative purposes.

Rabies has been reported in Tanzania, mainly in the southern highland regions, since 1954 [63]. To date, rabies is endemic in all districts in Tanzania and efforts are being made to contain the disease. It was determined that mass vaccination of at least 70% of an animal population is most effective, in terms of profitability and cost, in reducing transmission of rabies [78].

[1]Wellcome Trust, University of Glasgow, Dr Katie Hampson

It is often challenging to obtain representative samples for such a vaccination process through household surveys in low-income countries such as Tanzania. The most popular sampling method adopted by WHO is the EPI (Expanded Programme of Immunisation) method. This method entails selecting clusters, such as communities or villages, with probabilities of selection, proportional to the size and then sampling an equal number of houses from each of the selected clusters. For each selected cluster, the EPI method selects:

1. a central point in the community
2. a random direction (through spinning a bottle or pen)
3. a random house along the random direction chosen in 2.

After this random house is surveyed, the nearest house is surveyed (provided it meets the criteria) and so on until the desired sample size, n is obtained. The main concerns regarding this approach are the improper use due to lack of understanding of its limitations. This method offers no optimisation of the sample size or the sampling strategy. [9]

Bostoen et al. [10] tested the T-square sampling strategy as an option when sampling frames are not available. The T-square sampling strategy is a two-stage process which first optimises the sample size and then selects an optimal pathway connecting all the sampling points. This process is optimal for a distribution of houses which can be described by a spatially homogenous Poisson process. The houses in the village to which the sampling discussed in this mini-dissertation is applied was tested to determine if it can be described as a spatially homogenous process. The null hypothesis of the village exhibiting a homogenous Poisson process is tested using `quadrat.test` in `R` [59]. The hypothesis is rejected at a 1% level of significance, which suggests that using the T-square sampling procedure may be suboptimal for the village in question. [10]

The current approach for vaccination in Tanzanian villages takes some features from the EPI method but is rather basic and unreliable. The vaccination station is set up in the middle of the village with the expectation that the villagers will bring their animals for the necessary treatment. This is a very ineffective approach. The expectation that the villagers will bring all their animals for vaccination is rather bold. Some villagers may not even be aware of the number of animals that are technically in their possession, let alone that a vaccination station is being set up. This is an ineffective approach to sampling and will hinder the efforts of the veterinarian to vaccinate the animals in the village and obtain herd immunity.

For this reason, an optimal road network, over which the vaccinator (driver) will travel to the animals, is proposed. This road network will be the optimal approach to driving through the village while ensuring 70% coverage of the animals within a village. The concept is for the driver to drive along the optimal

route from location to location, which will be referred to as stopping points. At each stopping point, the driver will exit the vehicle and walk among the houses in the most optimal way so as to minimise walking distance, while vaccinating the animals at the sampled houses. The driver will then repeat this process by driving to the next stopping point, walking among the houses, vaccinating the animals and so on.

1.2 Overview

The structure of this mini-dissertation is as follows; Chapter 2 discusses a proposed optimal driving route along a road network between the houses of villages. This chapter also discusses optimising the walking route between sampled houses in the village such that the total walking distance of a vaccination process is minimised. Chapter 3 deals with traditional sampling techniques, discussing the theory of simple random, stratified and cluster sampling. Within each of these sections is a subsection of application where the theory of the sampling strategies is immediately applied to the data. The sampling strategies are bootstrapped to obtain cost distribution statistics, where the cost of a sampling procedure is the total distance walked between houses. Chapter 4 introduces the reader to spatial sampling techniques that align quite closely to traditional sampling. These techniques are uniform random, stratified and cluster sampling. After the theory of each of these sampling strategies is explained, they are applied to the dataset once again, calculating the distribution of the cost of the sampling strategies each time. Chapter 5 discusses the strengths and shortcomings of the proposed sample design, as well as the potential for future work. Chapter 6 contains concluding remarks regarding this mini-dissertation, and the `R` code which was employed for the application of the techniques is presented in the Appendix.

1.3 Objectives

The theory of traditional and spatial sampling is to be presented and discussed. Graph theory and its application in obtaining an optimal road network as well as optimal walking paths between houses will be introduced. A combination of sampling and graph theory will be applied to a village in Tanzania. The cost of the sampling strategies will be calculated and compared to determine if consideration of the spatial nature of data results in more efficient sampling.

Chapter 2

Optimal Road Network

This chapter will outline the theory and methodology used in obtaining the optimal driving route between the houses of the Buchanchari village. High density collections of houses are mapped using kernel density estimation and the optimal driving route is developed through graph theory. Explanation is also provided in terms of how graph theory will be utilised to attain the optimal walking route between sampled houses.

The information regarding the exact location of houses and total animals at each house has been obtained for 78 villages in the Mara region of Tanzania. For illustrative purposes only one village, Buchanchari, will be used for the remainder of this mini-dissertation. In order to achieve at least 70% coverage of the Tanzanian village Buchanchari in the Mara region, a kernel density (discussed in Section 2.1) will be estimated for the distribution of the houses of the village. This density estimate will allow for the identification of areas within Buchanchari with a dense collection of houses and therefore potential stopping points along the driving route. For this mini-dissertation, houses which lie within $200m$ of the stopping point will be included, as it has been decided that houses which lie more than $200m$ away are an unattainable walking distance. Once these stopping points have been located, graph theory (Section 2.2) will be utilised to determine the optimal route to drive so as to have access to as many houses as possible, allowing for vaccination of a minimum of 70% of the Buchanchari animal population against rabies. Graph theory will also be applied to the calculation of the optimal walking path among the houses at a particular stopping point such that the walking distance is a minimum.

2.1 Density estimation

Consider Figure 2.1, a plot of the houses of Buchanchari laid over a Google Earth image of the area. It is not very easy to identify the spatial distribution and where possible stopping points should be. A kernel density map (also known as a heat map) is a useful approach for visualising the spatial distribution of the point data [40].

Kernel density estimation is a non-parametric approach to estimating the probability density function of sample data $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots \boldsymbol{x_n}$. The bivariate kernel density estimator at $\boldsymbol{x}_0$ is defined as [64]:

$$\hat{f}_X(\boldsymbol{x}_0) = \frac{1}{nh^2} \sum_{i=1}^{n} K(\boldsymbol{x}_0, \boldsymbol{x}_i),$$

where $\boldsymbol{x}_0 = (x_{01}, x_{02})$, $\boldsymbol{x}_i = (x_{i1}, x_{i2})$, h =bandwidth and K is the kernel. The bandwidth is sometimes also referred to as the smoothing parameter and exhibits a strong influence on the resulting estimate. When h is chosen too small, the density estimate is *undersmoothed* and for h too large, the estimate is *oversmoothed* to the sample data. There are many functions that can be the kernel such as, Gaussian, Epanechnikov and triangular. The Gaussian kernel function will be used for this density estimate. The Gaussian kernel density estimate is defined as [64]:

$$\hat{f}_X(\boldsymbol{x}_0) = \frac{1}{n(2h^2\pi)} \sum_{i=1}^{n} e^{-\frac{1}{2}(\boldsymbol{x}_0 - \boldsymbol{x}_i)'(\boldsymbol{x}_0 - \boldsymbol{x}_i)}.$$

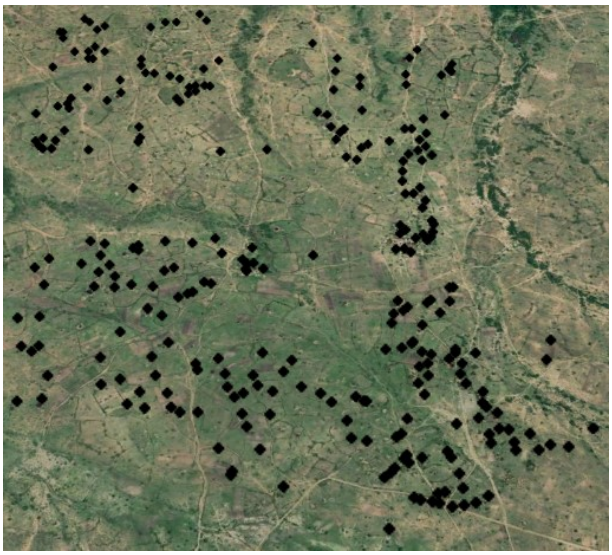


Figure 2.1: A Google Earth image with the locations of the 280 houses in the village of Buchanchari, Tanzania

The `density` function in `R` [59] is used with a convex hull window around the data points to fit the kernel density map. A convex hull is the smallest convex polygon P such that each point of a set Q is either

inside of P or on the boundary [18]. Figure 2.2 illustrates the convex hull computed using `R` for the house locations in Figure 2.1. The heat map is then calculated and plotted for the data. A bandwidth of 0.3 was deemed to be best since any smaller would result in an over-fitted density.

Figure 2.2: Plot of Buchanchari houses in Tanzania within an `R`-generated convex hull

Figure 2.3 shows the density plot of the houses when an adjusted bandwidth of 0.3 is used. It is quite apparent from this image where the hot spots lie and where there are little or no houses. The `locator` function in `R` [59] is used to pinpoint these locations.

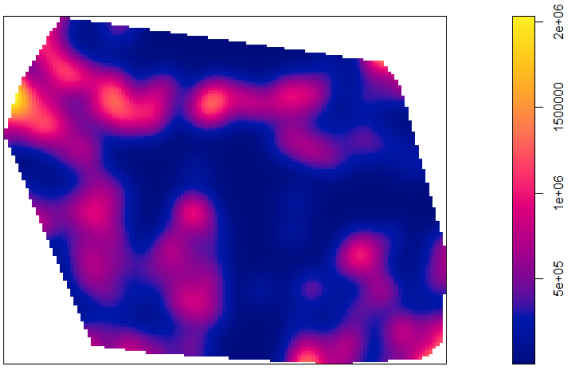


Figure 2.3: Density map of houses in Buchanchari with an adjusted bandwith of 0.3, intensity values are counts per 100 square kilometers

After obtaining these stopping points, it is necessary to determine the road network, as the roads within this village are all informal dirt roads. The digitising of the road network is achieved through Google Earth. The stopping points are then mapped to the road network to ensure that the coordinates align. The road network, houses and stopping points are all illustrated in Figure 2.4.

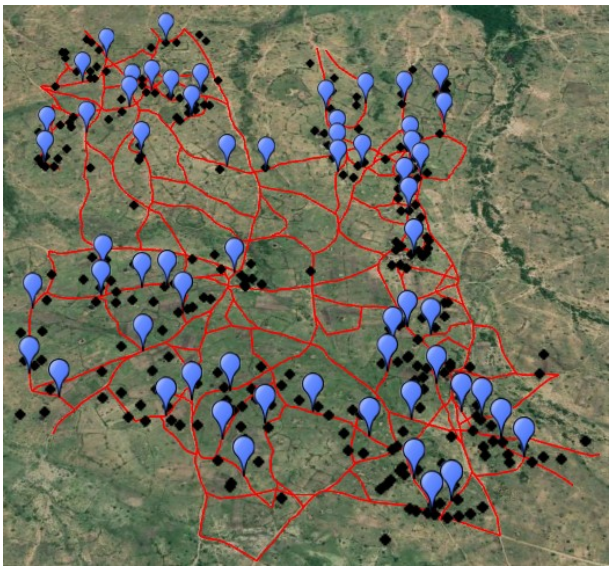


Figure 2.4: Google Earth image of the Buchanchari houses, stopping points and the digitised road network

Now that all the stopping points and their connections have been determined, the optimal path to travel such that at least 70% of the animal population is accessible will be calculated through graph theory.

2.2 Graph theory

Graph theory was first developed in 1736 when Leonhard Euler solved The Königsberg Bridge problem (Königsberg is pictured in Figure 2.5 [1]). This puzzle posed the following question: "Is it possible for a person to walk around the town of Königsberg (now Kaliningrad in Russia) beginning and ending at the same location while crossing each of the seven bridges only once?" Euler's paper, entitled "Seven Bridges of Königsberg", provided a solution using methods that are now referred to as graph theory [29]. Graphs are extremely useful in situations where various pairs of elements are related through some property [70]. For example, electrical networks, telephone communication systems, road maps, oil pipelines and subway systems are all examples of graphs.

[1] http://www-history.mcs.st-andrews.ac.uk/Extras/Konigsberg.html

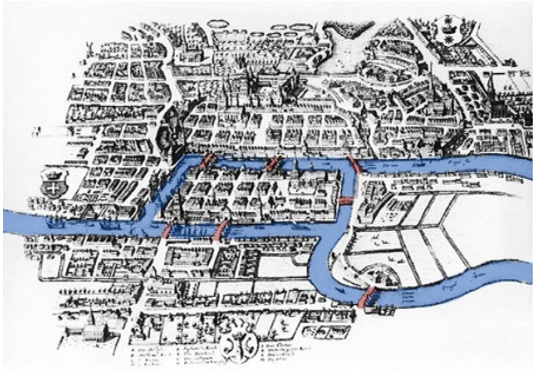
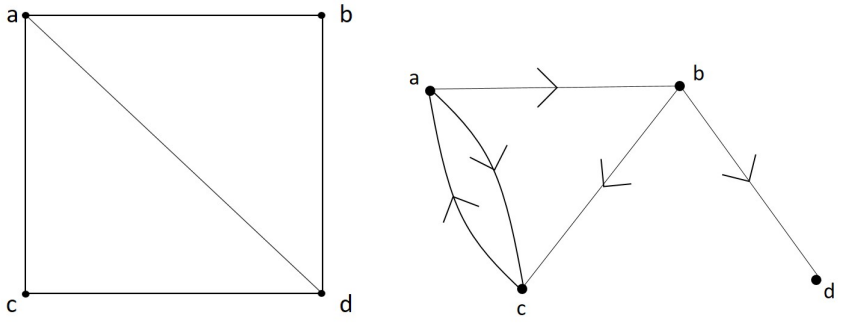


Figure 2.5: The Seven Bridges of Königsberg

A graph is defined as $G = (V, E)$ where V is a finite set of vertices or points and E is a set of edges or lines joining certain pairs of points in V [70]. Consider Figure 2.6a, this *undirected* graph has the set of vertices $V = \{a, b, c, d\}$ and set of edges $E = \{(a, b), (a, c), (a, d), (b, d), (c, d)\}$. This particular graph does not allow two edges to join the same two vertices, nor does it allow loop edges which begin and end at the same vertex. The two ends of an *undirected* edge can be written in any order, i.e. (a, b) or (b, a). It is common to say that vertex a is adjacent to vertex b when there exists an edge from a to b. [70]

Looking at Figure 2.6b, it is apparent that this is a *directed* graph, with vertices $V = \{a, b, c, d\}$. Notice now, that the edges are $E = \left\{(\vec{a}, b), (\vec{a}, c), \left(\vec{b}, c\right), \left(\vec{b}, d\right), (\vec{c}, a)\right\}$. Consider the edge denoted by $(\vec{a}, b)$, this implies that the edge is a path from a to b, notice also that there isn't a path from b to a. The degree of valency of vertex is the number of edges that touch that vertex. For example, the degree of vertex a in Figure 2.6b is 3, denoted as $\deg(a) = 3$.

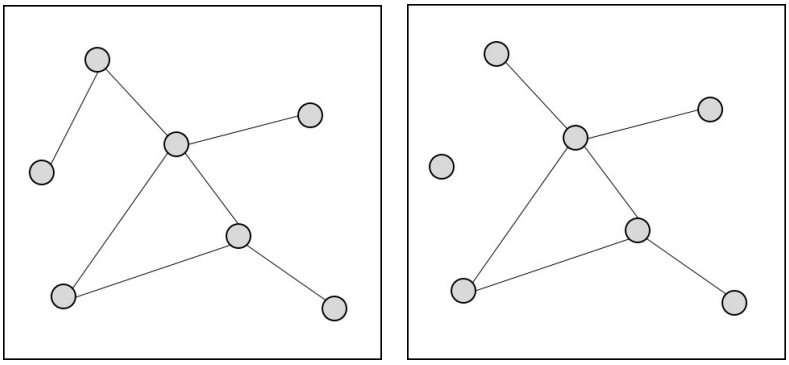


(a) An undirected graph with vertices $V = \{a, b, c, d\}$ and edges $E = \{(a, b), (a, c), (a, d), (b, d), (c, d)\}$

(b) A directed graph with vertices $V = \{a, b, c, d\}$ and edges $E = \left\{(\vec{a}, b), (\vec{a}, c), \left(\vec{b}, c\right), \left(\vec{b}, d\right), (\vec{c}, a)\right\}$

Figure 2.6: Examples of an undirected and a directed graph with vertices V and edges E

Alan Tucker outlines many different examples of graphs in *Applied Combinatorics* [70, page 4-11]. For the sake of this mini-dissertation, the focus will be given to directed graphs known as trees. In graph theory, a tree is a connected graph. When a graph is connected, there exists a path between every pair of vertices, that is to say, that no vertex is unreachable. Figure 2.7



(a) A connected graph (b) A disconnected graph

Figure 2.7: Examples of connected and disconnected graphs in graph theory

2.2.1 Road network

The minimum spanning tree (MST) is among the easiest combinatorial optimisations [1]. Given a set of distinct, positive edge weights E and the complete graph G with vertices V, the aim of the MST problem is to obtain T, the unique subgraph of G with vertex set V (i.e. all vertices of the complete graph are included in the subgraph) that minimises the total weight of the edges, $\sum_{e \in E} w_e$. The most commonly used MST algorithms are Kruskal's and Prim's algorithms.

Kruskal's Algorithm [70]:

1. Order the edges from smallest to largest, in terms of weight.
2. The edge with the smallest weight is added to subgraph T. If the edge forms a circuit with edges already in T, it will not be included and the next smallest edge with be considered.
3. Step 2. is repeated until all the vertices of G are included in T.

Note that a circuit is a sequence of vertices $(x_1, x_2, \ldots x_n)$, where $x_1 = x_n$ and x_i is adjacent to x_{i+1}. A vertex may not appear more than once in a circuit, except for the first and last vertex.

Prim's Algorithm:

1. Given an initial vertex i, the edge with the smallest weight connecting vertex i to another vertex in G is added to subgraph T.
2. At the newly added vertex j, the smallest edge to include another vertex to T is added, provided that the edge does not form a circuit with any other edge already in T. If no such edge exists, the algorithm will include the smallest edge to connect a vertex already in T to another vertex in G.
3. Step 2. is repeated until all the vertices of G are included in T.

If there is a tie for the shortest edge, both algorithms will select either of the edges that form the tie. [70]. The proof for the minimality of Prim's algorithms is presented in *Applied Combinatorics* [70, page 134]. Here Prim's algorithm will be used due to the manner in which successive edges are added to the graph.

The connections between the stopping points along the road network in Buchanchari are illustrated in Figure 2.8, where each line represents travelling in both directions. That is to say that the line between nodes 1 and 2 represents travelling from stopping point 1 to stopping point 2 as well as travelling from stopping point 2 to stopping point 1. The distances between the stopping points are considered negligible since the time taken to walk between houses will have more of an impact on the overall sampling procedure than the time taken to drive between stopping points. The weights of the edges are therefore measured as the number of houses at the destination node. For example, if the driver were to travel from node 2 to 3, he would have access to 3 houses at node 3, therefore the weight of the edge $\left(\vec{2}, 3\right)$ is $\frac{1}{3} \times 100 = 33$ (since graphs aim to minimise the total weight of the edges travelled along, the reciprocal is used). However, should the driver travel along the $\left(\vec{3}, 2\right)$, they will have access to 4 houses at node 2, which means the weight of the edge is $\frac{1}{4} \times 100 = 25$. Since the inclusion of houses that fall within $200m$ is sufficient in obtaining more than 70% coverage of animals in the village, any house that falls more than $200m$ away from a stopping point will be excluded from the sample and are deemed inaccessible. This optimal driving path would also be a logical choice that the driver could make on their own. Seeing that there are more houses along a road would naturally result in the driver deciding to drive along that route so as to samples as many houses as possible.

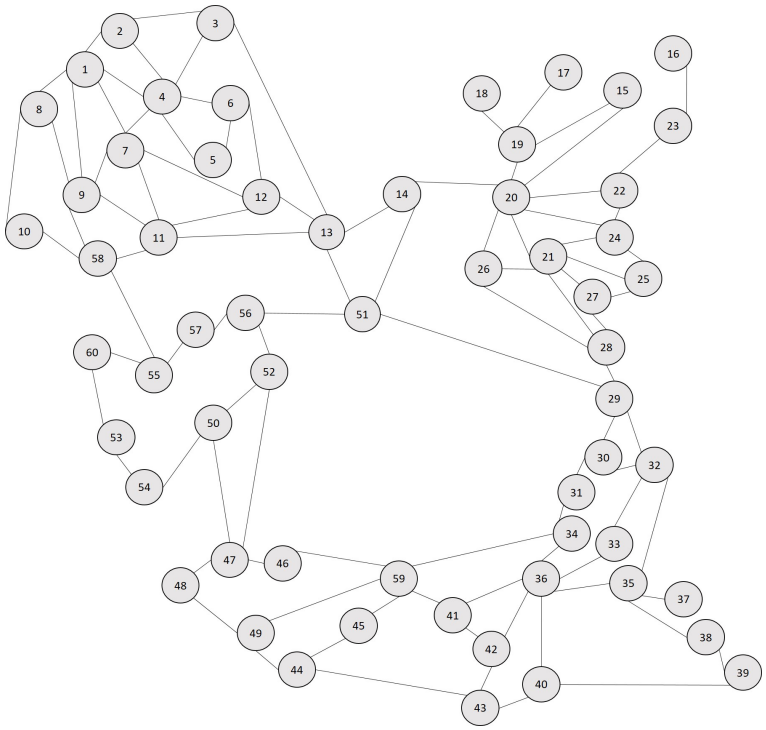


Figure 2.8: A graph of the stopping points and connections in the Buchanchari village

Using the `optrees` package in `R` [31], the MST is determined and is plotted in Figure 2.9. The optimal connections are highlighted in red with a direction. The attainable houses have 614 animals, which is 88% of the population value of 701. Therefore, the stopping points that were developed are sufficient in allowing access to at least 70% of the animals in the village of Buchanchari. The list of houses that are within reach will henceforth be referred to as the accessible population and will be the population from which samples are drawn. This optimal route through the stopping points is an illustration of the possibility of optimisation, however, the driver will be given a list of locations (the stopping points) to reach and may naturally decide which route will be best.

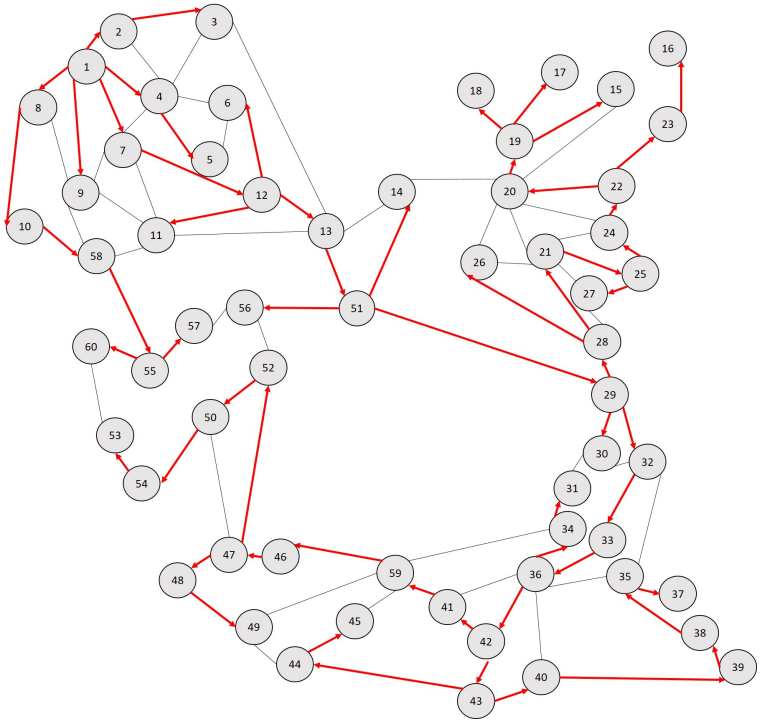


Figure 2.9: An illustration of the minimum spanning tree (MST) for the graph of in the Buchanchari village

2.2.2 Optimal walking route

Now that the percentage of accessible animals in the village is sufficiently larger than the herd immunity value of 70%, samples will be drawn from this accessible population using both traditional (Chapter 3) and spatial sampling (Chapter 4) techniques. In order to minimise the walking time between houses, graph theory will again be used, where the vertices of the graph are the houses and the edges are weighted according to the distance between the houses. The total of these weights will become the cost of the particular sampling scheme and will serve as the determining factor for which sampling approach is most efficient.

Consider the houses around stopping point number 50. There are five houses, the plot in Figure 2.10 shows the house locations as well as the location of the stopping point. These houses and the stopping point will become the vertices of a graph, Figure 2.11a. Figure 2.11b shows the optimal walking path between the houses, beginning and ending at node 1, stopping point 50. This optimal walking path was developed using the `searchWalk` function in the `optrees` package within `R` [31]. The total walking distance from this optimal walk is, therefore, $148 + 87 + 253 + 226 + 251 + 150 = 1115m = 1.115km$.

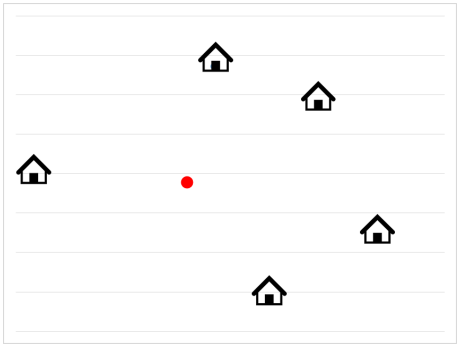
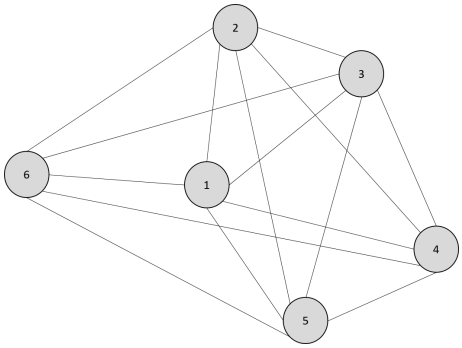
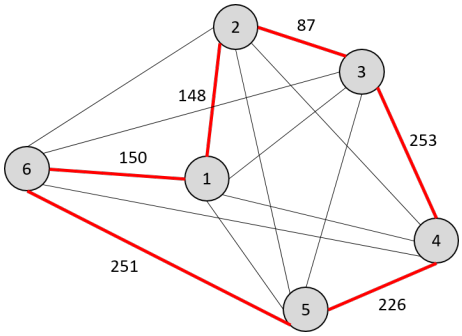


Figure 2.10: Plot of the houses surrounding stopping point 50



(a) Graph representation of stopping point 50, where the houses and the stopping point are the vertices and the edges are weighted according to the distance between the house



(b) Graph representation of stopping point 50, where the houses and the stopping point are the vertices and the edges are weighted according to the distance between the houses

Figure 2.11: The optimal walk between the houses

2.3 Summary

Kernel density estimation was discussed in this chapter and used to determine where the houses are situated in the village of Buchanchari. This estimation allows stopping points to be developed based on the collection of houses in areas. In this chapter, graph theory was also introduced to obtain an optimal road network that can be driven by the vaccinator. This theory is also utilised in subsequent chapters to determine the optimal walking path between sampled houses. The optimal driving path is merely a suggestion and not the focus of the proposed sampling strategy, as it is expected that a logical driver would see a collection of houses and sample to that location for maximum animal vaccination.

In the chapters to follow, traditional and spatial sampling will be discussed and applied to the Buchanchari village in Tanzania. This process of sampling stopping points and determining an optimal walking path

will be followed for all of the sampling strategies presented in this mini-dissertation. The cost of each sampling scheme (total distance walked between houses) will be recorded and analysed so as to determine if a spatial strategy is more efficient than a traditional sampling approach.

Chapter 3

Traditional Sampling

This chapter aims to discuss traditional sampling and its application to the vaccination of animals at the houses of the Tanzanian village of Buchanchari. The concept of traditional sampling is discussed in Section 3.1 with emphasis on notation and terminology. Sections 3.2, 3.3 and 3.4 deal with simple random, stratified and cluster sampling respectively. The theory of these sampling strategies is discussed and used to obtain samples from the Buchanchari village for obtaining 70% vaccination coverage. Within simple random, stratified and cluster sampling (Sections 3.2, 3.3 and 3.4), are application subsections (3.2.1, 3.3.3 and 3.4.4) which provide an illustration of the application of the sampling procedure to the Buchanchari village. The general approach to sampling, as outlined in Chapter 2, is to sample houses according to the specified strategy (simple random, stratified or cluster) taking note of the stopping point each sampled house belongs to. Graph theory is then used to determine the optimal route to walk between houses by constructing graphs at each stopping point. The total walking distance for the entire sample is then calculated and is treated as the cost of the sample. Each sampling strategy is bootstrapped 1000 times using `R` software [59] to obtain distribution functions of the cost, and comparisons between the strategies are made.

3.1 Philosophy of sampling

Sampling has occurred in society since very early human history [74]. The first known attempt to obtain information about a population based on a sample dates back to the English merchant John Graunt and his desire to analyse the London population [20]. Researchers may have once been inclined to only accept the results of a census, however, sampling poses advantages that no longer make this the only feasible option [15].

Some **advantages of sampling** are [15]:

- ✓ reduced cost
- ✓ greater summarising speed of data
- ✓ larger scope and flexibility of the data that can be obtained
- ✓ increased accuracy of results.

When sampling for a study the researcher measures characteristics of the sample observations which are of interest and are denoted by $y_1, y_2, \ldots, y_n$. These characteristics are analysed, estimates obtained and inferences made about the population.

It is important to note that an **estimator** is a formula or rule by which an estimate is obtained (e.g. $\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$) while an **estimate** is the numeric value obtained from the specific sampled data (e.g. $\bar{X} = 206.3$).

The mean square error (MSE) of an estimator $\hat{\theta}$ is:

$$E\left[\left(\hat{\theta} - \theta\right)^2\right] = \text{var}\left(\hat{\theta}\right) + \left[E\left(\hat{\theta}\right) - \theta\right]^2 = \text{var}\left(\hat{\theta}\right) + [\text{Bias}]^2,$$

where the bias of an estimator is defined as the difference between the expected value of the estimator and the value of the parameter [25]. An estimator is unbiased (bias = 0) if the average value of the estimate (obtained over all possible samples of size n) equals the true population value. A biased estimator can be adjusted to make it unbiased. The MSE allows for comparison of biased and unbiased estimators, or for the comparison of two estimators with different biases.

Estimators are **consistent** if as $n \to N$ they approach the true population value. An inconsistent estimator may sometimes be relevant, but only when n is small.

In statistical sampling, two terms are commonly used and need to be formally defined and differentiated in order to proceed.

The **precision** of an estimator $\hat{\theta}$ is the size of the deviation from the **mean obtained through repeated sampling**, that is var $\left(\hat{\theta}\right)$; and **accuracy** is the size of deviation from the **true mean**, that is MSE $\left(\hat{\theta}\right)$ [15].

These concepts will be described using the following example which can be found in Lohr [49, page 32].

Three archers have six attempts at a target. Figure 3.1 shows the shots of Archers A, B and C.

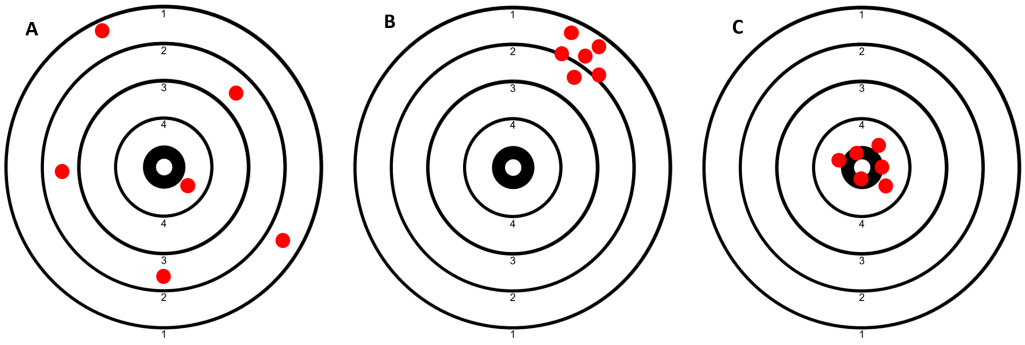


Figure 3.1: An example of an unbiased (A), a precise (B) and an accurate (C) archer

In this example, bull's eye can be interpreted as the true population value. Archer A is **unbiased** since the overall average position of the arrows is the bull's eye. Archer B is **precise** but not unbiased as all the arrows are close to each other but are all consistently away from the bull's eye. Archer C is **accurate** because all of the arrows are close together and near the bull's eye [49].

The probability that a unit i is included in a sample is denoted by π_i. These are known as **inclusion probabilities** (sampling fraction [15]) and are used in probability sampling to calculate point estimates. These inclusion probabilities are known to the investigator before the start of the survey. [49]

The i^{th} **sampling weight** (expansion factor [15]) is the number of population units represented by the i^{th} unit [49]. This is calculated as the reciprocal of the inclusion probabilities, that is $w_i = \frac{1}{\pi_i}$.

Sampling strategies are either **design-** or **model-based**. Design-based strategies are based on classical sampling theory while model-based approaches are based on the assumption of an underlying model [11]. In design-based theory, the only relationship between sampled and non-sampled units is that the non-sampled units could have been included in the sample had the random number generation started at a different random number, while the model-based approach supplies a connection between the sampled and non-sampled units [49]. A model-based sample can obtain a value which is a single realisation of an underlying stochastic process and may never be realised again [74]. Design- and model-based sampling are discussed in more detail in Chapter 4 in the spatial context.

Herein the focus will be design-based models where it is assumed that the n observations of the probability sample come from a finite population j [6]. The y_i's are fixed but unknown values which are an indication of which population units are present in the sample [49] and can be used to estimate the population mean $\bar{y}_U$. This U notation will be used throughout the remainder of this document. It indicates universe and therefore refers, in our case, to the population. Ultimately, the decision between model or design-based strategies depends on whether or not the assumed model is true. If it is, then the model-based strategy is optimal, however, if there is a possibility of misspecification in the model then design-based methods are

preferred [6]. The three traditional sampling techniques which will be discussed in this chapter are simple random sampling (Section 3.2), stratified sampling (Section3.3) and cluster sampling (Section 3.4).

3.2 Simple random sampling

Simple random sampling is a process whereby n units are selected at random from a population of size N. The samples are chosen without replacement such that each combination ($C_n^N = \frac{N!}{n!(N-n)!}$) has an equal chance of being drawn. Consider one sample of size n. When the first observation is drawn for the sample it has an **inclusion probability** $\pi_1 = \frac{n}{N}$ of being drawn. Since this is sampling without replacement, the second observation has an inclusion probability of $\pi_2 = \frac{n-1}{N-1}$, and so on. Therefore, the probability that all n units are selected in n draws is

$$\frac{n}{N} \cdot \frac{(n-1)}{(N-1)} \cdot \frac{(n-2)}{(N-2)} \cdots \frac{1}{(N-n+1)} = \frac{n!(N-n)!}{N!} = \frac{1}{C_N^n}.$$

All the **sampling weights** are the same, therefore each unit represents itself as well as $\frac{N}{n-1}$ of the non-sampled units. This kind of sample is known as a **self-weighting** sample. Sampling with replacement is sometimes viable, particularly for more complex sampling strategies since calculating the variance and the estimated variance is much simpler for sampling with replacement than without.

Cochran [15, page 19] illustrates how random samples are obtained using a table of 1000 random numbers. Other literature, such as Rand Corporation [19] or Kendall and Smith [46] offer tables with 1 million and 100 000 random digits respectively.

The population statistics of interest and their estimators are presented in Table 3.1, while Table 3.2 shows more statistics of interest and their formulae for a simple random sample.

Population Statistic	**Symbol**	**Estimator**
Mean	$\bar{y}_U$	$\bar{y} = \frac{\sum_{i \in S} y_i}{n}$
Total	t	$\hat{t} = N\bar{y}$
Proportion	p	$\hat{p} = \frac{\hat{t}}{n}$

Table 3.1: Population statistics and their estimators for simple random sampling

Statistic description		Formula
Estimated population variance		$s^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$
Standard error of estimated:	population mean	$\sigma_{\bar{y}_U} = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$
	population total	$\sigma_{\hat{t}} = \frac{Ns}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$
Confidence limits:	population mean	$\left(\bar{y} - \frac{z_{\frac{\alpha}{2}}\, s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}, \bar{y} + \frac{z_{\frac{\alpha}{2}}\, s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}\right)$
	population total	$\left(N\bar{y} - \frac{z_{\frac{\alpha}{2}} Ns}{\sqrt{n}}\sqrt{\frac{N-n}{N}}, N\bar{y} + \frac{z_{\frac{\alpha}{2}} Ns}{\sqrt{n}}\sqrt{\frac{N-n}{N}}\right)$

Table 3.2: Estimated variance, standard error and confidence limits of the statistics of a population based on a simple random sample

The quantity $\sqrt{\frac{(N-n)}{N}}$ is known as the **finite population correction**. In practice, this correction can be ignored if $\frac{n}{N} < 0.05$, however, it should be noted that ignoring this correction results in the overestimation of the standard errors. The $z_{\frac{\alpha}{2}}$ used in the confidence limit formulas is the standard normal variable associated with the desired confidence limit, since $\bar{y}$ and $\hat{t}$ are normally distributed around the population values. This fact is discussed by Cochran [15, page 39]. Should the sample size be less than 50, then $z_{\frac{\alpha}{2}}$ may be approximated with $t_{\frac{\alpha}{2},(n-1)}$ taken from the Student's t table with degrees of freedom $(n-1)$. The interested reader is referred to an example on estimation and confidence limits in Cochran [15, page 27].

Estimation of the standard error from a sample is used for [15]:

1. Comparing the accuracy of simple random sampling with other sampling methods
2. Estimating the sample size needed in a survey
3. Estimating the attained accuracy from a completed survey.

Advantages of simple random sampling [15]:

✓ Provides a sample which is highly representative of the population

✓ Allows for statistical inferences regarding the population.

Disadvantages of simple random sampling [15]:

× Only possible if the population list is complete

× Need to ensure an adequate proportion of the sample participates in the research and recontacting non-respondents is time-consuming.

Sampling methods other than simple random sampling are often adopted to make the selection process easier, obtain extra information or to increase the confidence in conclusions [25].

3.2.1 Application

The houses in the Tanzanian village of Buchanchari are presented in a list for sampling. The simple random sample of the houses is drawn using the `sample` function from the `sampling` package in `R` [68] and the size of the sample n is selected such that $\frac{a}{A_{true}} \geqslant 0.7$, where a is the total number of animals at the sampled houses, and $A_{true} = 701$ represents the true number of animals in the village (remembering that only 614 animals are accessible according to the constraints set in Chapter 2).

For each simple random sample that is taken, the number of animals at the included houses is noted so as to determine whether or not the sample meets the 70% coverage requirement. If the number of animals at the sampled houses make up less than 70% of the population, then another house is added in the same manner as a simple random sample to increase the coverage. The sample size n is increased by 1 until 70% coverage is obtained. That is to say that n is selected such that the number of animals sampled exceeds the herd immunity value of 70%. The sampled houses will be treated as nodes in graphs constructed at each stopping point, as was discussed in Chapter 2. The cost of a sample taken is measured as the total optimal walking distance between the sampled houses at a stopping point. Figure 3.2 shows the location of all of the houses in the Buchanchari village, the stopping points as well as which houses are included in the simple random sample.

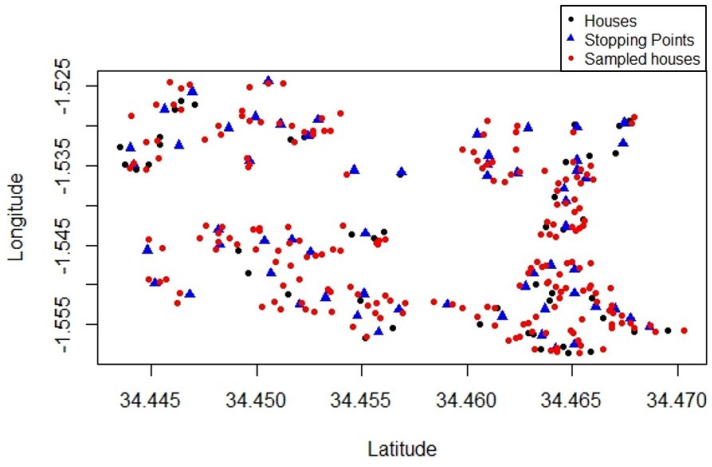


Figure 3.2: Plot of the houses in the Buchanchari village with marked stopping points and an illustration of a simple random sample of the houses to obtain 70% coverage

This sampling procedure is bootstrapped in R software 1000 times to obtain a distribution of the cost as well as the average coverage of the village. Figure 3.3 is a graphical representation of the distribution of the cost function for the simple random sampling procedure with an average cost of $21.91km$ and a standard deviation of $0.98km$. Table 3.3 shows the summary statistics for the cost of simple random samples and the average population coverage of these simple random samples is 70.32%.

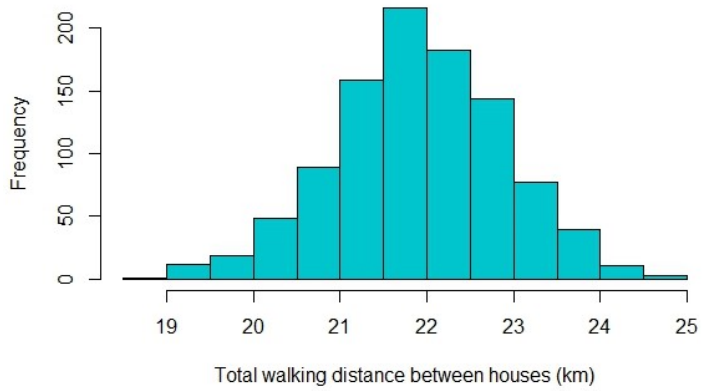


Figure 3.3: Distribution of cost function of 1000 simple random samples

Summary statistics km	
Mean	21.91
Standard deviation	0.98
Minimum	18.94
Q_1	21.30
Median	21.91
Q_3	22.56
Maximum	24.74

Table 3.3: Summary statistics of the cost distribution of simple random samples

3.3 Stratified random sampling

Stratified sampling involves dividing the population of size N into H subpopulations or ***strata*** denoted by $N_1, N_2, \ldots, N_H$ according to stratification variables (defining variables) of the population data. For example, the average weight of males and females in a population is of interest to a researcher, the population would logically be divided into two strata; males and females. These stratification variables need to be correlated with the measurements which define the population characteristics of interest,

implying that the strata are homogeneous within themselves, and must effectively partition the population into the desired subgroups, i.e. major subgroups should be identified by one or more strata [43]. The strata are non-overlapping, therefore $N_1 + N_2 + \ldots + N_H = N$, where N is the number of units in the population [15].

Stratified sampling is a commonly used sampling technique since [49] [14]:

- If specific data are required for different subgroups within a population, each subdivision is then ideally treated as its own "population", reducing the chance of having an unrepresentative sample
- Administrative reasons may push the sampling procedure towards a stratified sampling approach
- Separate strata may pose different sampling problems and are then *grouped* accordingly
- Stratification produces a gain in precision in population estimates
- There is a natural appeal of the dispersion of the sample across the population.

The process of stratified sampling can be described in a few short steps. First, the population is divided into the strata according to the stratification variables. A sample size is chosen and, considering the proportions of the variables that exist in the population, the size of the sample from each stratum is calculated. The samples are then drawn from each stratum using simple random sampling ((Section 3.2)) or systematic sampling ((Subsection 3.4.3)).

Before the estimators of the population statistics can be defined, some new notation needs to be clarified. Tables 3.4 and 3.5 show the new symbols referring to the stratum h.

Symbol	Description
N_h	Total number of units in stratum h
n_h	Number of units in sample from stratum h
y_{hi}	Value obtained for the i^{th} unit from stratum h
$W_h = \frac{N_h}{N}$	Stratum weight of the h^{th} stratum
$f_h = \frac{n_h}{N_h}$	Inclusion probabilities in stratum h

Table 3.4: Stratified sampling symbols

Stratum Statistic	**Symbol**	**Estimator**
Population total in stratum h	t_h	$\hat{t}_h = N_h \bar{y}_h$
Population mean for stratum h	$\bar{y}_{hU}$	$\bar{y}_h = \sum_{i \in S_h} \frac{y_{hi}}{n_h}$
Population Statistic	**Symbol**	**Estimator**
Population total	t	$\hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h$
Population mean	$\bar{y}_U$	$\bar{y}_{str} = \frac{\hat{t}_{str}}{N}$

Table 3.5: Stratum statistics and their estimators for stratified sampling

Table 3.6 shows the estimated population variance of each stratum, as well as the standard error and confidence limits of the overall population mean and total. The z value is representative of the standard normal variable and, if $n < 50$, can be replaced by a student's t statistic with $(n - H)$ degrees of freedom [49].

Statistic description		**Formula**
Estimated population variance of stratum h		$s_h^2 = \frac{\sum_{i \in S_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1}$
Standard error of estimated:	population mean	$\sigma_{\bar{y}_{str}} = \sqrt{\sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h}}$
	population total	$\sigma_{\hat{t}_{str}} = \sqrt{\sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{S_h^2}{n_h}}$
Confidence limits:	population mean	$\left(\bar{y}_{str} - z_{\frac{\alpha}{2}} \sigma_{\bar{y}_{str}}, \bar{y}_{str} + z_{\frac{\alpha}{2}} \sigma_{\bar{y}_{str}}\right)$
	population total	$\left(\hat{t}_{str} - z_{\frac{\alpha}{2}} \sigma_{\hat{t}_{str}}, \hat{t}_{str} + z_{\frac{\alpha}{2}} \sigma_{\hat{t}_{str}}\right)$

Table 3.6: Estimated variance, standard error and confidence limits of the statistics of a population based on a stratified random sample

3.3.1 Number of strata

The stratification variables divide a population into subgroups, such as gender or age. In populations with multiple stratification variables, the most important or most relevant variables are chosen for the sampling procedure [49]. The most useful variables on which to stratify are intuitively the characteristics that are being measured, however, this is not always feasible in practice, therefore the most highly correlated data is the next best option and will reduce the variance of the estimates. Costs involved in obtaining the data need to be considered when deciding on the different strata for the sampling, which usually simplifies the stratification [38]. It sometimes happens that data is available from previous studies and, so long as this information can be trusted and is relevant to the study, it is viable to use this in calculating population estimates.

3.3.2 Allocation to strata

Here, two methods of allocating observations to strata are discussed: **proportional** and **optimal** allocation. Proportional allocation is rather simple and requires no knowledge of the variances of the strata or the relative sampling cost, while optimum allocation needs this information which is usually unavailable [3]. Allocation can also be done based on specified precision within the strata such as a desired margin of error. Pandey and Verma [56] used a mixture of these two allocation methods to sample the impact of a development programme on a population. This so-called 'mixture allocation' was developed in the article *Mixed allocation in stratified sampling* [62].

Proportional allocation results in equal inclusion probabilities for each stratum, since the inclusion probability of element j in stratum h is $\pi_{hj} = \frac{n_h}{N_h} = \frac{n\frac{N_h}{N}}{N_h} = \frac{n}{N}$. For example, if there exists a population of 2000 men and 1200 women, a proportional allocation with a 10% sample will be 200 men and 120 women so that each man and each woman in the sample represents 10 men or women in the population. The probability of an observation being included in the sample is still the same as under simple random sampling methods ($\frac{n}{N}$) however, the risk of having a sample with only men or only women has been removed. Proportional allocation is best used when the variances S_h^2 can be assumed to be approximately equal to each other, otherwise optimal allocation will be used. [49]

Optimal allocation considers the most cost-effective sample size and results in better estimation than that of proportional allocation [43]. The total cost function used in this process is: $C = c_0 + \sum_{h=1}^{H} c_h n_h$, where c_0 denotes the overhead cost and c_h the cost associated with stratum h [4].

Allocation of stratum sizes (n_h) is chosen to either: a) minimise $\sigma^2_{\bar{y}_{str}}$ given a certain cost:

$$n_h = \frac{\frac{(C-C_0)W_h s_h}{\sqrt{c_h}}}{\sum_{h=1}^{H} W_h s_h \sqrt{c_h}},$$

or b) minimise C for a given $V = \sigma^2_{\bar{y}_{str}}$

$$n_h = \frac{\sum_{h=1}^{H} W_h s_h \sqrt{c_h}}{V + \frac{1}{N}\sum_{h=1}^{H} W_h s_h^2} \frac{W_h s_h}{\sqrt{c_h}}.$$

For the interested reader, the derivation of these formulas can be found in Barnett [4, page 118].

It may occur that the sampling costs are equal, then $C = c_0 + nc$. The optimum allocation then becomes: $n_h = \frac{W_h s_h}{\sum_{h=1}^{h} W_h s_h} n$ and $n_h = \frac{\sum_{h=1}^{H}(W_h s_h)\cdot W_h s_h}{V + \frac{1}{N}\sum_{h=1}^{H} W_h s_h^2}$ for cases (a) and (b) respectively. This special case is referred to as **Neyman allocation** [4].

The ultimate goal of stratification is to have a lower within strata variance and a higher between strata variance. That is, the observations in a stratum need to differ slightly from each other and differ significantly enough from observations in the other strata.

Advantages of stratified sampling [38, 49]:

- ✓ Better representation of measurements which are to be estimated
- ✓ Promote adequate sample size for analysis of specific subgroups e.g. race/gender/age
- ✓ Variance is reduced when dividing the population into subgroups
- ✓ Effective for populations with extreme values
- ✓ Increases precision.

Disadvantages of stratified sampling [49]:

- × A complete list of the population is required
- × Each unit in the population must belong to only one stratum
- × Adds complexity to the sampling procedure.

3.3.3 Application

The houses in the Buchanchari village are sampled using stratified sampling with proportional allocation. The stopping points and the houses in the vicinity of that stopping point are interpreted as the strata so that $h = 60$, and each house is assigned an inclusion probability according to the number of houses at the stopping point. Once again, the sample size n is selected such that at least 70% of the animals in the village are sampled and therefore vaccinated against rabies. The `strata` function from the `sampling` package in `R` is used to obtain the optimal samples and the graphs at the stopping points are again constructed with the edge weights being the distance between the houses. Figure 3.4 is a plot of the sampled houses according to the stratified sampling procedure.

The process of stratified sampling is bootstrapped 1000 times so as to obtain a distribution of the cost of the sampling procedure. Figure 3.5 is a histogram of the total walking distance for 1000 stratified samples generated in `R`. The average walking distance between houses for stratified sampling is $21.34km$ with a standard deviation of $0.55km$ and the average coverage of the animal population is 71.53%. Table 3.7 shows the summary statistics for the cost of stratified samples.

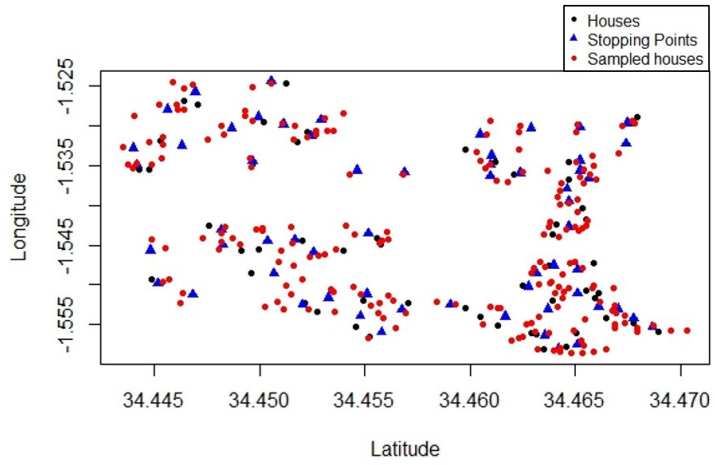


Figure 3.4: Plot of a the houses in the Buchanchari village with marked stopping points and an illustration of a stratified sample of the houses to obtain 70% coverage

Figure 3.5: Distribution of cost function of 1000 stratified samples

The difference between the simple random and stratified sampling approaches are that for the simple random samples, some of the stopping points will not be visited at all since none of their surrounding houses formed part of the sample, whereas with stratified sampling there is always at least one house that will be sampled at each stopping point. The stratified sampling approach yields only slightly better results than simple random sampling in terms of the coverage as well the average walking distance between houses.

Summary statistics km	
Mean	21.34
Standard deviation	0.55
Minimum	18.89
Q_1	20.99
Median	21.35
Q_3	21.73
Maximum	22.84

Table 3.7: Summary statistics of the cost distribution of stratified random samples

3.4 Cluster sampling

Individual units within a population may not be well-defined, even if the target population is, therefore the need arises for another sampling strategy, namely cluster sampling. Cluster sampling can either be performed as one, two or multistage sampling. One-stage cluster sampling is when every element within a selected cluster is included in the sample, while two-stage cluster sampling is when the selected clusters are subsampled so that only a portion of the selected cluster's elements are included in the sample [49]. A practical example of multistage cluster sampling is a survey of students in a province; a sample can first be taken of schools in the province, then samples of homeroom classes within the selected schools, and finally samples of students within selected homeroom classes [44].

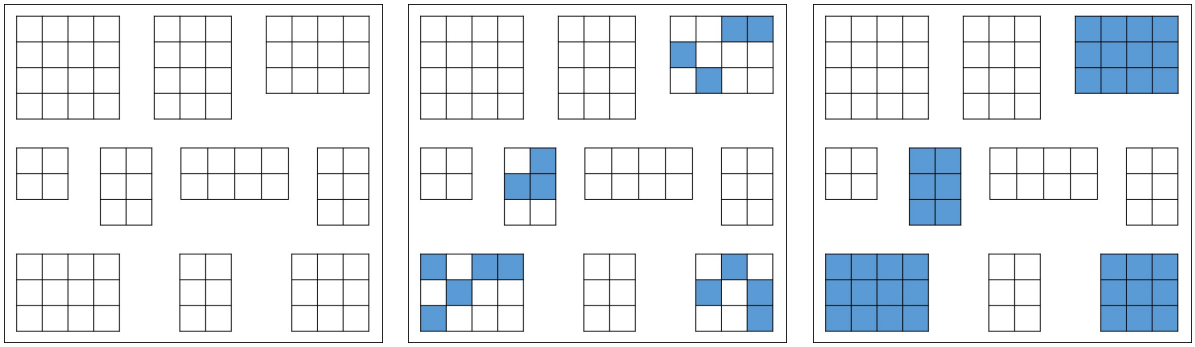
Before going into detail regarding the two types of cluster sampling, notation needs to be defined. In simple random sampling, the units that are sampled and the elements that are observed are one and the same. However, in cluster sampling, the sampling units are the clusters or **primary sampling units (*psu*)** and the elements observed are the elements within the clusters or the **secondary sampling units (*ssu*)** [49]. Table 3.8 shows the population statistics and their estimators for cluster sampling. U is the population of N *psu's* and S is the sample space consisting of *psu's* chosen from the population. S_i is the sample of *ssu's* chosen from the i^{th} *psu*. Therefore, y_{ij} is the j^{th} element from the i^{th} *psu*.

Symbol	Description
Population - primary sampling units	
N	number of *psu's* in the population
M_i	number of *ssu's* in *psu* i
$M_0 = \sum_{i=1}^{N} M_i$	total number of *ssu's* in the population
$t_i = \sum_{j=1}^{M_i} y_{ij}$	total in *psu* i
$t = \sum_{i=1}^{N} t_i$	population total
$S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(t_i - \frac{t}{N}\right)^2$	population variance of the *psu* totals
Population - secondary sampling units	
$\bar{y}_U = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$	population mean
$\bar{y}_{iU} = \frac{t_i}{M_i}$	population mean in *psu* i
$S^2 = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_U)^2}{M_0 - 1}$	population variance (per *ssu*)
$S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{iU})^2}{M_i - 1}$	population variance within *psu* i
Sample	
n	number of *psu's* in the sample
m_i	number of *ssu's* in the sample from *psu* i
$\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m_i}$	estimated mean of *psu* i
$\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij}$	estimated total for *psu* i
$\hat{t}_{unb} = \sum_{i \in S} \frac{N}{n} \hat{t}_i$	unbiased estimator for population total
$s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{unb}}{N}\right)^2$	sample variance of population total
$s_i^2 = \sum_{j \in S_i} \frac{(y_{ij} - \bar{y}_i)^2}{m_i - 1}$	sample variance within *psu* i
w_{ij}	sampling weight for *ssu* j in *psu* i

Table 3.8: Cluster sampling symbols for population and sample statistics

3.4.1 One-stage cluster sampling

One stage cluster sampling is closely related to stratified sampling, however, instead of selecting observations from within each stratum, observations will be selected from a full stratum (now known as a cluster) [3]. That is, either all or none of the elements from a cluster are included in a sample. One-stage cluster sampling is used when the cost involved with sampling the *ssu's* is negligible compared to the costs involved in sampling the *psu's* [49]. Figure 3.6 compares the idea behind stratified and cluster sampling. Figure 3.6a shows the strata in stratified sampling and the clusters in cluster sampling [49]. Each block represents a stratum or a cluster and each square within each block is an observation belonging to the particular stratum or cluster. In Figure 3.6b, the blue blocks represent the observations which will be



(a) Possible subgroups in the population (b) Stratified samples (c) Cluster samples

Figure 3.6: Stratified and cluster sampling

included in the samples from stratified sampling and in Figure 3.6c, the observations from the cluster sampling. The distinction between stratified sampling and cluster sampling becomes clear here; for stratified sampling, observations are selected (via simple random sampling [4]) from every stratum, while in cluster sampling, an entire cluster is selected to be included in the sample. Sampled elements, therefore, need to be representative of unsampled elements and so the elements within the clusters need to be heterogenous [44].

In stratified sampling, the variance of $\bar{y}_U$ depends on the *within* strata variance, while in cluster sampling, the variance of $\bar{y}_U$ depends on the variability *between* clusters [49].

The *psu's* in cluster sampling can either be of equal or unequal size. Unequal sized clusters are more common in real-life sampling, except in the case of agricultural and industrial sampling, where equal sized clusters are obtainable [49].

Equal sized clusters:

In terms of estimation under equal sized clusters, the same concepts as simple random sampling will be followed, where the *psu* totals (or means) are used as the observations [49]. The intraclass correlation coefficient (ICC) is a measure of how similar elements in the same cluster are to each other [49]. In the case of equal sized clusters, $M_i = M, \forall i \in S$.

$$ICC = 1 - \frac{M}{M-1}\left(\frac{SSW}{SST}\right),$$

where

$$SSW = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \bar{y}_{iU}\right)^2 = \text{Sum of errors within } psu\text{'s}$$

$$SST = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \bar{y}_{U}\right)^2 = \text{Sum of total errors about } \bar{y}_U.$$

Lohr [49, page 174] shows that $-\frac{1}{M-1} \leq ICC \leq 1$. Therefore, if the elements in a cluster are perfectly homogeneous (that is $SSW = 0$), then $ICC = 1$. Also, if the ICC is positive, this implies similarity within the cluster and it would be more efficient to use simple random sampling. If the ICC is negative, the elements within a cluster are more dispersed than a randomly chosen group would be.

Unequal sized clusters:

The variation among individual cluster totals t_i will most probably be large for unequal sized clusters (different number of *ssu's*). Ratio estimation is used to introduce an alternative estimator for $\bar{y}_U$ for clustering with unequal clusters [49]:

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}}{\sum_{i \in S} \sum_{j \in S_i} w_{ij}},$$

where the standard error of the estimator is [49]

$$\sigma_{\hat{\bar{y}}_r} = \sqrt{\left(1 - \frac{n}{N}\right) \frac{1}{n\bar{M}^2} \frac{\sum_{i \in S} M_i^2 \left(\bar{y}_i - \hat{\bar{y}}_r\right)^2}{n-1}}.$$

If M_0 is known, ratio estimation can be used to estimate the population total; $\hat{t}_r = M_0 \hat{\bar{y}}_r$. $\hat{t}_{unb}$ doesn't have the limitation that $\hat{t}_r$ (Table 3.8) does.

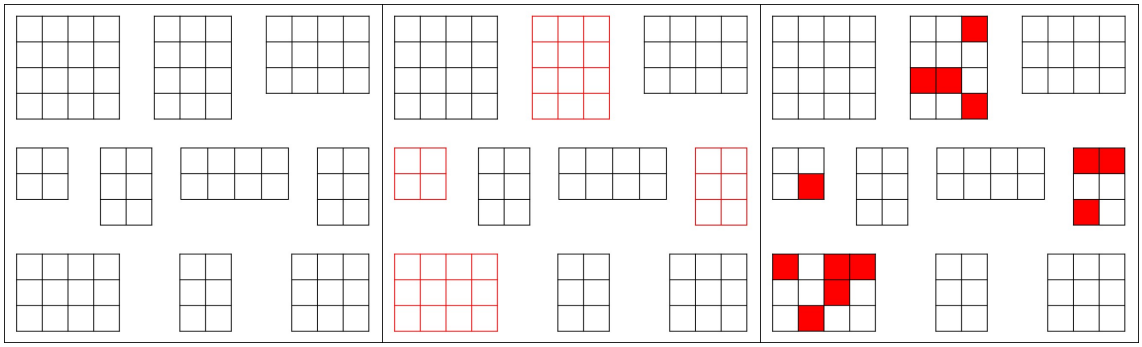
ICC cannot be calculated for clusters of unequal sizes, therefore an alternative measure of homogeneity is the adjusted R^2 denoted by R_a^2 which is interpreted as the relative variability in the population explained by the *psu* means, adjusted for the degrees of freedom [49]:

$$R_a^2 = 1 - \frac{\frac{SSW}{N(M_0 - 1)}}{S^2}.$$

Note: R_a^2 can be negative. This implies elements in the same cluster are less similar than randomly selected elements from the population [49]. That is, cluster sampling is more efficient than simple random sampling.

3.4.2 Two-stage cluster sampling

In many situations, the elements within a cluster are similar resulting in sampling of all subunits within a *psu* being a waste of resources and possibly expensive [49]. Therefore, subsampling within a *psu* is beneficial. Two-stage cluster sampling involves: (1) Selecting a simple random sample S of n *psu's* from the population of N *psu's*, (2) selecting a simple random sample S_i of *ssu's* from each selected *psu*.Figure 3.7 explains the process of two-stage cluster sampling. Figure 3.7a shows the N clusters in the population, Figure 3.7b is a simple random sample of n *psu's* (clusters) and Figure 3.7c shows how within each of the sample *psu's*, another simple random sample is taken of m_i *ssu's* in *psu* i.



(a) Clusters for two-stage cluster sampling

(b) Sampled clusters for two-stage cluster sampling

(c) Sampled *ssu's* of sampled clusters for two-stage cluster sampling

Figure 3.7: Two-stage cluster sampling

In two-stage sampling, the weight of an element is $w_{ij} = \frac{NM_i}{nm_i}$ which is interpreted as element j in *psu* i representing w_{ij} elements in the population. The interested reader is referred to Lohr [49, page 184] for the mathematical explanation.

Advantages of cluster sampling [4, 30]:

- ✓ Provides more information per unit cost than any other sampling method
- ✓ Administrative convenience.

Disadvantages of cluster sampling [47, 30]:

- × Produces less precise estimates than simple/stratified random sampling
- × High probability of sampling error since a significant portion of the population is unsampled
- × Need to recruit more study samples.

3.4.3 Systematic sampling

A special case of cluster sampling is when $n = 1$, that is the total number of clusters sampled is 1, this is known as systematic sampling. Systematic sampling involves ordering and labeling the observations of a population from 1 to N. Selecting a sample of size m requires taking a random observation from the first k ordered observations and then every k^{th} observation thereafter. To illustrate, suppose it is decided that every 10^{th} person who walks into a shop will be asked to complete a questionnaire, starting at person number 3. Then, since $k = 10$, the next observation included in the sample will be the 13^{th} person, then

the 23^{th} and so on. This concept is illustrated by Figure 3.8 and is known as an *every* k^{th} systematic sample [15]. The observations (customers) have been colour coded according to the cluster they belong to, that is to say, that since observation number 1 and 11 belong to the same cluster they have both been coloured blue. Observations 2 and 12 also belong to the same cluster and are both orange and so on. The cluster that will be sampled for this example is the green one, namely observation 3, 13 and so on.

Figure 3.8: An example of systematic sampling

Another method suggested by Madow [52] is selecting the observation at the center of the 'stratum'. This is efficient when the population exhibits a monotone decreasing autocorrelation plot (correlogram) [52]. Cochran [15] proved that random start systematic sampling is more efficient than stratified sampling when the population has a concave upwards and decreasing autocorrelation plot. Madow then pointed out that centered systematic sampling is more powerful than random start systematic sampling under these same conditions [52].

Advantages of systematic sampling [15]:

- ✓ Drawing a sample is easier and often has fewer mistakes than simple random sampling
- ✓ More precise than simple random sampling and, in certain instances, stratified sampling.

Disadvantages of systematic sampling [49]:

- × Does not necessarily result in a representative sample
- × Does not give a truly random sample
- × Many of the inclusion probabilities are zero
- × Sampling in line with any cyclical or seasonal patterns will result in inadequate estimates.

3.4.4 Application

One-stage proportional cluster sampling is performed on the Buchanchari village. The stopping points are the clusters in this approach and each cluster is assigned an inclusion probability proportional to the number of houses accessible via the particular cluster or stopping point. For example, there are 7 houses within $200m$ of stopping point 10. The inclusion probability of stopping point 10 as a cluster in the sample is, therefore, $\frac{7}{247} = 0.03$. Figure 3.9 is a plot of the houses in the Buchanchari village, the stopping points or clusters and the sampled houses using the cluster sampling approach. The red points are those houses that are included in the sample and the black points are houses in the village of Buchanchari that are not included in the sample.

The reason why one-stage cluster sampling is used is that it seems illogical to be at a stopping point and not sample all of the houses within its vicinity. One can argue that this occurs during stratified sampling too, however the supporting argument is that with stratified sampling, all stopping points or strata will be included in the sample, while with cluster sampling, some stopping points or clusters will not be included, so to sample a step further and still obtain 70% coverage will require more clusters to be included in the sample, increasing total walking distance. Since walking distance is considered more costly than the driving distance in this application, one-stage sampling is the preferred approach to cluster sampling.

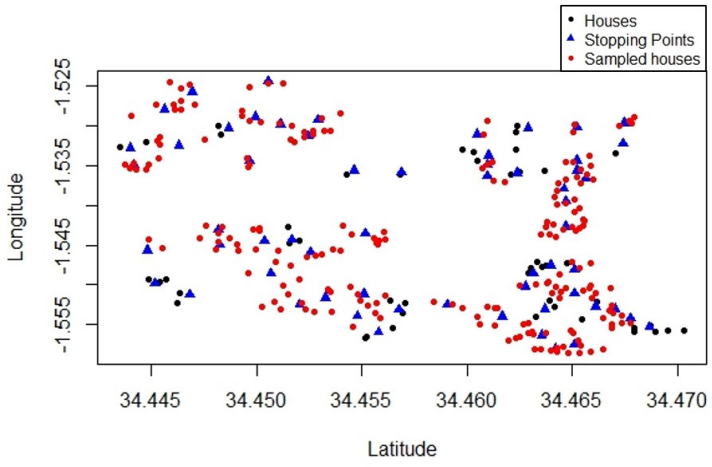


Figure 3.9: Plot of a the houses in the Buchanchari village with marked stopping points and an illustration of a cluster sample of the houses to obtain 70% coverage

Figure 3.10 illustrates the distribution of the cost of cluster sampling in the village. Table 3.9 describes the summary statistics of the cost of 1000 cluster random samples. The average cost of the cluster random samples is $23.86km$ with a standard deviation of 0.9. The average animal coverage of these samples is 72.23%. The cost of the cluster samples is higher than the stratified or simple random sampling. This is

because the clusters or stopping points with more houses have a higher probability of being included in the sample and, since they have more houses within their $200m$ radius, will have more walking between the houses.

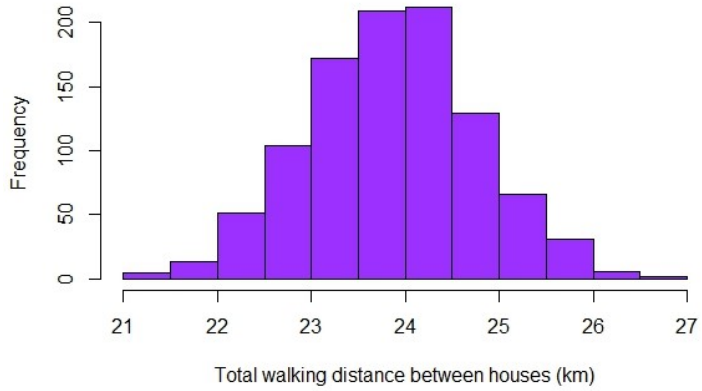


Figure 3.10: Distribution of cost function of 1000 cluster samples

Summary statistics km	
Mean	23.86
Standard deviation	0.9
Minimum	21.02
Q_1	23.26
Median	23.88
Q_3	24.44
Maximum	26.84

Table 3.9: Summary statistics of the cost distribution of cluster random samples

Comparing Figure 3.4 and Figure 3.9, the distinction between stratified and cluster sampling is apparent. In Figure 3.4, the stopping points in the middle of the plot are included in the sample since stratified sampling includes all strata, whereas in Figure 3.9, the middle stopping points are not included since there is only one house within their vicinity and will therefore have a lower probability of being included in the sample as a cluster.

3.5 Summary

In this chapter, approaches to traditional sampling were discussed, namely: simple random, stratified, and cluster sampling. Samples of houses were obtained from the Buchanchari village through simple random, spatial and cluster sampling techniques. For stratified and cluster samples, the stopping points are the strata and clusters respectively. The cost of each sampling strategy is measured as the walking distance between houses which is minimised using graph theory. The sampling strategies were all bootstrapped 1000 times so as to plot the distributions of the cost of the sampling procedures and compare them. Cluster sampling resulted in the largest mean animal coverage, 72.23% but also in the highest average walking distance, $23.86km$. Stratified samples had a slightly less average walking distance compared to simple random samples, as well as a lower variance. While these approaches all yielded similar results, none of them takes into consideration the spatial component of the data. Therefore, in Chapter 4, the spatial characteristic will be accounted for by taking spatial random samples as opposed to traditional samples.

Chapter 4

Spatial Sampling

Classical sampling assumes that the sampled data is independent and identically distributed. For this reason, different sampling strategies need to be used for spatial data where autocorrelation is present between observations. Including two points in the sample which are geographically very close and have no covariate information justifying them to provide significant observed values is not a desirable property of a sample.

This chapter outlines the theory behind spatial sampling and presents strategies which are closely linked to the traditional sampling approaches discussed in Chapter 3. The strategies discussed are, uniform (simple) random Section 4.3, stratified Section 4.4, and cluster sampling Section 4.5. The costs of each of the sampling strategies are measured as the total walking distances along the optimal walking paths between the sampled houses as determined by graph theory. Each of these spatial sampling techniques are then bootstrapped 1000 times to, once again, obtain distributions of the costs of the sampling procedures. Comparisons are made in this chapter between traditional and spatial sampling as well as comparisons of the spatial sampling techniques.

4.1 Why use spatial sampling?

Suppose a researcher is interested in determining the variety of plant-life that is present in a forest and that n sites of a uniform random sample are chosen independently, as for simple random sampling in Chapter 2, each with a uniform distribution over the region R. That is to say that each point has an equal and independent probability of being sampled [60]. Practically, two random numbers K_i and K_i' are generated from a Uniform$(0, 1)$ distribution [24]. The point $u_i = (x_i, y_i)$ is then sampled such that $x_i = K_i L$ and $y_i = K_i' L$, where L denotes the length of the study area [24]. This procedure is repeated

until m samples are obtained [24]. Figure 4.1 is an illustration of a uniform random sample of size $n = 10$ where 10 K_i and K_i' independent random uniform variables are generated and the resulting $\boldsymbol{u_i} = (x_i, y_i)$'s are plotted on the xy-axis.

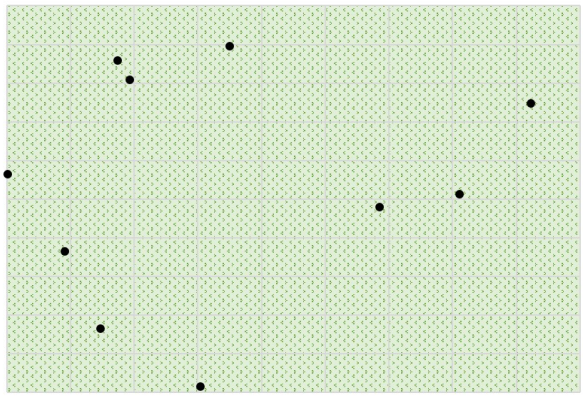


Figure 4.1: Scatter plot of independently selected uniform random pairs plotted over the region of interest

Suppose now that the locations of the population of plants in the area are available and different plant species are depicted by different colours in Figure 4.2. It is evident that many different species have been left out of the sample; an entire group in the top left corner has not been included in the sample, nor have the darker plants in the center.

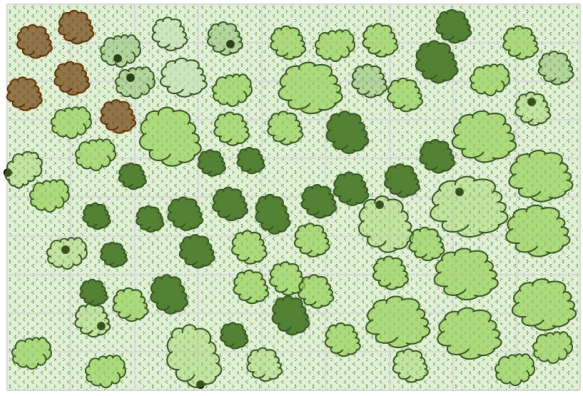


Figure 4.2: Plot of the population of plants in the region of interest

Upon further investigation, it is discovered that there is, in fact, a river running through this forest (Figure 4.3), which offers a possible explanation for the apparent grouping of the darker trees along what is now known to be a river bank. Neighbouring trees will undoubtedly compete with each other for nutrients and are influenced in the same way by the physical features of the landscape as well as any pollutants that

may be present in the air or perhaps the water supply. As with most spatial data, nearby units interact with each other and this information is pertinent to the efficiency of the sampling plan for the region.

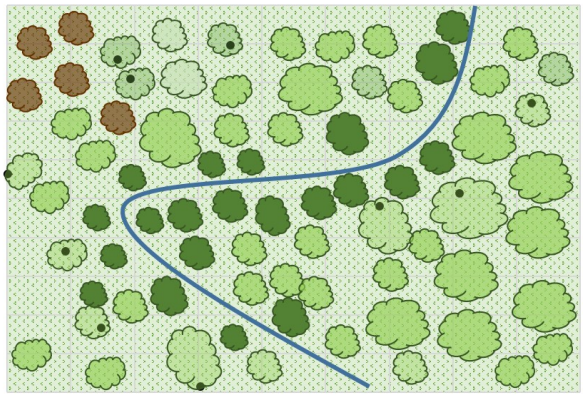


Figure 4.3: Plot of the population of plants in the region of interest as well as the river

Consider now a sampling of streams and rivers for the purpose of water quality and biological assessment as is done by the Indiana Department of Environmental Management. In traditional sampling, the resources can be represented by discrete units by taking fixed length intervals of the waterbody. Performing a simple random sample of these units may result in an unbalanced sample. Longer rivers will have more units and therefore, have an increased probability of being sampled, while shorter streams will have a reduced probability. Unequal probability sampling is a possible solution, however, discretising the points of the population ignores the very nature of the data; one-dimensional continuous observations within a two-dimensional space. Stevens and Olsen [66] suggested viewing the rivers or streams as linear networks, thereby preserving the continuous nature of the data and producing a population of uncountably infinite points. [66]

In the agricultural industry monitoring of soil nutrients is an ongoing process, however, the entire population of soil in a field cannot be practically sampled since this is an example of continuous, two-dimensional data and it is generally inconceivable to obtain observations for an entire region due to financial, administrative and practical constraints [21]. In one field it is also possible to have a high spatial variation of the soil; pH levels, organic matter or nutrient content. As a result, an overall mean for the entire area will not suffice an estimate for the soil nutrients. In order to incorporate the 'spatial component' of data, it is possible to do stratified sampling with unequal probabilities [49]. That is to say, given the coordinates of a set of observations, the distance between observations can be measured and recorded and then observations with larger distances between them may be assigned higher inclusion probabilities than observations that are closer to each other. By assigning unequal probabilities to sample units, the sampling variance is decreased [49]. Recall that selection bias, occurs when sample units are unintentionally sampled at different rates [49]. An example of selection bias could be if a sample of students is acquired by standing

in front of the library and asking every fifth student who walks by to answer a questionnaire. Students who are at the library often are more likely to be included in the sample while those who do not come to that side of campus will not be included. It is not possible to know how many students are represented by the sample chosen and therefore no way of correcting for the unequal probabilities in the estimates [49]. In unequal sampling selection bias will arise in the case of using unequal probabilities that are unknown and cannot be estimated.

Walvoort et al. [73] also suggested stratification of the region, specifically using the k-means algorithm (see Section 4.4).

What should be noticed about these examples is that the spatial distribution of geographical data should not be ignored when designing a survey or attempting to monitor a process in an efficient manner [66]. By performing the appropriate sampling strategy, the researcher can ensure a sample which is representative of the spatial population. For spatial sampling there are three major distinctions between the different types of populations obtainable [66]:

1. Zero-dimensional (point-like,finite)
2. One-dimensional (linear)
3. Two-dimensional (areal).

Finite populations are those with discrete, distinctly identifiable units. Examples of such populations are trees in a forest with a characteristic of interest or perhaps an attribute of lakes. In such an instance, the centre point of the lake could be used as the location for sampling purposes. Linear populations are those observations that are only found in a linear network, such as streams or rivers or roads. These are often sampled as finite observations by dividing the network into discrete units. Finally, areal resources are continuous points that are present everywhere within the bounded area such as soil or air. [66]

The purpose of spatial sampling is to make inferences regarding a population whose elements have a location parameter. It is either required to estimate global parameters of an area, such as the mean crop yield, or to construct maps and make predictions about an area, such as rainfall [37]. Another call for spatial sampling is the detection of extreme values of an attribute in a region [37], for example, areas which have been contaminated. Spatial autocorrelation is inherent in spatial populations [74], explaining the need to take Tobler's law (see Chapter 1) into account when working with spatial data. Spatial autocorrelation disobeys the important assumption of independence in traditional sampling, therefore new sampling techniques are developed which are very specific to spatial data.

In general, there is a continuous surface of observed characteristics $Z(\boldsymbol{u})$ where $\{\boldsymbol{u_i} \in A; i = 1, \ldots, n\}$ for which the mean (or total) of the domain space (population space) A are of interest [60, 67]. $\boldsymbol{u_i} = (x_i, y_i)$ is a spatial longitude and latitude coordinate.

For example, if it is desired to measure ground-level air quality, there are an infinite number of possible locations where the measurements can be taken [37], therefore, it is up to the researcher to choose a sample of size n by selecting n points within A in such a way that the spatial inhomogeneities are accounted for. The values Z can then be measured at each of the points.

Within `R` [59], there is a package `sp` [8], which is described as *Classes and methods for spatial data*, the function `spsample` performs spatial sampling of point locations on a spatial line or within a spatial polygon [7, 57].

For a discrete population, the value of interest could be, for example, the mean

$$\mu = \frac{1}{N} \sum_{i=1}^{N} Z(\boldsymbol{u_i}), \tag{4.1}$$

while in the case of a continuous population, the mean will be calculated using

$$\mu_A = \frac{1}{|A|} \int_A Z(\boldsymbol{u})\, d\boldsymbol{u}, \tag{4.2}$$

where $|A|$ is the area (or volume, depending on the dimension of the space) of the region of interest [37]. It is important to note that in practice, by sampling, the continuous space is reduced to a discrete space.

When drawing samples from geographical data certain questions need to be answered in order to decide on a sampling design such as [37]:

1. What is being estimated?
2. What sample size will achieve the desired level of precision?
3. What locations should be used in the sample?
4. What estimator should be used?
5. What measures should the sampling seek to minimise?

It is important to know if the characteristic being estimated is a spatial property or not [37]. For example, the mean level of an attribute in an area is considered a non-spatial property since the question being asked is: *how much* rather than *where* [11]. It is this *where* question that leads to map construction and makes spatial sampling necessary.

There are two strategies which can be followed in spatial sampling; design-based and model-based spatial sampling. While both will be introduced here, the focus will be given to design-based spatial sampling for the remainder of the text.

4.2 Design- and model-based spatial sampling

It was previously believed that design-based sampling is unsuitable for study areas which exhibit autocorrelation. However, Brus and de Gruijter [11] concluded that, while design-based sampling strategies closely follow the rules of classical sampling theory (assumes data are independent and identically distributed), *independence* in this context has a different meaning and is determined by the sampling design. The comparative article [11] also determined that the effectiveness and efficiency of the two approaches are dependent on certain factors, namely:

- the reason for sampling
- desire for unbiased estimates
- the need for unique estimates of prediction variance
- quality of the model
- autocorrelation present between observations
- sample size.

Design-based sampling

Design-based sampling involves viewing the population of values in the region as being fixed but unknown [37, 74]. The source of randomness is the random selection of locations through a chosen sampling plan where the probability of selection is determined by the chosen sampling plan and the advantages of the particular sampling plans are dependent on the organisation of the spatial variation in the population [11, 22]. Possible sampling plans are discussed further on in this chapter.

Design-based estimators assign weights to the observations in the sample according to their associated probability of being included in the sample [37]. This results in estimators of quantities such as (4.1) or (4.2) exhibiting poor properties in the absence of spatial information. For example, a design-based sample with one observation per strata is seen in Figure 4.4 [37]. Without any spatial intelligence, the point marked by $\times$ in the middle strata will have to be estimated based on the observation within that same strata, although it would make more sense to use the samples in the neighbouring strata since those observed points are geographically closer to the desired $\times$.

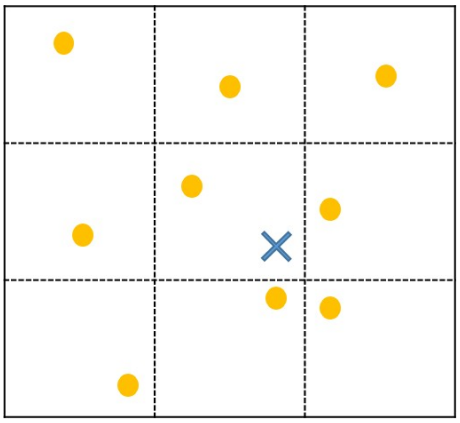


Figure 4.4: Using spatial information for estimation from sample strata

Models which describe the population can be used in design-based sampling. They differ from those used in model-based sampling in that model-based models describe the data generating the process as will be discussed now.

The variables are denoted by $z(\boldsymbol{U}_j)\,(i = 1, 2, \ldots, n)$. So the value z is fixed (lower case) and the location or point $\boldsymbol{U}$ is random, where the randomness is determined by the sampling design p. [22]

The independence (or lack thereof) between $z(\boldsymbol{U}_i)$ and $z(\boldsymbol{U}_j)$ is determined by sampling design and NOT the spatial variation in the region being sampled, R. The inference is therefore based on the sampling design chosen by the researcher and requires no further assumptions. Design-based sampling is useless if for any reason probability sampling is not practical. [22]

Model-based sampling

Model-based sampling strategies are different from design-based in that the population values are regarded as a single outcome of a stochastic model [11, 37, 74]. In other words, the population values of the characteristic of interest are modelled as a stochastic process and the source of randomness is the stochastic model used. This means that the sample need not be selected through some random sampling plan as in design-based strategies since the simulated values are already randomised [11].

Still with the goal of making an inference regarding the mean of the population, (4.1) and (4.2) are seen as the mean of a single realisation. Should another realisation be generated from the model, the value of (4.1) or (4.2) would be different since they are both functions of random variables and are therefore random themselves [37]. Therefore, the characteristics of interest are *predicted* when working with model-based sampling strategies as opposed to being *estimated* with design-based strategies.

The predictors provided through model-based sampling strategies are dependent on the efficiency of the model. It is advisable to use predictors when the underlying structure of the population is known since

this will be the optimal, although rare, scenario [53]. For a more in-depth look into model-based sampling strategies, the reader is referred to [11, 37, 74].

The model-based approach treats the value Z associated with the location $\boldsymbol{u}$ as random, therefore the set of all values is one realisation of a stochastic process. Some features of this stochastic process are known and are represented in a geostatistical model. Models developed for practical purposes account for spatial variation and Tobler's law as was mentioned in Chapter 1. Model-based sampling is not applicable if a sufficient and reliable model cannot be developed for any reason such as due to a lack of data. [22]

Table 4.1 [11] shows the different sources of randomness that arise in sampling and suggests which strategies to follow in each case. When it comes to deciding between which strategy to utilise, the following advice is offered by Robert Haining [37]: model-based strategies should be used for predicting values at locations and estimating the parameters of the underlying model of an area while design-based strategies should be used when estimating global properties of the population such as those in (4.1) and (4.2). de Gruijter and ter Braak [22] found the design-based approach for spatial sampling to be more advantageous than the model-based approach with respect to robustness (the ability of a model to provide insight to the spatial distribution of an area, even if its assumptions are altered or violated).

If both model- and design-based sampling are feasible, then the choice that needs to be made is the quality criterion which alternative strategies will be judged on. One such measure relates to what would happen should the sampling be repeated in the same region but using different sampling arrangements. This leads the researcher to follow design-based sampling techniques. The other criterion is associated with repeated sampling as before, however, the sampling arrangement remains the same and the region being sampled is changed, in this instance, model-based strategies would be desirable. [22]

		Values at given locations	
		Fixed	**Random**
Sample locations	**Fixed**	Fully deterministic	Model-based
	Random	Design-based	Fully random

Table 4.1: Types of sampling strategies defined by two sources of randomness

The focus of this report will be design-based spatial sampling and its applications.

If the characteristic that is of interest is the spatial mean (4.2), then the commonly used estimator is the unweighted sample mean,

$$T_{um} = \frac{1}{n} \sum_{i=1}^{n} z\left(\boldsymbol{U}_i\right).$$

The simple random sampling strategy is said to be p-unbiased, since the expected value of (4.2) denoted by $E_p[T_{um}]$ is equivalent to (4.2). Note that E_p implies expected value over repeated sampling under the sampling design p [22].

This unbiasedness results in equality of the p-MSE and p-variance of 4.2:

$$E_p = [T_{um} - \mu_A]^2 = V_p(T_{um}) = \frac{\sigma_A}{n},$$

where σ_A is the variance between points in A:

$$\sigma_A = \frac{1}{|A|}\int_A [z(\boldsymbol{u}) - \mu_A]^2 \, d\boldsymbol{u}.$$

In order to increase efficiency and accommodate the practical application of sampling, other sampling designs can be implemented.

4.3 Uniform (simple) random sampling

Simple random sampling in a spatial environment can also be referred to as uniform random sampling [21, 33, 60].

In `R` software [59], there exists a package which consists of classes and methods for spatial data, namely `sp`. Within this package there is a function `spsample` which samples point locations within a square or ring area, or on a spatial line using regular or random sampling [57, page 99]. The geometry used in the methods of this sampling are planar and therefore require coordinate data i.e. (x, y).

The method/theory behind this function is based on the explanation of uniform random sampling in Brian Ripley's *Spatial Statistics* [60]. At the beginning of the chapter, uniform random sampling was explained and the result was plotted.

The error variance of an estimator is the variance due to external factors or measurement error [17], or the mean squared error (MSE), defined in Chapter 2. The error variance or MSE calculation differs for each sampling technique in the spatial environment. For the calculation of error variance the following notation is upheld [60]:

$$\mu = E[Z(\boldsymbol{u})] \qquad \sigma^2 = var[Z(\boldsymbol{u})],$$

and

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z(\boldsymbol{u_i}) \quad \text{is used to estimate} \quad \tilde{Z}(R) = \frac{1}{A}\int_R Z(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u}.$$

In order to calculate the error variance for a uniform random sample, the expected value of the random positioning of the observations $\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_n}$ is derived [60]:

$$E\left[\bar{Z} - \tilde{Z}\left(R\right)|Z\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{A}\int_R Z\left(\boldsymbol{u}\right)\,\mathrm{d}\boldsymbol{u} - \frac{1}{A}\int_R Z\left(_u\right)\,\mathrm{d}\boldsymbol{u} = 0.$$

Also,

$$\begin{aligned}
\mathrm{var}\left(\bar{Z} - \tilde{Z}\left(R\right)|Z\right) &= \frac{1}{n^2}\sum \mathrm{var}\left(X\left(\boldsymbol{u_i}\right)\right) \quad \textit{by independence} \\
&= \frac{1}{n}\,\mathrm{var}\left(Z\left(\boldsymbol{u_1}\right)\right) \\
&= \frac{1}{nA}\int_R\left[Z\left(\boldsymbol{u}\right) - \tilde{Z}\left(R\right)\right]^2\,\mathrm{d}\boldsymbol{u} \\
&= \frac{1}{nA}\int_R\left[Z\left(\boldsymbol{u}\right) - \mu\right]^2\,\mathrm{d}\boldsymbol{u} - \frac{1}{nA^2}\int_R\int_R\left[Z\left(\boldsymbol{x}\right) - \mu\right]\left[Z\left(\boldsymbol{y}\right) - \mu\right]\,\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.
\end{aligned}$$

Therefore,

$$\mathrm{var}\left(\bar{Z} - \tilde{Z}\left(R\right)\right) = E\left[\mathrm{var}\left(\bar{Z} - \tilde{Z}\left(R\right)|Z\right)\right] = \frac{1}{n}\left[\sigma^2 - E\left[C\left(\boldsymbol{X}, \boldsymbol{Y}\right)\right]\right], \tag{4.3}$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, uniformly distributed points in R and $C\left(\boldsymbol{X}, \boldsymbol{Y}\right)$ is the covariance between the points $\boldsymbol{X}$ and $\boldsymbol{Y}$. For R large in comparison to the effective range of the covariance function, the second term will be negligible.

An unbiased estimator for the sampling error variance is $\frac{s^2}{n}$, where $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left[Z\left(\boldsymbol{u_i}\right) - \bar{Z}\right]^2$.

4.3.1 Application

Within `R` there is a sampling function `spsample`, which draws continuous samples from a spatial polygon according to the method specified by the user. For uniform random sampling, the specification is `random`. This will produce n random coordinates within the spatial polygon. The distance between these locations and the positions of the houses in the Buchanchari village are measured and the house which is closest to the location generated by `R` is sampled. This is done in such a way so as to ensure that no house is sampled twice and n is increased in the same fashion as in Section 3.2 to ensure 70% coverage of the animals in the village. Figure 4.5 illustrates the houses of the Buchanchari village with the red points being the houses included in the sample and the black points are the unsampled houses. The blue triangles are the stopping points along the road network.

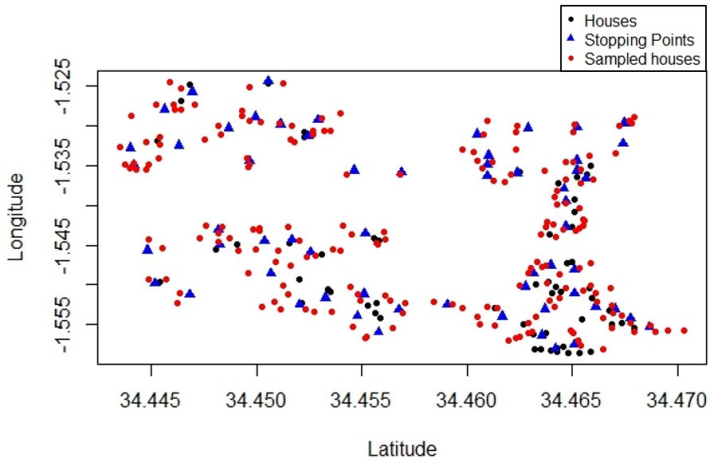


Figure 4.5: Plot of the houses in the village of Buchanchari (black), stopping points along the road network (blue) and the sample houses (red) generated in R

The strategy of sampling randomly from the Buchanchari village is bootstrapped 1000 in R so as to obtain the cost of sampling the houses through this method. Once again, the cost is measured as the total walking distance between houses within the vicinity of the specified stopping points. Figure 4.6 is a histogram representing the distribution of the cost of spatial random sampling in the Buchanchari village. Table 4.2 contains the summary statistics of the cost of this sampling procedure. The average walking distance is $22.41km$ and the average coverage of the animals in the village is 70.87%.

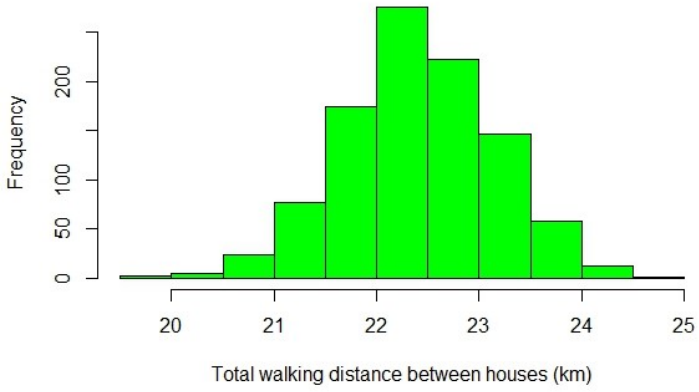


Figure 4.6: Distribution of cost function of 1000 spatial random samples

Summary statistics km	
Mean	22.41
Standard deviation	0.74
Minimum	19.93
Q_1	21.93
Median	22.39
Q_3	22.93
Maximum	24.69

Table 4.2: Summary statistics of the cost distribution of spatial random samples

4.4 Stratified sampling

Stratified sampling is used when local estimates are required [37]. The region of interest R is divided into non-overlapping strata and a simple random sample is drawn from each stratum [21]. Samples are collected from within in strata such that the sum of the samples across all the strata is n. Prior knowledge of the region and the underlying process helps the researcher determine the shape and size of the strata [24].

If R is partitioned into m square strata $S_1, S_2, \ldots, S_m$ each with area s, then the expected value of the random sample is

$$E\left[\bar{Z}|Z\right] = E\left[\frac{1}{m}\sum_{i=1}^{m}\bar{Z}_i|Z\right] = \frac{1}{m}\sum_{i=1}^{m}\tilde{Z}\left(S_i\right) = \tilde{Z}\left(R\right).$$

So that

$$\operatorname{var}\left(\bar{Z} - \tilde{Z}\left(R\right)|Z\right) = \frac{1}{m^2}\sum_{i=1}^{m}\frac{1}{ks}\left[\int_{S_i}\left(Z\left(\boldsymbol{u}\right)-\mu\right)^2 \mathrm{d}\boldsymbol{u} - \int_{S_i}\int_{S_i}\left(Z\left(\boldsymbol{x}\right)-\mu\right)\left(Z\left(\boldsymbol{y}\right)-\mu\right)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}\right],$$

and

$$\begin{aligned}\operatorname{var}\left(\bar{Z} - \tilde{Z}\left(R\right)\right) &= E\left[\operatorname{var}\left(\bar{Z} - \tilde{Z}\left(R\right)|Z\right)\right] \\ &= \frac{1}{m^2 k}\sum\left[\sigma^2 - E\left[C\left(\boldsymbol{X_i}, \boldsymbol{Y_i}\right)\right]\right] \\ &= \frac{1}{n}\left[\sigma^2 - \overline{E\left[C\left(\boldsymbol{X_i}, \boldsymbol{Y_i}\right)\right]}\right]. \end{aligned} \tag{4.4}$$

$\boldsymbol{X_i}$ and $\boldsymbol{Y_i}$ are random points in the stratum i. The idea behind stratification is that the strata are chosen such that $E\left[C\left(\boldsymbol{X_i}, \boldsymbol{Y_i}\right)\right]$ is as large as possible so that (4.4) is less than (4.3). That is to say that the sampling error variance under a stratified sampling plan is less than the error variance under a uniform random sampling plan.

For stratified sampling with k points per strata where $k \geqslant 2$, an unbiased estimator for the sampling error variance would be the average within stratum variance [60]. However, the sampling error variance is minimised when small, compact strata are chosen ($k = 1$ is ideal); there is, therefore, no information that can be used to estimate the sampling error variance.

A proposed solution to random-grid or uniform random sampling is using geographically dense sub-areas of the region being sampled as strata [73]. The problem comes in when trying to determine how to split the region into these sub-areas such that they are as densely populated as possible. As a solution to this problem Walvoort, Brus and de Gruijter developed an `R` package for designing spatial coverage samples (space-filling designs) where the mean squared shortest distance (MSSD) was chosen as the objective function [73]. Brus et al. [12] determined that minimising the MSSD results in spatial coverage samples that are sufficient. A space-filling or coverage design is a spatial sampling plan which optimises a distance-based objective function [61].

In order to minimise the MSSD (equivalent to minimising the trace of the pooled within-cluster variance) k-means cluster analysis is utilised. The easiest way to briefly describe k-means clustering is through an algorithm [51]:

1. Place K points into the region being clustered. These are the initial cluster centroids
2. The distances between all the objects in the region and the centroids are then calculated and *groups* are formed by assigning the objects to their nearest centroid
3. Once all the objects have been assigned to a *group* the centroids of the groups are recalculated.
4. Steps 2 and 3 are repeated until the centroids converge

This algorithm is the basic idea behind the `spcosa` package.

For a more in-depth explanation of k-means the reader is referred to *The Elements of Statistical Learning* [32, page 460].

For the purpose of this package, algorithms for equal as well as for unequal partitioning of an area are outlined [73].

Unequal area partitioning [73]:

Step 1 If k sample points are required, then select k cells randomly from the grid. The midpoints of these cells are the temporary centroids of the k *clusters* (clusters in this context is not that sample as a cluster in cluster sampling). The unselected cells are then allocated to a partition based on the nearest centroid. Once all of the cells have been allocated to a partitioning, the *cluster* centroids are recalculated so that there are k two-dimensional vectors representing the k centroids: $\bar{\boldsymbol{x}}_{\boldsymbol{1}} \ldots \bar{\boldsymbol{x}}_{\boldsymbol{k}}$.

Step 2 It now needs to be determined whether or not cells need to be reallocated to different *clusters* based on the new set of centroids. Suppose cell $\boldsymbol{u_1}$, belonging to *cluster* is the first cell being considered for reallocation, the distance from cell u_1 to the k *cluster* centroids is calculated; $d^2(\boldsymbol{u_1}, \bar{\boldsymbol{x}}_{\boldsymbol{1}}), d^2(\boldsymbol{u_1}, \bar{\boldsymbol{x}}_{\boldsymbol{2}}), \ldots d^2(\boldsymbol{u_1}, \bar{\boldsymbol{x}}_{\boldsymbol{k}})$. If $d^2(\boldsymbol{u_1}, \bar{\boldsymbol{x}}_{\boldsymbol{1}}) > d^2(\boldsymbol{u_1}, \bar{\boldsymbol{x}}_{\boldsymbol{i}})$ for any $i \neq 1$ then the cell is reallocated to *cluster* i and the centroids for *cluster* 1 and i are recalculated.

Step 3 Step 2 is repeated, beginning at the first cell again ($\boldsymbol{u_1}$), until no cells are reallocated. This final partition and the resulting *cluster* centroids is the final solution.

Equal area partitioning [73] :

Step 1 An initial partition of k *clusters* with size N is made by running through all of the cells (randomly) and assigning cell i to *cluster* $mod(i, k)$. For example, if cell 12 is being considered and $k = 10$, then $\frac{12}{10} = 10$ remainder 2, so cell 12 will be allocated to *cluster* 2. Once all the cells have been allocated, the initial centroids are calculated and are denoted the same way as in Step 1 of unequal area partitioning.

Step 2 It is now determined if the first cell, with coordinate vector $\boldsymbol{u}$, of the first *cluster* (1) should be swapped with the first cell, $\boldsymbol{v}$, of another *cluster* (2) by calculating the distances between the cells and the *cluster* centroids. $d^2(\bar{\boldsymbol{x}}_{\boldsymbol{1}}, \boldsymbol{u}), d^2(\bar{\boldsymbol{x}}_{\boldsymbol{2}}, \boldsymbol{u}), d^1(\bar{\boldsymbol{x}}_{\boldsymbol{1}}, \boldsymbol{v}), d^2(\bar{\boldsymbol{x}}_{\boldsymbol{2}}, \boldsymbol{v})$. If $d^2(\bar{\boldsymbol{x}}_{\boldsymbol{1}}, \boldsymbol{u}) + d^2(\bar{\boldsymbol{x}}_{\boldsymbol{2}}, \boldsymbol{v}) > d^2(\bar{\boldsymbol{x}}_{\boldsymbol{1}}, \boldsymbol{v}) + d^2(\bar{\boldsymbol{x}}_{\boldsymbol{2}}, \boldsymbol{u})$, then the swap is carried out and the centroids are recalculated.

Step 3 Step 2 is repeated for all cells until no swaps occur. This final partitioning is the solution.

The usual method of k-means clustering updates the centroids at the end of each cycle, however, this often results in empty clusters. Therefore the methods mentioned here updated the centroids immediately after a swop or transfer took place. The k-means algorithm is deterministic, which means that the choice of the initial centroids influences the final clustering solution, therefore, multiple initial centroid processes could be tried until the most desirable clustering is found. [73]

4.4.1 Application

A spatially stratified sample in the `spsample` function is generated by using a grid which divides the spatial polygon into equal partitions or strata and then drawing a random coordinate from within each block or cell of the grid. This is similar to traditional stratified sampling, however, the strata are now the blocks of the grid as opposed to the stopping points along the road network. As was done with uniform random sampling, the distance between the `R` generated point and the location of the houses in the village is measured and the houses closest to the generated points then become part of the spatial

stratified sample. Figure 4.7 illustrates the houses that are sampled from the village according to the spatial stratified technique. Once again, the sampled houses are red, the unsampled houses are black and the stopping points are represented by blue triangles.

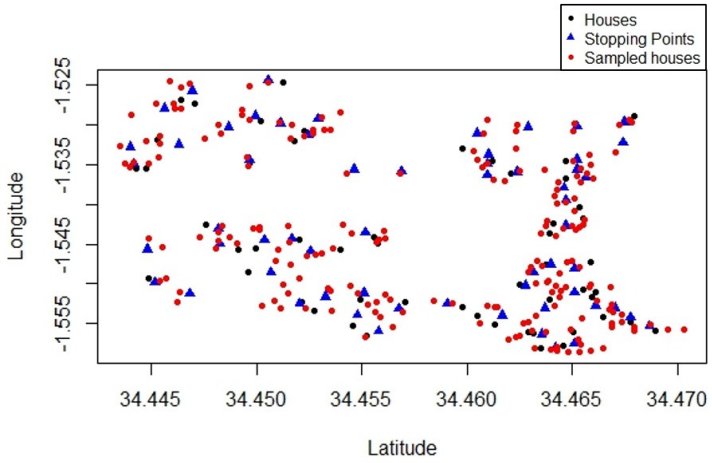


Figure 4.7: Plot of the houses in the Buchanchari village, sampled (red) and unsampled (black) according to spatial stratified sampling procedures in R

1000 spatial stratified samples are generated in R so as to obtain the distribution of the cost of the sampling strategy where cost is measured as the total walking distance between the houses in the sample. Figure 4.8 depicts the distribution of the cost function and Table 4.3 contains the summary statistics of the distribution. The average walking distance for this sampling strategy is 20.67km and the average coverage of the animals in the village is 70.8%

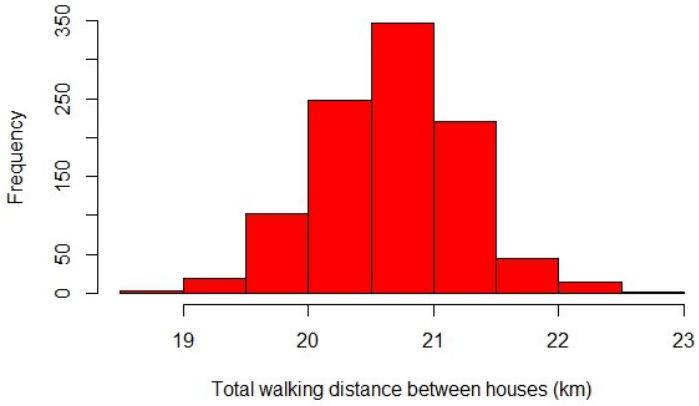


Figure 4.8: Distribution of cost function of 1000 spatial stratified samples

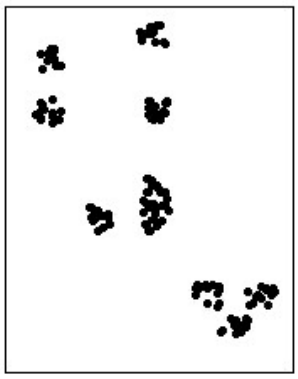
Summary statistics km	
Mean	20.67
Standard deviation	0.58
Minimum	18.82
Q_1	20.31
Median	20.69
Q_3	21.04
Maximum	22.79

Table 4.3: Summary statistics of the cost distribution of spatial stratified samples

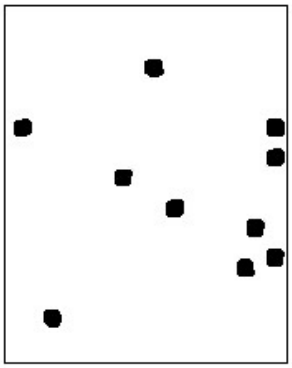
4.5 Cluster sampling

Cluster sampling in the spatial context is much like the traditional cluster sampling. Areas are randomly selected such that the variation within areas is small [21, 24]. Entire clusters are drawn where the probability of a cluster being drawn is equal for all the clusters [37]. Cluster sampling is particularly effective with discrete cases where an entire list of the population is unattainable but a complete list of the clusters is [24].

Within `spsample`, there is the option to do cluster sampling, however, this proved to be ineffective. The function randomly selects a block or cell of the grid as clusters and then takes a simple random sample of locations inside the cell. The number of clusters, as well as the sample size within each cluster, may be specified. Figure 4.9 shows examples of what the cluster option produces in `R` and highlights the impracticality of this approach for the Buchanchari village, particularly due to the reasoning behind sampling; obtaining a minimum coverage. Figure 4.9a is a sample of 10 clusters with 150 observations in each cluster and Figure 4.9b is a sample of 10 clusters with 1000 samples in each cluster. Through 1000 iterations of this kind of sampling, the result was that none of the cluster samples generated by the function met the minimum requirements to ensure 70% coverage of the village. This is therefore not an effective procedure for sampling houses in the village of Buchanchari. However, systematic cluster sampling, which is to be discussed next, results in sufficient samples.



(a) Spatial cluster sample with 10 clusters and 150 observations in each cluster



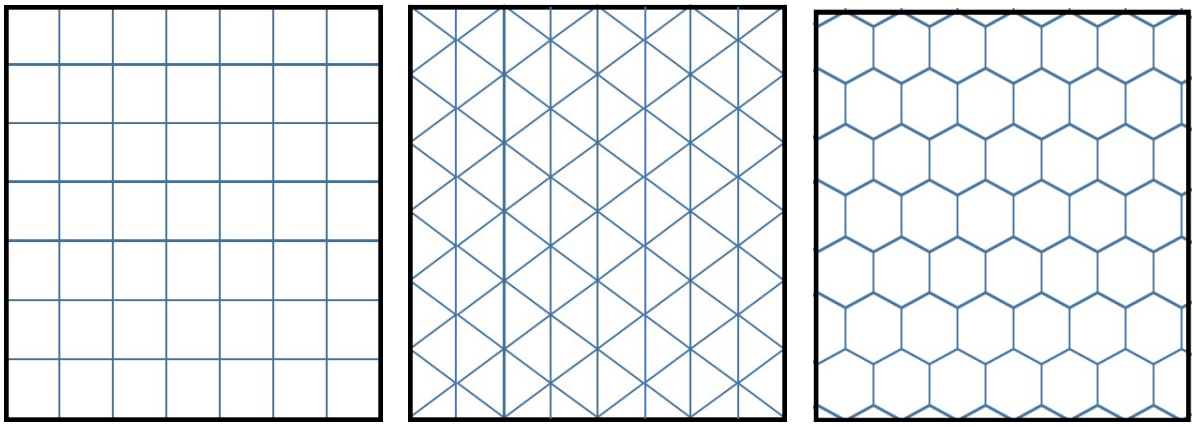
(b) Spatial cluster sample with 10 clusters and 1000 observations in each cluster

Figure 4.9: The optimal walk between the houses

4.5.1 Systematic sampling

Once again, systematic sampling is a special case of cluster sampling where only one observation is drawn from each cluster That is to say that the area of interest is divided according to a grid and only one observation is drawn from within each cell of the grid as opposed to drawing multiple samples in a cell. This is the more commonly chosen design of cluster sampling for spatial sampling due to superiority in comparison to uniform random and stratified sampling [27, 60].

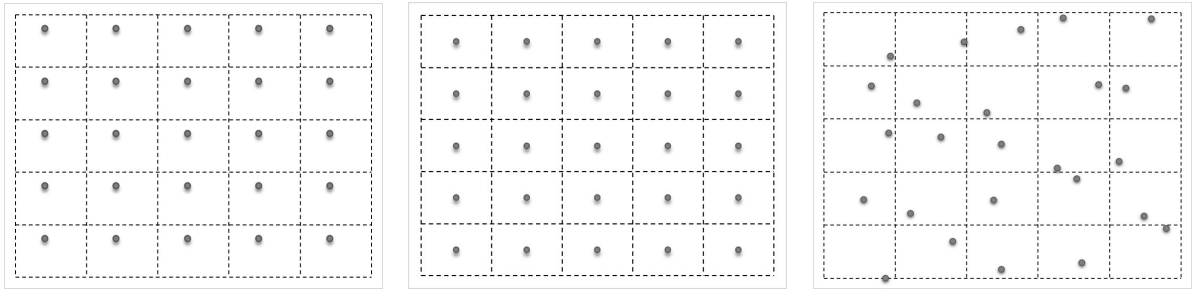
The population or area of interest is divided into n subgroups of equal size. Figure 4.10 shows the common shapes used to divide a domain space into subgroups . Figure 4.10a shows a region divided by means of squares (also known as quadrats [27]). For simplicity, this type of division will be the focus for the remainder of the section. Figure 4.10b shows division using equilateral triangles and Figure 4.10c is a hexagonal division [21]. The regions used in these images are of equal size, therefore, it is clear that using the equilateral triangle or the hexagon grid will require a higher number of sampling sites to cover the entire region in comparison to the square grid [21]. Specifically, 1.15 and 0.77 times more sampling units respectively [21, 77].



(a) Square divisions (b) Equilateral triangle divisions (c) Hexagonal divisions

Figure 4.10: Common systematic sampling grids

Suppose R can be approximated by a finite set of N sites and is divided into N nonoverlapping quadrats of equal area [27]. The first element of the sample is chosen from each subgroup either randomly or with specific intention [24]. There are three types of systematic spatial sampling techniques which are based on the selection of the first element. If the first element is chosen at random, it results in a **random** systematic sample where the remaining $n-1$ elements are allocated by some regular interval $\Delta = \frac{L}{\sqrt{n}}$ [21, 24]. This type of sampling can be seen in Figure 4.11a. If the first element is not chosen at random, then it is a **regular** systematic sample and, if the first element is chosen at the center of the first interval then the resulting sampling scheme is **centric** systematic sampling [24], Figure 4.11b. Figure 4.11c shows unaligned systematic sampling, which is where the elements within each cell are chosen using uniform random sampling. There are many different ways to perform unaligned systematic sampling [60, 24, 76], however, taking a simple random sample in each of the cells is the simplest method.



(a) Random systematic (b) Centric systematic (c) Unaligned systematic

Figure 4.11: Types of systematic sampling grids

The error variance for systematic samples is derived as follows:

Since,

$$E\left[\bar{Z}\right] = \frac{1}{n}\sum_{\boldsymbol{u}} E\left[Z\left(\boldsymbol{u}\right)\right],$$

$$\begin{aligned}
\operatorname{var}\left(\bar{Z} - \tilde{Z}\left(R\right)\right) &= \frac{1}{n^2} E\left[\sum_{\boldsymbol{u}}\left\{Z\left(\boldsymbol{u}\right) - \tilde{Z}\left(R\right)\right\}\right]^2 \\
&= \frac{1}{n^2}\sum_{\boldsymbol{u},\boldsymbol{v}} C\left(\boldsymbol{u},\boldsymbol{v}\right) - \frac{2}{n}\sum_{\boldsymbol{u}} E\left[\left(Z\left(\boldsymbol{u}\right) - \mu\right)\left(\tilde{Z}\left(R\right) - \mu\right)\right] + E\left[\tilde{Z}\left(R\right) - \mu\right]^2 \\
&= \frac{1}{n^2}\sum_{\boldsymbol{u},\boldsymbol{v}} C\left(\boldsymbol{u},\boldsymbol{v}\right) - 2\sum_{\boldsymbol{u}} \frac{1}{An}\int_R C\left(\boldsymbol{u},\boldsymbol{v}\right)\,\mathrm{d}\boldsymbol{y} + \frac{1}{A^2}\int_R\int_R C\left(\boldsymbol{x},\boldsymbol{y}\right)\,\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}. \qquad (4.5)
\end{aligned}$$

In order to get even coverage of a region, systematic point sampling is often used [26]. However, there is the risk that the interval between points may correspond to some recurrence in the data such as landforms, man-made landscapes or spacing of crops in a field.

A method to estimate the sampling error variance for a systematic random sample is to regard $\frac{s^2}{n}$ as the error variance or to average s^2 over the artificial strata and dividing by n, effectively using the stratified sampling formula [60]. Milne [54] published an article where the effect of treating a centric systematic sample as if it were random was investigated. He concluded that, while random sampling cannot be ignored, "*one will not go very far wrong, if wrong at all, in treating centric systematic area-sampling as if it were random*".

4.5.2 Application

The aforementioned function `spsample` in the `R` package `spatstat` is used to generate a systematic random pattern of points within the Buchanchari village. A grid is *placed* over the region and a single observation is chosen from each window of the grid. This point may be chosen regularly i.e. equal spacing between points in adjacent blocks or nonaligned which is when the first point is chosen at random in the first grid and thereafter an algorithm specifies where in the adjacent block the next point will be sampled. Another variation on the systematic sampling in `spsample` is that a hexagonal grid may be used. The points are then generated regularly but their alignment differs slightly from the `regular` argument as the grid is hexagonal as in Figure 4.10c.

The following samples were drawn using `R`:

- `regular`

  Regular spatial sampling, also known as centric systematic sampling, places a grid over the region of interest and generates regular, continuous points from the center of the cells of the grid.

  Figure 4.12 illustrates a spatial regular sample obtained in `spsample` where the points are systematically aligned in the centre of the cells. The black points are the houses in the Buchanchari village which are not included in the sample and the red points are the sampled houses. The blue triangles are the stopping points along the road network.

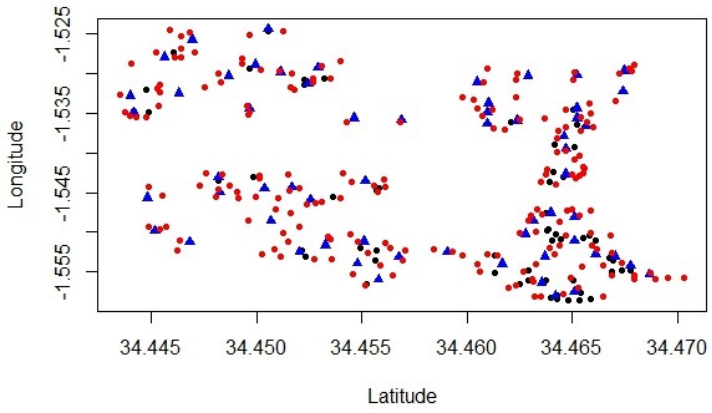


Figure 4.12: Plot of the houses in the village of Buchanchari (black), stopping points along the road network (blue) and the sample houses (red) generated in `R`

The spatial regular samples are bootstrapped in `R` to obtain a distribution of the cost of the sampling procedure. Figure 4.13 illustrates this distribution, which has an average optimal walking distance of $20.81km$ and an average animal coverage of 70.64%.

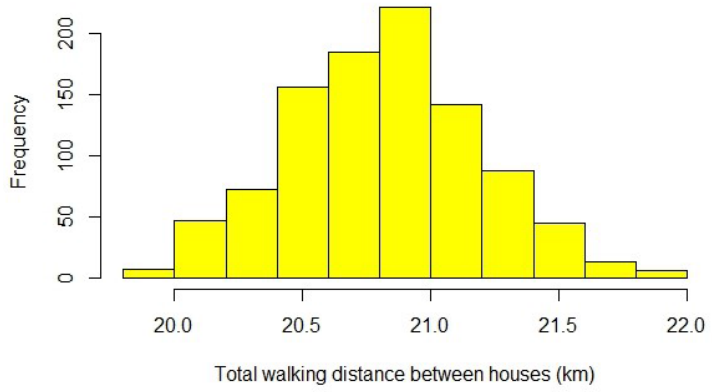


Figure 4.13: Distribution of cost function of 1000 spatial regular samples

Summary statistics km	
Mean	20.81
Standard deviation	0.37
Minimum	19.85
Q_1	20.55
Median	20.82
Q_3	21.04
Maximum	21.91

Table 4.4: Summary statistics of the cost distribution of spatial regular samples

- `nonaligned`

  For the nonaligned samples, `R` generates n random y coordinates and n random x coordinates which then form locations within each cell of the grid. The nonaligned systematic samples generated with `spsample` are plotted in Table 4.14.

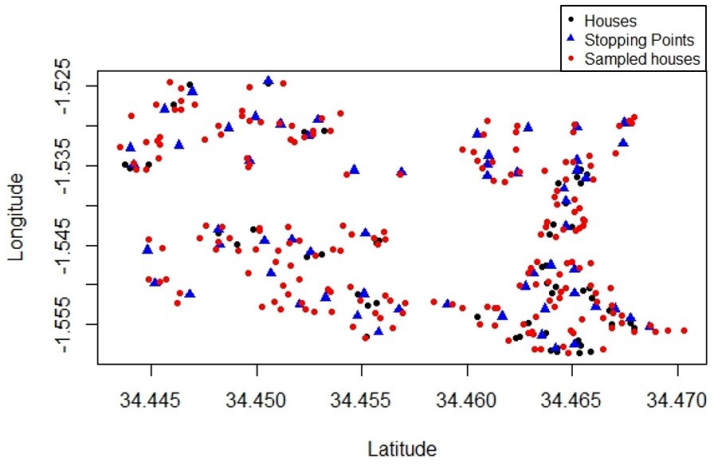


Figure 4.14: Plot of the houses (black), stopping points (blue) and sampled houses (red) according to nonaligned systematic spatial sampling

Once again this is bootstrapped 1000 times so as to obtain a distribution for the cost of the sampling strategy. The average walking distance is $20.91km$ and the average coverage is 70.75% of the animals in the village. The summary statistics are presented in Table 4.5.

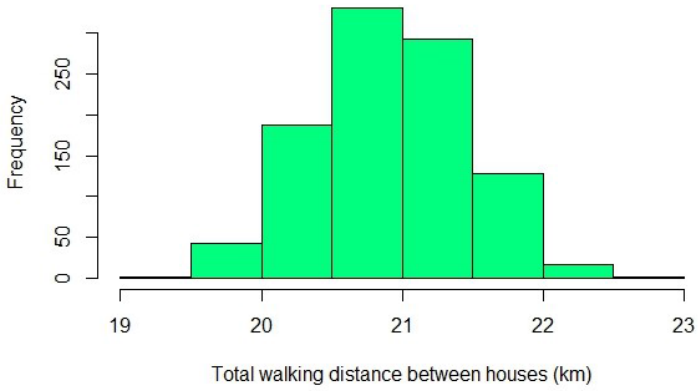


Figure 4.15: Distribution of cost function of 1000 spatial nonaligned samples

Summary statistics km	
Mean	20.91
Standard deviation	0.54
Minimum	19.25
Q_1	20.55
Median	20.89
Q_3	21.28
Maximum	22.57

Table 4.5: Summary statistics of the cost distribution of spatial nonaligned samples

- `hexagonal`

  The hexagonal samples in `spsample` are generated by sampling a single point from within each cell of the hexagonal grid like the one presented in Figure 4.10c. A plot of a systematic sample using a hexagonal grid can be seen in Figure 4.16.

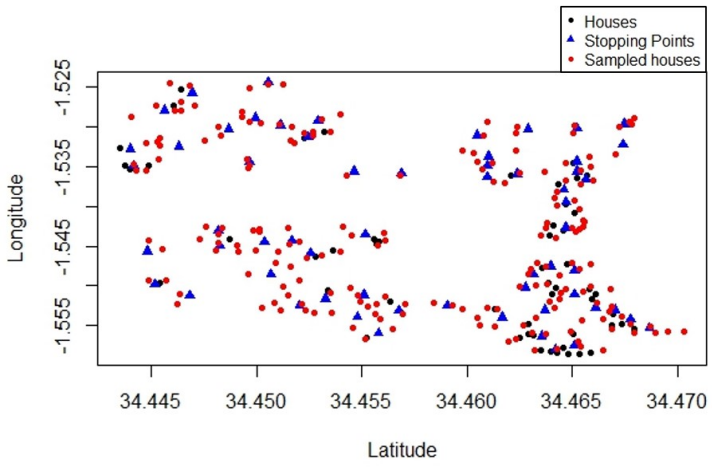


Figure 4.16: Plot of the houses (black), stopping points (blue) and sampled houses (red) according to hexagonal systematic spatial sampling

The hexagonal sample is bootstrapped 1000 times to obtain a distribution for the cost of the sampling strategy. Figure 4.17 shows the distribution of this cost with an average walking distance of 20.73km and an average animal coverage of 70.68%. The summary statistics for this distribution can be seen in Table 4.6

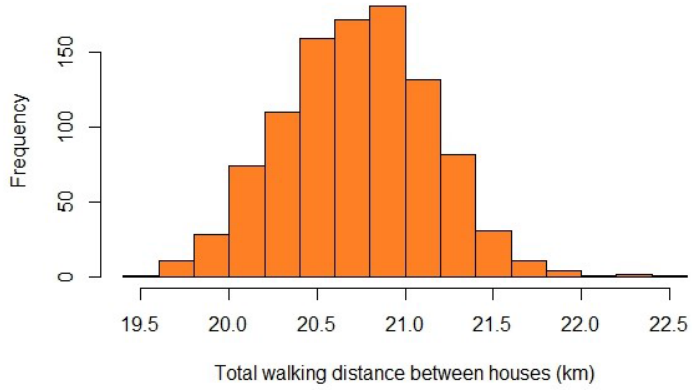


Figure 4.17: Distribution of cost function of 1000 spatial hexagonal samples

Summary statistics km	
Mean	20.73
Standard deviation	0.43
Minimum	19.47
Q_1	20.44
Median	20.73
Q_3	21.01
Maximum	22.48

Table 4.6: Summary statistics of the cost distribution of spatial hexagonal samples

4.6 Other spatial sampling strategies

Presented here are alternative options for obtaining spatial samples. Line intersect sampling is effective in analysing an area containing various objects of interest and fixed- and variable-radius plot sampling are area sampling strategies which are often used in estimation with forest inventories. These strategies were not applied to the Buchanchari village due to the nature of this mini-dissertation and its aim to provide a parallel comparison between traditional samples and spatial samples which share similar underlying theory.

4.6.1 Line intersect sampling

Line intersect sampling was developed based on the idea of *Buffon's Needle Problem.* An eighteenth-century French naturalist, George Louis Leclerc, who went more commonly by the title *Comte de Buffon* [36, 72] posed a question which is considered to be the first problem associated with geometrical probability [71, page 251]. Suppose there is a surface marked with parallel lines which are separated by a distance d, as is depicted in Figure 4.18 and a needle of length $l < d$ is dropped randomly on the surface. In 1733 Buffon wished to determine the probability of this needle crossing a line on the surface [36]. The solution was determined to be $p = \frac{2l}{\Pi d}$.

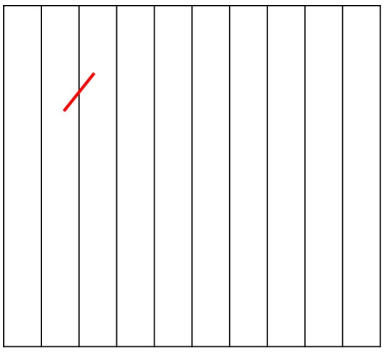


Figure 4.18: The Buffon Needle Problem

Now suppose the needle is thrown m times,

$$t_i = \begin{cases} 1 & \text{if the } i^{th} \text{ needle thrown intersects a line} \\ 0 & \text{otherwise.} \end{cases}$$

From this equation it is clear that t_i follows a Bernoulli distribution with probability of success (i.e. intersection with a line) equal to p. If m_0 represents the number of intersections ($m_0 \leq m$), then $m_0 \sim$ BIN (m, p). The reader is referred to *Sharpening Buffon's Needle* [58] for further reading regarding Buffon's Needle Problem.

Suppose now that there exists a planar region R with an area equal to A and there are N fixed particles in the region denoted by $P_1, P_2, \ldots P_N$ which are to be sampled (such as roads, hedges, forest canopies [72]) [42]. Also let θ represent some direction in the plane where $\theta = 0$ is the base direction.

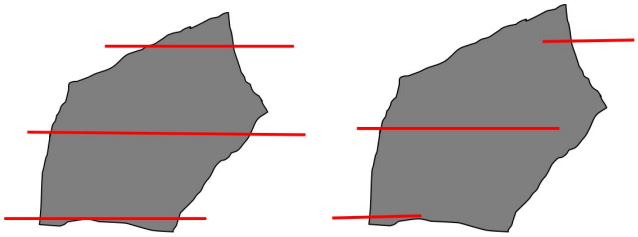
Samples are obtained as follows [42]:

1. Choose a random point uniformly in R
2. This chosen point is the midpoint of a transect with length L

3. The direction of the transect is $\theta \sim \text{UNI}(0, \pi)$
4. The midpoint and θ are independent
5. The particle P_i is sampled if the transect intersects the particle completely
6. If a particle is partially intercepted, a side of the transect is chosen at random and only particles intersecting with the chosen side are sampled [50] (see Figure 4.19).

Timothy Gregoire [35] noted that the orientation of the transect (fixed or random) is a non-issue in terms of countering the lack of randomness of the population particles. The shape of these particles need not be convex (summation of interior angles equal to 180°) but are assumed to be a set of connected points [42]. Complete and partial particle intersection with transects are pictured in Figure 4.19.

It is up to the researcher if the sampling will be done with or without replacement, depending on the reason for the sampling. For example, Battles et al. [5] sampled canopy gaps in a forest of trees using line intersect sampling. The study required that the sampling is done without replacement so the researchers placed transects a minimum of 50 meters apart in order to avoid sampling any gap twice.



(a) Complete Intersection (b) Partial Intersection

Figure 4.19: Example of complete and partial intersection of a transect with a non-convex particle

Consider the particle P_i in Figure 4.20. If the transect has a direction equal to θ then the particle will be intercepted completely if the midpoint of the transect is within the shaded region which is bounded by broken lines. The area of the region is $Lw_i(\theta) - a_i$ with a_i the area os particle P_i. If the transects midpoints is in either of the shaded regions then P_i is sampled with probability 0.5 [42]. Lee Kaiser [42] therefore revealed that the conditional probability of the j^{th}, P_j particle being included in the sample is

$$\begin{aligned} P(t_j = 1|\theta) &= \frac{Lw_j(\theta) - a_j}{A} + \frac{1}{2}\left(\frac{2a_j}{A}\right) \\ &= \frac{w_j(\theta) L}{A}, \end{aligned} \tag{4.6}$$

where A is the area of the region R and $w_j(\theta)$ is the maximum perpendicular distance between parallel tangents to P_j [42, page 967].

The unconditional probability of particle P_j being included in the sample is

$$P\left(t_j = 1\right) = E_\theta\left[w_j\left(\theta\right)\right]\frac{L}{A}, \tag{4.7}$$

where $E\left[w_j\left(\theta\right)\right] = \frac{1}{\pi}\int_0^\pi w_j\left(\theta\right)\mathrm{d}\theta$. The probability of P_i being sampled is proportional to the size of P_i. [42]

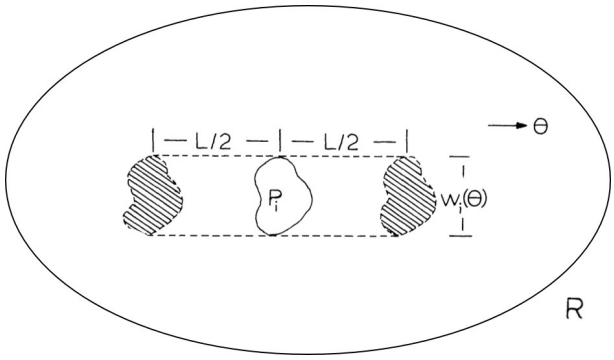


Figure 4.20: An illustration of the formula $P\left(t_i = 1|\theta\right) = \frac{Lw_i(\theta)}{A}$, with $t_k = 1$ $U|left(0, w_k\left(\theta\right)$ [42, page 967]

For particles which are less than a distance of $0.5L$ of the boundary of R, equation (4.6) will not be true for all values of θ. In order to make equation (4.6) usable, provided that R is convex, is to *bring* the portion of the transect that lies outside of R *into* R by extending the transect a perpendicular distance d away from the portion of the transect already in R. This distance must be greater than the maximum value of $w_i\left(\theta\right)$ so that no transect intercepts the same particle twice [42]. Figure 4.21 illustrates an example where the portion of a transect lying outside of R is brought back on the right-hand side of the portion in R. The side(left or right of the transect) that the portion is brought back in on does not matter. Equation (4.6) now holds for particle P_j.

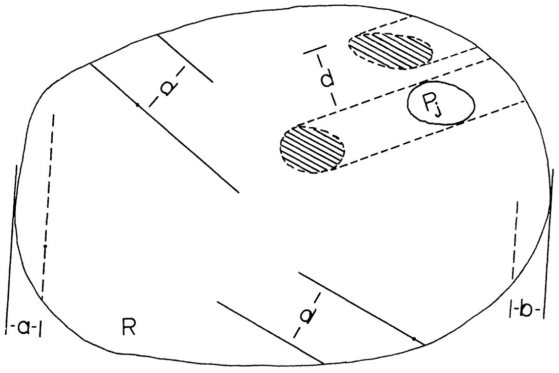


Figure 4.21: An illustration of a scheme for bringing into R any portion of the transect which lies outside of R. Note that for the dashed transect, the distances a and b between the transect and the tangent lines to R are such that $a + b = d$ [42, page 968]

An important feature of this sampling scheme is that, given the direction of the transect θ, and that $t_k = 1$ (i.e. particle k is sampled), the perpendicular distance U between the left-hand tangent line to P_k and the transect is such that $U \sim \text{UNI}(0, w_k(\theta))$.

In order to show this, consider the following definition of conditional probability,

$$P(U \in (u, u+du) | \theta, t_k = 1) = \frac{P(U \in (u, u+du), t_k = 1 | \theta)}{P(t_k = 1 | \theta)}. \tag{4.8}$$

Let z be the length of the intercept of P_k in Figure 4.22 with a transect $U = u$. Given θ, U will be in $(u, u+du)$ and P_k will be sampled with probability 0.5 if the midpoint is in either of the shaded strips, which have total area $2zdu$. Alternatively, if the midpoint is in the unshaded strip which has an area equal to $(L-z)\,du$, then the particle Pk will be sampled with probability 1. [42, page 968]

Therefore,

$$P(U \in (u, u+du), t_k = 1 | \theta) = \frac{(L-z)\,du + 0.5\,(2zdu)}{A} \tag{4.9}$$

$$\frac{Ldu}{A}. \tag{4.10}$$

By combining equations (4.6),(4.8) and (4.10), it is shown that

$$P(U \in (u, u+du) | \theta, t_k = 1) = \frac{du}{w_k(\theta)},$$

which means $U|theta, t_k = 1 \sim \text{UNI}(0, w_k(\theta))$. An identical argument is true for particles *near* the boundary of R.

Two unbiased estimators for λ_x are,

$$\hat{\lambda}_1 = \frac{1}{A} \sum \frac{t_i y_i}{E(t_i, y_i)} x_i,$$

and

$$\hat{\lambda}_2 = \frac{1}{A} \sum \frac{t_i y_i}{E(t_i, y_i | \theta)} x_i.$$

If the distribution of θ is degenerate, that is to say that probability distribution of θ only takes a single value, then $\hat{\lambda}_1 = \hat{\lambda}_2$.

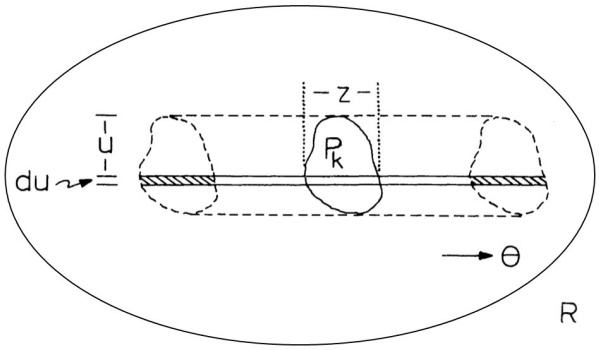


Figure 4.22: An illustration of $P(U \in (u, u+du) | \theta, t_k = 1) = \frac{du}{w_k(\theta)}$ [42, page 967]

4.6.2 Fixed- and variable-radius plot sampling

Fixed- and variable-radius plot sampling are commonly used in forest inventories. This is the collection of data regarding forests for analysis which could, for example, be used to sustain ecosystems [55].

Consider M independent sampling locations within a forested region R. Each location represents the centre of the circle which has either a fixed or variable radius depending on the sampling requirements [35]. It is important to note that all M locations must either all be fixed-radius or all be variable-radius plots, there cannot be a mixture within one sampling scheme [35, page 1441].

Suppose there are N trees within R denoted by $(\tau_1, \tau_2, \ldots, \tau_N)$, each with a biomass denoted by $(y_1, y_2, \ldots, y_N)$ and diameters $(d_1, d_2, \ldots, d_N)$. The parameter of interest for the trees is $T_y = \sum_{k \in U} y_k$ and U represents the population of N discrete elements. [35]

Let Z_m denote the location of the m^{th} circle centre and t_{km} be the distance between τ_k and Z_m. (Note that since Z_m is random, t_{km} will also be random.) ν_k is the limiting distance defined as:

$$\nu_k = \begin{cases} r, & \text{when using fixed-variable plots} \\ r_k, & \text{when using variable-radius plots,} \end{cases}$$

where r is the plot radius and $r_k = \phi d_k$, with ϕ the plot radius factor.

At each of the M locations, the trees will be selected into the sample using the following distance rule:

$$\tau_k \textit{ will be included in the sample if } t_{km} \leq \nu_k.$$

Therefore, if I_{km} indicates whether or not τ_k is included in the sample, then

$$I_{km} = \begin{cases} 1 & \text{if } t_{km} \leq \nu_k \\ 0 & \text{otherwise,} \end{cases}$$

so that $p_k = P(I_{km} = 1)$.

The probability p_k of the k^{th} tree being included in the sample can also be interpreted as the proportion of the region R within which a plot can be located such that $t_{km} \leq \nu_k$. p_k is a constant since it is a property of τ_k fixed by the sampling design and not the sampling location Z_m. If τ_k is closer to the border of R than the limiting distance ν_k, then p_k will be less than if τ_k was further from the edge than ν_k. [35] Plots that are located near the edge have no effect on p_k, the reader is referred to an article by Timothy Gregoire [34] for further explanation of this phenomena, known as 'boundary overlap'.

The Horvitz-Thompson estimator of T_y from the sample at Z_m is,

$$\hat{T}_{y,m} = \sum_{k \in U} \frac{y_k I_{km}}{p_k}.$$

Consider,

$$\begin{aligned} E\left[\hat{T}_{y,m}\right] &= \sum_{k\in U} \frac{y_k}{p_k} E\left[I_{km}\right] \\ &= \sum_{k\in U} \frac{y_k}{p_k} p_k \\ &= \sum_{k\in U} y_k \\ &= T_y. \end{aligned}$$

This proves that $\hat{T}_{y,m}$ is an unbiased estimator of T_y [35].

The variance of the estimator is

$$\begin{aligned} \text{var}\left(\hat{T}_{y,m}\right) &= \sum_{i\in U}\sum_{j\in U} \frac{y_i y_j}{p_i p_j} Cov\left(I_i, I_j\right) \\ &= \sum_{i\in U}\sum_{j\in U} \frac{y_i y_j}{p_i p_j} \left(p_{ij} - p_i p_j\right), \end{aligned}$$

where $p_{ij} = E\left[I_{im} I_{jm}\right]$ (the probability of both trees i and j being included in the sample.

When the sample size is fixed, the variance of the estimator can be expressed as

$$\text{var}\left(\hat{T}_{y,m}\right) = \sum_{k\in U} y_k^2 \left(\frac{1-p_k}{p_k}\right) + \sum \sum_{i\neq j\in U} y_i y_j \left(\frac{p_{ij} - p_i p_j}{p_i p_j}\right).$$

4.7 Summary

Uniform (simple), stratified and systematic spatial samples of houses in the Buchanchari village were drawn using functions within `R` software. The functions allow for more even coverage of a specified region by laying a grid over the area and sampling from within the cells. The grid can have square, hexagonal, or even triangular divisions. As had been done in traditional sampling, the cost of each sampling procedure was calculated as the total optimal walking distance between houses surrounding stopping points. All of the applicable sampling strategies were bootstrapped 1000 times so as to obtain a distribution of the cost as well as average coverage of the animal population in the village.

From the results of this chapter, it is evident that both stratified and systematic spatial sampling yield better outcomes than random spatial sampling. Stratified spatial sampling resulted in a slightly lower cost and a slightly higher coverage than the systematic spatial samples. These two spatial sampling approaches also produced improved results when compared to traditional spatial samples, thereby confirming that taking spatial characteristics into consideration when sampling is beneficial.

Chapter 5

Discussion

The applications that were performed in this mini-dissertation will be discussed and critically compared here. The distributions of the cost functions, as well as the summary statistics, are set side by side for easier interpretation and the challenges of this work as well as possible shortcomings are deliberated.

The dataset considered here is an extensive census with information regarding the village in Tanzania, the location of the houses within the village and the number and type of animals at each of the houses (cat or dog, younger or older than 3 months, vaccinated or not vaccinated etc.). From this comprehensive dataset, the validity of each of the applied sampling techniques could be verified, in that the true number of animals at the sampled houses is known. In order to achieve herd immunity, at least 70% of the animals in an area need to be vaccinated against rabies and it was determined that consideration of the spatial component of the data is useful in achieving this minimum coverage. The sampling techniques presented, not only ensure sufficient coverage of the area of interest but also ensure that the walking distance of the vaccinator is minimised.

Figure 5.1 contains the distributions of the costs of all of the sampling techniques that were applied to the Buchanchari village of Tanzania. These cost functions are calculated as the total optimal walking distance, in kilometres, between sampled houses and all have a minimum animal coverage of 70%, as is suggested by the World Health Organisation [45, 78, 16, 13] as being the most profitable and cost-effective. Studies which compare the monetary cost of different vaccination studies have been performed [28] and could be combined with the work done here. However, the focus for these spatial samples was to minimise distance walked by the veterinarian rather than the money spent, thereby focusing on the time constraint of the approach only.

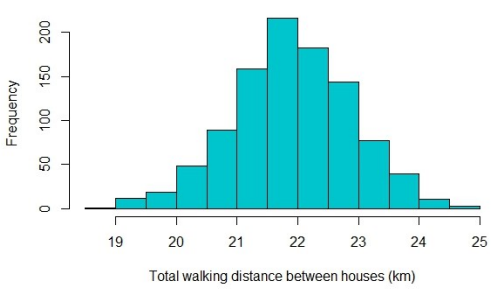
Each sampling strategy was bootstrapped 1000 times in order to obtain and plot the distributions. In general, the average cost for the spatial sampling procedures was less than the cost of the traditional

samples. The sampling strategy which yielded the lowest cost after 1000 samples was the spatial stratified sampling strategy which had an average walking distance of $20.73km$. The spatial systematic samples had the next lowest total cost. For the hexagonal, regular and nonaligned approaches, the costs were $20.73km$, $20.81km$ and $20.91km$ respectively. These results support the original belief that the spatial approach to sampling is beneficial and appropriate. Figure 5.2 contains all the summary statistics of the costs of the sampling strategies that were used in this mini-dissertation. The standard deviations of each of the costs are all less than $1km$. This shows that the costs of the bootstrapped samples do not vary too drastically through each iteration. These minimum values are highlighted in red, emphasising that the spatial stratified sampling techniques yields the most efficient results in terms of minimising the cost. The average of 1000 spatially stratified samples is $20.67km$ with a standard deviation of $0.58km$. The technique of spatial systematic regular(Table 5.2f), nonaligned (Table 5.2g) and hexagonal (Table 5.2h) sampling have lower variance with only a slightly larger mean cost, thus one could also argue that these techniques are optimal to use. The results presented in Table 5.2e support the argument that stratified spatial sampling is the optimal choice since it has the smallest mean, first quartile and median values.

It can be argued that spatial intelligence was used in the design of the traditional samples in that stopping points were used as the strata and clusters in the sampling procedures and graph theory was utilised to minimise the walking distance between houses. However, this is often done in traditional sampling, for example when sampling from South Africa, the country is often divided into strata based on the provinces, a spatial characteristic for traditional sampling. It is important to note that the spatial characteristics of the data were used in developing the cost of the sampling and not the actual sampling itself.

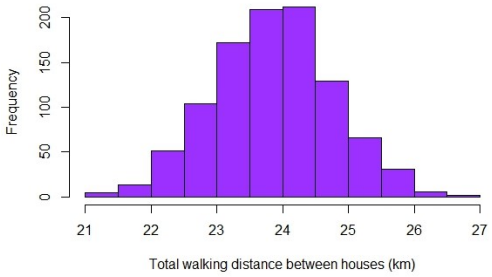
The function `spsample` used in `R` for the methods used in Chapter 4 generates continuous data points and the houses form a discrete dataset. The implementation selects the sampled houses as those nearest to the continuous data point selected by the `spsample` function. There exists a function within `R` called `discrete.sample` which draws discrete samples from a spatial data set with an imposed minimum distance between observations. While this is a useful function, it would not have served its purpose in this context as the distance between the houses should not affect whether or not they are included in the sample, but should rather influence the cost of the sampling strategy. That is to say, if the vaccinator is at a house and there is another house $10m$ away, it would not make logical sense for the animals at that house not to be vaccinated as well.

Another aspect which could be questioned is the way in which the `R` function accounts for the spatial component of the data. A simple grid may not be sufficient in dividing the region being sampled. However, the samples produced did yield meaningful results and were not too computationally advanced for implementation. The grid approach is valuable as it ensures equal dispersion across the area of interest.

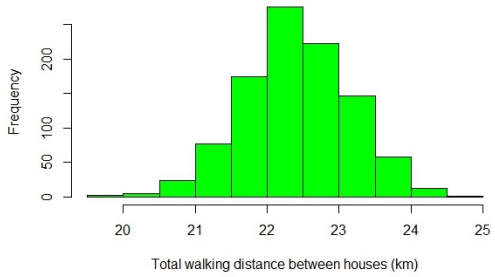


(a) Traditional - simple random samples

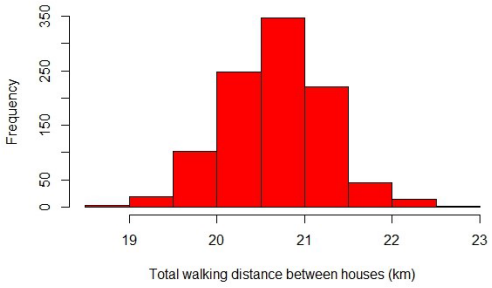
(b) Traditional - stratified samples



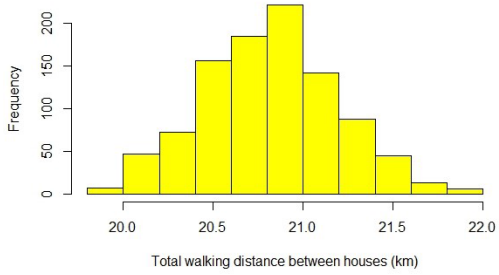
(c) Traditional - cluster samples



(d) Spatial - uniform random samples

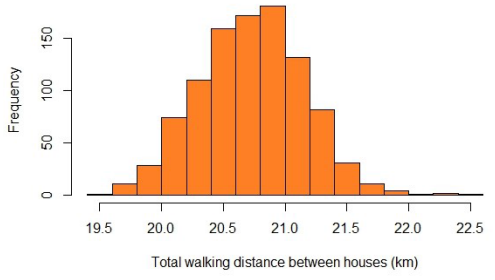


(e) Spatial - straitifed samples



(f) Spatial - systematic regular samples

(g) Spatial - systematic nonaligned samples



(h) Spatial - systematic hexagonal samples

Figure 5.1: Distributions of the costs of traditional and spatial sampling strategies

Summary Statistics km	
Mean	21.91
Standard deviation	0.98
Minimum	18.94
Q_1	21.3
Median	21.91
Q_3	22.56
Maximum	24.74

(a) Traditional - simple random sampling

Summary Statistics km	
Mean	21.34
Standard deviation	0.55
Minimum	18.89
Q_1	20.99
Median	21.35
Q_3	21.73
Maximum	22.84

(b) Traditional - stratified sampling

Summary Statistics km	
Mean	23.86
Standard deviation	0.9
Minimum	21.02
Q_1	23.26
Median	23.88
Q_3	24.44
Maximum	26.84

(c) Traditional - cluster sampling

Summary Statistics km	
Mean	22.41
Standard deviation	0.74
Minimum	19.93
Q_1	21.93
Median	22.39
Q_3	22.93
Maximum	24.69

(d) Spatial - uniform random sampling

Summary Statistics km	
Mean	20.67
Standard deviation	0.58
Minimum	18.82
Q_1	20.31
Median	20.69
Q_3	21.04
Maximum	22.79

(e) Spatial - stratified sampling

Summary Statistics km	
Mean	20.81
Standard deviation	0.37
Minimum	19.85
Q_1	20.55
Median	20.82
Q_3	21.04
Maximum	21.91

(f) Spatial - systematic regular sampling

Summary Statistics km	
Mean	20.91
Standard deviation	0.54
Minimum	19.25
Q_1	20.55
Median	20.89
Q_3	21.28
Maximum	22.57

(g) Spatial - systematic nonaligned sampling

Summary Statistics km	
Mean	20.73
Standard deviation	0.43
Minimum	19.47
Q_1	20.44
Median	20.73
Q_2	21.01
Maximum	22.48

(h) Spatial - systematic hexagonal sampling

Figure 5.2: Summary statistics of the cost distributions of traditional and spatial sampling techniques

Within the application of this work, driving distance between houses was deemed negligible since the distance the vaccinator drives is not as taxing as the distance walked to reach the animals. It is understood that the approach used currently does not require the veterinarian to drive between houses or even walk, but the addition of these factors allows for more efficient samples. A study [48] suggested that at any given time 59% of animals should be vaccinated in order to control rabies and only an annual coverage of 70%. Therefore, the strategy suggested in this mini-dissertation could be divided into trips over the whole year. By doing this, the veterinarian need not walk the full $20km$ in a day to vaccinate all the animals, but would rather do multiple trips, walking shorter distances and still achieving the desired population coverage.

The geography of the land, in terms of hills and slopes, was not considered in obtaining these samples. The distances between houses were measured as the crow flies and may, therefore, be more costly than is presented here. Future work could include considering these slopes and factoring them into the calculation of the cost of the sample. However, the Buchanchari village that was used in this application was low lying with no noticeable slopes.

5.1 Summary

In this discussion chapter, the results of the application of traditional, as well as spatial sampling, were presented and compared. Upon further investigation of the cost distributions of each of the sampling strategies which were applied, it was determined that a spatial stratified approach yields the most efficient samples in terms of average cost (walking distance) for a specified population coverage.

Chapter 6

Conclusion

This mini-dissertation discussed both traditional and spatial sampling strategies with specific application to vaccination of animals in a village in Tanzania. Currently, animals are vaccinated by placing a vaccination station in the centre of the village and waiting for villagers to bring their animals for treatment. This mini-dissertation proposes a more efficient sampling strategy which takes into consideration the spatial component of the data. Both traditional and spatial sampling techniques were applied to the data in order to make comparisons and justify the use of one approach over another.

The objectives of this mini-dissertation were outlined in Chapter 1. The theory of the traditional and spatial sampling techniques was covered with emphasis on design-based sampling so as to draw a clear relationship between the sampling strategies under traditional and spatial sampling theories. These approaches to sampling were applied to a census data set regarding a village of Tanzania, Buchanchari. Within this extensive data set is the location of the houses within the village, as well as the number of animals at each of the houses. It is desirable to access at least 70% of the animal population in a village in order to vaccinate them against rabies. This minimum percentage is believed to achieve herd immunity and decrease the transmission of rabies among the population of the village.

The state veterinarian currently vaccinates the animal population of a village in Tanzania by setting up a vaccination station in the middle of the village for a few days and hoping the villagers will bring their animals for vaccination. Obviously, this is inefficient and unreliable. This mini-dissertation proposed an improvement in this strategy by taking spatial samples in the villages which incorporated graph theory in obtaining optimal walking paths and vaccinating accordingly. A road network was developed around the houses in order to determine which houses are accessible and which are not based on their distance from the road. Stopping points were then located according to a kernel density estimation heat map which revealed areas with a large number of houses. These stopping points served as nodes in graphs which

were constructed around each point to determine how far the veterinarian would need to walk between the sampled houses to vaccinate the animals. Each sampling technique was bootstrapped 1000 times in order to obtain a cost distributions for each approach and make comparisons between traditional and spatial techniques. The results revealed that a stratified spatial sample resulted in the smallest walking distance between houses with the minimum 70% animal coverage, but the strategies of systematic regular, nonaligned and hexagonal spatial sampling did achieve slightly lower variation with a similar mean walking distance. The results of the work done here can be applied to other geographical fields which require reliable samples.

Future research should involve covariate data, such as terrain slope, in order to obtain improved spatial samples. The effectiveness of other spatial sampling functions within statistical software should also be looked at. Space-filling designs, which aim to fill a region with points as uniformly as possible [23], may also be utilised in an attempt to ensure even coverage of areas being sampled. The extension of this study to model-based designs should reveal interesting results.

Accounting for the spatial component of data proved to be advantageous in obtaining efficient samples of the village Buchanchari. The application of these sampling strategies is far-reaching as they can be used in disease control, as was done here, population control, environmental monitoring and environmental surveys. The procedures can also be applied to geographic observations where little is known regarding spatial properties and samples need to be obtained.

Appendix

This Appendix contains the R code which was employed for the application of the sampling techniques discussed in this mini-dissertation.

```
library(readr)
library(rgdal)
library(sampling)
library(igraph)
library(optrees)
library(spatstat)
library(sp)

houses <- read_csv("C:\\Users\\Hayley␣Reynolds\\Dropbox\\Masters\\Year␣2
\\Semester␣2\\WST␣895\\Dataset\\BuchanchariAccessible.csv")
stopping_points <- read_csv("C:\\Users\\Hayley␣Reynolds\\Dropbox\\Masters\\Year␣2
\\Semester␣2\\WST␣895\\Dataset\\SPacc.csv")
stratified <- read_csv("C:\\Users\\Hayley␣Reynolds\\Dropbox\\Masters\\Year␣2
\\Semester␣2\\WST␣895\\Dataset\\strata.csv")

##### Simple Random Sampling ####
fix(SRS)
SRS(houses,stopping_points,1)
##### Stratified Sampling ####
fix(Stratified)
Stratified(houses,stopping_points,stratified,1000)
##### Cluster Sampling ####
fix(Cluster)
Cluster(houses,stopping_points,1000)
#### Spatial random ####
```

```
fix(spat_rand)
spat_rand(houses,stopping_points,1000)
#### Spatial regular ####
fix(spat_reg)
spat_reg(houses,stopping_points,1000)
#### Spatial stratified ####
fix(spat_strat)
spat_strat(houses,stopping_points,1000)
#### Spatial nonaligned ####
fix(spat_non)
spat_non(houses,stopping_points,1000)
#### Spatial hexagonal ####
fix(hexagonal)
hexagonal(houses,stopping_points,1000)
#### Spatial clustered ####
fix(clustered)
clustered(houses,stopping_points,1000)

### Simple random sampling ###
function (houses,stopping_points,q)
{total_walk <- matrix(1, nrow = 60, ncol = q)
  aaa <- matrix(0,nrow = q,ncol = 1)
  for(j in 1:q){
    set.seed(j+60)
    x <- 1:247
     #obs for sampling
    s <- sample(x,247)
    sample <- houses[s,]
     #indexing sample
    CumAnimals <- cumsum(sample[,5])
    obs_num <- min(which(CumAnimals/701 > 0.7))
     #determine if 70% reached
    Act_sample <- sample[1:obs_num,]
     #Save the observations that allows the 70% coverage
    aaa[j] <- as.numeric(CumAnimals[obs_num,]/701*100)
```

```
    for(i in 1:60) {
      subbie <- subset(Act_sample,SP==i)
       #Working with individual stopping points
      SP <- stopping_points[i,]
      new <- rbind(SP[,2:3],subbie[,3:4])
       #Combine to work with graph
      distance <- dist(new,"euclidean",diag=FALSE,upper=FALSE)
      weig <- t(as.vector(distance))
      g <- graph.full(nrow(new))
      E(g)$weight <- weig
       #edge weights
      form <- cbind( get.edgelist(g) ,E(g)$weight)
       #Gettin graph info into a format compatible for optrees
      nodes <- 1:nrow(new)

      if (length(nodes)==1) {
        total_walk[i,j] <- 0
      } else {
        walk <<- searchWalk(nodes,form,directed = FALSE,
        start.node = nodes[1],end.node = length(nodes))
        total_walk[i,j] <- (sum(walk$walk.arcs[,3]))*100
         #Total walking distance for each stopping point in metres

     }
    }
  }
  cost <<- data.matrix(colSums(total_walk))
  hist(cost,col="turquoise3",xlab = "Total walking distance between houses (km)")
}

### Traditional stratified function ###
function (houses,stopping_points,stratified,q)
{total_walk <- matrix(1, nrow = 60, ncol = q)
  aaa <- matrix(0,nrow = q,ncol = 1)
  for(j in 1:q){
```

```
 set.seed(j+60)

 prop <- c(0.024291498,0.016194332,0.012145749,0.016194332,
               0.008097166,0.008097166,0.008097166,0.008097166,
               0.016194332,0.028340081,0.012145749,0.032388664,
               0.004048583,0.004048583,0.008097166,0.016194332,
               0.008097166,0.008097166,0.016194332,0.012145749,
               0.012145749,0.016194332,0.004048583,0.008097166,
               0.016194332,0.008097166,0.008097166,0.024291498,
               0.052631579,0.016194332,0.008097166,0.016194332,
               0.044534413,0.016194332,0.008097166,0.012145749,
           0.016194332,0.020242915,0.024291498,0.024291498,
           0.032388664,0.032388664,0.028340081,0.012145749,
           0.012145749,0.024291498,0.008097166,0.028340081,
           0.016194332,0.020242915,0.032388664,0.020242915,
           0.012145749,0.008097166,0.020242915,0.012145749,
           0.020242915,0.016194332,0.012145749,0.008097166)

 for (n in 124:247){
   size <- round(prop*n)
   ss <- strata(stratified,c("SP"),size, method="srswor")
   ID <- ss[,2]
   p <- stratified[ID,3]
   c <- sum(p)
   test <- c/701
   if(test>0.7){break}
}
 aaa[j] <- as.numeric(test*100)
 for(i in 1:60) {
   work <- subset(ss,SP==i)
    #Working with individual stopping points
   subbie <- houses[work$ID_unit,3:4]
   SP <- stopping_points[i,2:3]
   new <- rbind(SP,subbie)
    #Combine to work with graph
   distance <- dist(new,"euclidean",diag=FALSE,upper=FALSE)
```

```
      weig <- t(as.vector(distance))
      g <- graph.full(nrow(new))
      E(g)$weight <- weig
       #edge weights
     form <- cbind( get.edgelist(g) ,E(g)$weight)
       #Gettin graph info into a format compatible for optrees
      nodes <- 1:nrow(new)

      if (length(nodes)==1) {
        total_walk[i,j] <- 0
      } else {
        walk <<- searchWalk(nodes,form,directed = FALSE,
        start.node = nodes[1],end.node = length(nodes))
        total_walk[i,j] <- (sum(walk$walk.arcs[,3]))*100
         #Total walking distance for each stopping point in metres
     }
    }
 }
  cost <<- data.matrix(colSums(total_walk))
  hist(cost,col="orange1",xlab = "Total walking distance between houses (km)")

#### Traditional cluster sampling ####

function (houses,stopping_points,q)
{total_walk <- matrix(0, nrow = 60, ncol = q)
  aaa <- matrix(0,nrow = q,ncol = 1)
  for(j in 1:q){

    set.seed(j+60)

    prop <- c(6,4,3,4,2,2,2,2,4,7,
              3,8,1,1,2,4,2,2,4,3,
              3,4,1,2,4,2,2,6,13,4,
              2,4,11,4,2,3,4,5,6,6,
              8,8,7,3,3,6,2,7,4,5,
              8,5,3,2,5,3,5,4,3,2)
```

```
 for (n in 1:247){
   cl <- cluster(houses, clustername = c("SP"),n,method = "poisson",pik = prop)
   if (length(cl)==0){
     next
   }
   ID <- cl[,2]
   p <- houses[ID,5]
   c <- sum(p)
   test <- c/701
   if(test>0.7){break}
}

 aaa[j] <- as.numeric(test*100)
 for(i in 1:60) {
   work <- subset(cl,SP==i)
    #Working with individual stopping points

   if (nrow(work)==0){
     next
   }
   subbie <- houses[work$ID_unit,3:4]
   SP <- stopping_points[i,2:3]
   new <- rbind(SP,subbie)
    #Combine to work with graph
   distance <- dist(new,"euclidean",diag=FALSE,upper=FALSE)
   weig <- t(as.vector(distance))
   g <- graph.full(nrow(new))
   E(g)$weight <- weig
    #edge weights
   form <- cbind( get.edgelist(g) ,E(g)$weight)
    #Gettin graph info into a format compatible for optrees
   nodes <- 1:nrow(new)
   #if (length(nodes)==1) {
   #  total_walk[i,j] <- 0
   #} else {
```

```
      walk <<- searchWalk(nodes,form,directed = FALSE,
      start.node = nodes[1],end.node = length(nodes))
      total_walk[i,j] <- (sum(walk$walk.arcs[,3]))*100
       #Total walking distance for each stopping point in metres

      #}
    }
 }
  cost <<- data.matrix(colSums(total_walk))
  hist(cost,col="purple1",xlab = "Total walking distance between houses (km)")

}

#### Spatial random sampling ####
function (houses,stopping_points,q)
{total_walk <- matrix(0, nrow = 60, ncol = q)
  aaa <- matrix(0,nrow = q,ncol = 1)
  for(j in 1:q){
    set.seed(j+60)
    x_coord <- c(34.44354,  34.44354,34.47031,   34.47031, 34.44354)
    y_coord <- c(-1.5585796, -1.5244685,-1.5244685,  -1.5585796, -1.5585796)
    xym <- cbind(x_coord, y_coord)

    p = Polygon(xym)
    ps = Polygons(list(p),1)
    sps = SpatialPolygons(list(ps))

    for (n in 2:1000){
      sample.random <- spsample(sps, n, "random")
      a <- as.data.frame(sample.random)
      colnames(a) <- c("Latitude", "Longitude")
      r <- rbind(a,houses[,3:4])
       #First 100 columns is the sample, remaining are the houses
      d <- as.matrix(dist(r))[1:n,(n+1):(n+247)]
      actual_sample <- houses[as.vector(apply(d, 1, which.min)),]
      actual_sample <- actual_sample[!duplicated(actual_sample),]
```

```
    s <- sum(actual_sample[,5])/701
    if(s>0.7){break}
  }

  if(s<0.7){next}
  aaa[j] <- as.numeric(s*100)
  for(i in 1:60) {
    work <- subset(actual_sample,SP==i)
     #Working with individual stopping points
    if (nrow(work)==0){
      next
    }
    subbie <- work[,3:4]
    SP <- stopping_points[i,2:3]
    new <- rbind(SP,subbie)
     #Combine to work with graph
    distance <- dist(new,"euclidean",diag=FALSE,upper=FALSE)
    weig <- t(as.vector(distance))
    g <- graph.full(nrow(new))
    E(g)$weight <- weig
     #edge weights
    form <- cbind( get.edgelist(g) ,E(g)$weight)
     #Gettin graph info into a format compatible for optrees
    nodes <- 1:nrow(new)
    #if (length(nodes)==1) {
    #  total_walk[i,j] <- 0
    #} else {
    walk <<- searchWalk(nodes,form,directed = FALSE,
    start.node = nodes[1],end.node = length(nodes))
    total_walk[i,j] <- (sum(walk$walk.arcs[,3]))*100
     #Total walking distance for each stopping point in metres
    #}
  }
}

b <- as.vector(aaa[apply(aaa, 1, function(row) all(row !=0 ))])
```

```
  cost <<- data.matrix(colSums(total_walk))
  hist(cost,col="green1",xlab = "Total walking distance between houses (km)")
}
#### Spatial regular sampling ####
function (houses,stopping_points,q)
{total_walk <- matrix(500, nrow = 60, ncol = q)
  aaa <- matrix(0,nrow = q,ncol = 1)
  for(j in 1:q){

    x_coord <- c(34.44354,  34.44354,34.47031,   34.47031, 34.44354)
    y_coord <- c(-1.5585796, -1.5244685,-1.5244685,  -1.5585796, -1.5585796)
    xym <- cbind(x_coord, y_coord)

    p = Polygon(xym)
    ps = Polygons(list(p),1)
    sps = SpatialPolygons(list(ps))

    for (n in 2:1000){
      sample.random <- spsample(sps, n, "regular")
      a <- as.data.frame(sample.random)
      colnames(a) <- c("Latitude", "Longitude")
      r <- rbind(a,houses[,3:4])
       #First 100 columns is the sample, remaining are the houses
      d <- as.matrix(dist(r))[1:nrow(a),(nrow(a)+1):nrow(r)]
      if(is.null(dim(d))==TRUE){break}
      actual_sample <- houses[as.vector(apply(d, 1, which.min)),]
      actual_sample <- actual_sample[!duplicated(actual_sample),]
      s <- sum(actual_sample[,5])/701
      if(s > 0.7) {break}
    }

    if(s<0.7){next}
    aaa[j] <- as.numeric(s*100)

    for(i in 1:60) {
      work <- subset(actual_sample,SP==i)
```

```
      #Working with individual stopping points
      if (nrow(work)==0){
        next
      }
      subbie <- work[,3:4]
      SP <- stopping_points[i,2:3]
      new <- rbind(SP,subbie)
      #Combine to work with graph
      distance <- dist(new,"euclidean",diag=FALSE,upper=FALSE)
      weig <- t(as.vector(distance))
      g <- graph.full(nrow(new))
      E(g)$weight <- weig
      #edge weights
      form <- cbind( get.edgelist(g) ,E(g)$weight)
      #Gettin graph info into a format compatible for optrees
      nodes <- 1:nrow(new)
      #if (length(nodes)==1) {
      #  total_walk[i,j] <- 0
      #} else {
      walk <<- searchWalk(nodes,form,directed = FALSE,
      start.node = nodes[1],end.node = length(nodes))
      total_walk[i,j] <- (sum(walk$walk.arcs[,3]))*100
      #Total walking distance for each stopping point in metres
     #}
    }
 }
 b <- as.vector(aaa[apply(aaa, 1, function(row) all(row !=0 ))])

  cost <<- data.matrix(colSums(total_walk))
  hist(cost,col="yellow1",xlab = "Total walking distance between houses (km)")
}

#### Spatial stratified sampling ####
function (houses,stopping_points,q)
{
total_walk <- matrix(0, nrow = 60, ncol = q)
```

```
aaa <- matrix(0,nrow = q,ncol = 1)

for(j in 1:q){
  x_coord <- c(34.44354,34.44354,34.47031,34.47031,34.44354)
  y_coord <- c(-1.5585796, -1.5244685,-1.5244685,-1.5585796,-1.5585796)
  xym <- cbind(x_coord, y_coord)

  p = Polygon(xym)
  ps = Polygons(list(p),1)
  sps = SpatialPolygons(list(ps))

  for (n in 100:1000){
    sample.random <- spsample(sps, n, "stratified")
    a <- as.data.frame(sample.random)
    colnames(a) <- c("Latitude", "Longitude")
    r <- rbind(a,houses[,3:4])
     #First 100 columns is the sample, remaining are the houses
    d <- as.matrix(dist(r))[1:nrow(a),(nrow(a)+1):nrow(r)]
    if(is.null(dim(d))==TRUE){break}
    actual_sample <- houses[as.vector(apply(d, 1, which.min)),]
    actual_sample <- actual_sample[!duplicated(actual_sample),]
    s <- sum(actual_sample[,5])/701
    if(s > 0.7) {break}
  }
  if(s<0.7){next}
  aaa[j] <- as.numeric(s*100)

  for(i in 1:60) {
   work <- subset(actual_sample,SP==i)
     #Working with individual stopping points
   if (nrow(work)==0){
      next
    }
   subbie <- work[,3:4]
    SP <- stopping_points[i,2:3]
    new <- rbind(SP,subbie)
```

```
      #Combine to work with graph
      distance <- dist(new,"euclidean",diag=FALSE,upper=FALSE)
     weig <- t(as.vector(distance))
     g <- graph.full(nrow(new))
     E(g)$weight <- weig
      #edge weights
      form <- cbind( get.edgelist(g) ,E(g)$weight)
      #Gettin graph info into a format compatible for optrees
     nodes <- 1:nrow(new)
     #if (length(nodes)==1) {
     #  total_walk[i,j] <- 0
     #} else {
     walk <<- searchWalk(nodes,form,directed = FALSE,
     start.node = nodes[1],end.node = length(nodes))
     total_walk[i,j] <- (sum(walk$walk.arcs[,3]))*100
      #Total walking distance for each stopping point in metres
    #}
 }
}

b <- as.vector(aaa[apply(aaa, 1, function(row) all(row !=0 ))])
cost <<- data.matrix(colSums(total_walk))
hist(cost,col="red1",xlab = "Total␣walking␣distance␣between␣houses␣(km)")
}
#### Systematic Nonaligned Sampling ####
function (houses,stopping_points,q)
{total_walk <- matrix(0, nrow = 60, ncol = q)
aaa <- matrix(0,nrow = q,ncol = 1)

for(j in 1:q){
  x_coord <- c(34.44354,34.44354,34.47031,34.47031,34.44354)
  y_coord <- c(-1.5585796, -1.5244685,-1.5244685,-1.5585796,-1.5585796)
  xym <- cbind(x_coord, y_coord)
  p = Polygon(xym)
  ps = Polygons(list(p),1)
  sps = SpatialPolygons(list(ps))
```

```
for (n in 100:1000){
  sample.random <- spsample(sps, n, "nonaligned")
  a <- as.data.frame(sample.random)
  colnames(a) <- c("Latitude", "Longitude")
  r <- rbind(a,houses[,3:4])
   #First 100 columns is the sample, remaining are the houses
  d <- as.matrix(dist(r))[1:nrow(a),(nrow(a)+1):nrow(r)]
  if(is.null(dim(d))==TRUE){break}
  actual_sample <- houses[as.vector(apply(d, 1, which.min)),]
  actual_sample <- actual_sample[!duplicated(actual_sample),]
  s <- sum(actual_sample[,5])/701
  if(s > 0.7) {break}
}

if(s<0.7){next}
aaa[j] <- as.numeric(s*100)
for(i in 1:60) {
  work <- subset(actual_sample,SP==i)
   #Working with individual stopping points
   if (nrow(work)==0){
    next
  }
  subbie <- work[,3:4]
  SP <- stopping_points[i,2:3]
  new <- rbind(SP,subbie)
   #Combine to work with graph
  distance <- dist(new,"euclidean",diag=FALSE,upper=FALSE)
  weig <- t(as.vector(distance))
  g <- graph.full(nrow(new))
  E(g)$weight <- weig
   #edge weights
  form <- cbind( get.edgelist(g) ,E(g)$weight)
   #Gettin graph info into a format compatible for optrees
  nodes <- 1:nrow(new)
   #if (length(nodes)==1) {
```

```
    #  total_walk[i,j] <- 0
    #} else {
    walk <<- searchWalk(nodes,form,directed = FALSE,
    start.node = nodes[1],end.node = length(nodes))
    total_walk[i,j] <- (sum(walk$walk.arcs[,3]))*100
     #Total walking distance for each stopping point in metres
     #}
  }
 }
b <- as.vector(aaa[apply(aaa, 1, function(row) all(row !=0 ))])

cost <<- data.matrix(colSums(total_walk))
hist(cost,col="springgreen1",xlab = "Total walking distance between houses (km)")
}
#### Hexagonal Systematic Sampling ####
function (houses,stopping_points,q)
{total_walk <- matrix(0, nrow = 60, ncol = q)
aaa <- matrix(0,nrow = q,ncol = 1)

for(j in 1:q){
  x_coord <- c(34.44354,34.44354,34.47031,34.47031,34.44354)
  y_coord <- c(-1.5585796, -1.5244685,-1.5244685,-1.5585796,-1.5585796)
  xym <- cbind(x_coord, y_coord)

  p = Polygon(xym)
  ps = Polygons(list(p),1)
  sps = SpatialPolygons(list(ps))
 for (n in 100:1000){
    sample.random <- spsample(sps, n, "hexagonal")
    a <- as.data.frame(sample.random)
    colnames(a) <- c("Latitude", "Longitude")
    r <- rbind(a,houses[,3:4])
     #First 100 columns is the sample, remaining are the houses
    d <- as.matrix(dist(r))[1:nrow(a),(nrow(a)+1):nrow(r)]
    if(is.null(dim(d))==TRUE){break}
    actual_sample <- houses[as.vector(apply(d, 1, which.min)),]
```

```
    actual_sample <- actual_sample[!duplicated(actual_sample),]
    s <- sum(actual_sample[,5])/701
    if(s > 0.7) {break}
  }
  if(s<0.7){next}
  aaa[j] <- as.numeric(s*100)

  for(i in 1:60) {
     work <- subset(actual_sample,SP==i)
     #Working with individual stopping points
   if (nrow(work)==0){
      next
    }
   subbie <- work[,3:4]
    SP <- stopping_points[i,2:3]
    new <- rbind(SP,subbie)
     #Combine to work with graph
   distance <- dist(new,"euclidean",diag=FALSE,upper=FALSE)
    weig <- t(as.vector(distance))
    g <- graph.full(nrow(new))
    E(g)$weight <- weig
     #edge weights
   form <- cbind( get.edgelist(g) ,E(g)$weight)
     #Gettin graph info into a format compatible for optrees
    nodes <- 1:nrow(new)
    #if (length(nodes)==1) {
    #  total_walk[i,j] <- 0
    #} else {
    walk <<- searchWalk(nodes,form,directed = FALSE,
    start.node = nodes[1],end.node = length(nodes))
    total_walk[i,j] <- (sum(walk$walk.arcs[,3]))*100
     #Total walking distance for each stopping point in metres
   #}
  }
}
```

```
b <- as.vector(aaa[apply(aaa, 1, function(row) all(row !=0 ))])

cost <<- data.matrix(colSums(total_walk))
hist(cost,col="chocolate1",xlab = "Total␣walking␣distance␣between␣houses␣(km)")
}

#### Clustered Spatial Sampling ####
function (houses,stopping_points,q)
{total_walk <- matrix(0, nrow = 60, ncol = q)
aaa <- matrix(0,nrow = q,ncol = 1)

for(j in 1:q){
  x_coord <- c(34.44354,34.44354,34.47031,34.47031,34.44354)
  y_coord <- c(-1.5585796, -1.5244685,-1.5244685,-1.5585796,-1.5585796)
  xym <- cbind(x_coord, y_coord)

  p = Polygon(xym)
  ps = Polygons(list(p),1)
  sps = SpatialPolygons(list(ps))
  ?spsample
  for (n in 100:1000){
    sample.random <- spsample(sps, n, "clustered",nclusters=80)
    a <- as.data.frame(sample.random)
    colnames(a) <- c("Latitude", "Longitude")
    r <- rbind(a,houses[,3:4])
     #First 100 columns is the sample, remaining are the houses
    d <- as.matrix(dist(r))[1:nrow(a),(nrow(a)+1):nrow(r)]
    if(is.null(dim(d))==TRUE){break}
    actual_sample <- houses[as.vector(apply(d, 1, which.min)),]
    actual_sample <- actual_sample[!duplicated(actual_sample),]
    s <- sum(actual_sample[,5])/701
    if(s > 0.7) {break}
}

  if(s<0.7){next}
  aaa[j] <- as.numeric(s*100)
```

```
 for(i in 1:60) {
    work <- subset(actual_sample,SP==i)
     #Working with individual stopping points
    if (nrow(work)==0){
      next
    }
    subbie <- work[,3:4]
    SP <- stopping_points[i,2:3]
    new <- rbind(SP,subbie)
     #Combine to work with graph
    distance <- dist(new,"euclidean",diag=FALSE,upper=FALSE)
    weig <- t(as.vector(distance))
    g <- graph.full(nrow(new))
    E(g)$weight <- weig
     #edge weights
    form <- cbind( get.edgelist(g) ,E(g)$weight)
     #Gettin graph info into a format compatible for optrees
    nodes <- 1:nrow(new)
    #if (length(nodes)==1) {
    #  total_walk[i,j] <- 0
    #} else {
    walk <<- searchWalk(nodes,form,directed = FALSE,
    start.node = nodes[1],end.node = length(nodes))
    total_walk[i,j] <- (sum(walk$walk.arcs[,3]))*100
     #Total walking distance for each stopping point in metres
   #}
 }
}
b <- as.vector(aaa[apply(aaa, 1, function(row) all(row !=0 ))])

cost <<- data.matrix(colSums(total_walk))
hist(cost,col="springgreen1",xlab = "Total␣walking␣distance␣between␣houses␣(km)")
}
```

Bibliography

[1] L. Addario-Berry, N. Broutin, C. Goldschmidt, and G. Miermont. The scaling limit of the minimum spanning tree of the complete graph. *The Annals of Probability*, 45(5):3075–3144, 2017.

[2] M. Armstrong. *Basic Linear Geostatistics*. Springer Science & Business Media, 1998.

[3] V. Barnett. *Elements of sampling theory*. The English Universities Press Ltd, 1974.

[4] V. Barnett. *Sample survey: principles and methods*. British Library Cataloguing in Publication Data, 1991.

[5] J. J. Battles, J. G. Dushoff, and T. J. Fahey. Line intersect sampling of forest canopy gaps. *Forest Science*, 42(2):131–138, 1996.

[6] D. Binder and G. Roberts. Design-based and model-based methods for estimating model parameters. *Analysis of Survey Data*, pages 29–48, 2003.

[7] R. Bivand, E. Pebesma, and V. GomezRubio. *Applied Spatial Data Analysis with R*. Springer, 2008.

[8] R. . P. Bivand. *Classes and methods for spatial data in R*.

[9] K. Bostoen and Z. Chalabi. Optimization of household survey sampling without sample frames. *International Journal of Epidemiology*, 35(3):751–755, 2006.

[10] K. Bostoen, Z. Chalabi, and R. F. Grais. Optimisation of the T-square sampling method to estimate population sizes. *Emerging Themes in Epidemiology*, 4(1):7, 2007.

[11] D. Brus and J. de Gruijter. Random sampling or geostatistical modelling? choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, 80(1):1–44, 1997.

[12] D. Brus, J. De Gruijter, and J. Van Groenigen. Designing spatial coverage samples using the k-means clustering algorithm. *Developments in Soil Science*, 31:183–192, 2007.

[13] S. Cleaveland, M. Kaare, P. Tiringa, T. Mlengeya, and J. Barrat. A dog rabies vaccination campaign in rural africa: impact on the incidence of dog rabies and human dog-bite injuries. *Vaccine*, 21(17):1965–1973, 2003.

[14] W. Cochran. Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics*, pages 164–177, 1946.

[15] W. Cochran. Sampling Techniques. *Wiley Series*, 1953.

[16] P. G. Coleman and C. Dye. Immunization coverage required to prevent outbreaks of dog rabies. *Vaccine*, 14(3):185–186, 1996.

[17] A. Colman. A Dictionary of Psychology. *Oxford University Press*, 2015.

[18] T. H. Cormen. *Introduction to Algorithms*. MIT press, 2009.

[19] R. Corporation. *A Million Random Digits with 100, 000 Normal deviates*. RAND, 1955.

[20] P. Crepel, S. Fienberg, J. Gani, C. Heyde, and E. Seneta. *Statisticians of the Centuries*. Springer Science & Business Media, 2013.

[21] N. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 1993.

[22] J. J. de Gruijter and C. ter Braak. Model-free estimation from spatial samples: A reappraisal of classical sampling theory. *Mathematical Geology*, 1990.

[23] A. Dean, M. Morris, J. Stufken, and D. Bingham. *Handbook of Design and Analysis of Experiments*, volume 7. CRC Press, 2015.

[24] E. Delmelle. Spatial sampling. In A. Fotheringham and P. Rogerson, editors, *The SAGE Handbook of Spatial Analysis*, chapter 10, pages 183 – 206. SAGE Publications, Ltd, 2011.

[25] J. Devore and K. Berk. *Modern Mathematical Statistics with Applications*. Springer Science & Business Media, 2012.

[26] C. J. Dixon and B. Leach. Sampling methods for geographical research. In *Geo Abstracts Norwich*, 1977.

[27] M. D'Orazio. Estimating the variance of the sample mean in two-dimensional systematic sampling. *Journal of Agricultural, Biological, and Environmental Statistics*, 8(3):280, 2003.

[28] S. Durr, R. Mindekem, Y. Kaninga, D. D. Moto, M. Meltzer, P. Vounatsou, and J. Zinsstag. Effectiveness of dog rabies vaccination programmes: comparison of owner-charged and free vaccination campaigns. *Epidemiology & Infection*, 137(11):1558–1567, 2009.

[29] S. S. Epp. *Discrete Mathematics with Applications*. Cengage Learning, 2010.

[30] M. Fahimi. Cluster sampling. In P. Lavrakas, editor, *Encyclopedia of Survey Research Methods*, page 99. Sage Publications, Inc., 2008.

[31] M. Fontenla. *optrees: Optimal Trees in Weighted Graphs*, 2014. R package version 1.0.

[32] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics, Berlin, 2001.

[33] A. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes. *Handbook of Spatial Statistics*. CRC Press, 2010.

[34] T. G. Gregoire. Notes: The unbiasedness of the mirage correction procedure for boundary overlap. *Forest Science*, 28(3):504–508, 1982.

[35] T. G. Gregoire. Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research*, 28(10):1429–1447, 1998.

[36] T. Gregorie and V. HT. *Sampling Strategies for Natural Resources and the Environment*. Chapman & Hall / CRC, 2007.

[37] R. Haining. *Spatial Data Analysis: Theory and Practice*. Cambridge University Press, 2003.

[38] M. Hansen, W. Hurwitz, and W. Madow. *Sample Survey Methods and Theory*, volume I. John Wiley & Sons, 1953.

[39] E. Isaaks and R. Srivastava. *Applied Geostatistics*, volume 2. Oxford University Press New York, 1989.

[40] J. Ježek, K. Jedlička, T. Mildorf, J. Kellar, and D. Beran. Design and evaluation of Web gl-based heat map visualization for big point data. In *The Rise of Big Spatial Data*, pages 13–26. Springer, 2017.

[41] A. Journel and C. Huijbregts. *Mining Geostatistics*. Academic press, 1978.

[42] L. Kaiser. Unbiased estimation in line-intercept sampling. *Biometrics*, pages 965–976, 1983.

[43] W. Kalsbeek. Stratified sampling. In P. Lavrakas, editor, *Encyclopedia of Survey Research Methods*, pages 850 – 851. Sage Publications, Inc., 2008.

[44] G. Kalton. *Introduction to Survey Sampling*, volume 35. Sage, 1983.

[45] U. Kayali, R. Mindekem, N. Yemadji, P. Vounatsou, Y. Kaninga, A. Ndoutamia, and J. Zinsstag. Coverage of pilot parenteral vaccination campaign against canine rabies in n'djamena, chad. *Bulletin of the World Health Organization*, 81(10):739–744, 2003.

[46] M. Kendall and B. Smith. Randomness and random sampling numbers. *Journal of the Royal Statistical Society*, 101(1):147–166, 1938.

[47] S. Killip, Z. Mahfoud, and K. Pearce. What is an intracluster correlation coefficient? crucial concepts for primary care researchers. *The Annals of Family Medicine*, 2(3):204–208, 2004.

[48] P. Kitala, J. McDermott, P. Coleman, and C. Dye. Comparison of vaccination strategies for the control of dog rabies in machakos district, kenya. *Epidemiology & Infection*, 129(1):215–222, 2002.

[49] S. Lohr. *Sampling: Design and Analysis*. Nelson Education, 2009.

[50] H. Lucas and G. A. F. Seber. Estimating coverage and particle density using the line intercept method. *Biometrika*, pages 618–622, 1977.

[51] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, volume 14, pages 281–297. Oakland, CA, USA., 1967.

[52] W. Madow. On the theory of systematic sampling, III. *The Annals of Mathematical Statistics*, pages 101–106, 1953.

[53] B. Matérn. *Spatial Variation*, volume 36. Springer Science & Business Media, 2013.

[54] A. Milne. The centric systematic area-sample treated as a random sample. *Biometrics*, 15(2):270–297, 1959.

[55] L. Moores, B. Pittman, and G. Kitchen. Forest ecological classification and mapping: their application for ecosystem management in newfoundland. *Environmental Monitoring and Assessment*, 39(1):571–577, 1996.

[56] R. Pandey and M. R. Verma. Samples allocation in different strata for impact evaluation of developmental programme. *Rev. Mat. Estat*, 26(4):103–112, 2008.

[57] E. Pebesma, R. Bivand, B. Rowlingson, V. Gomez-Rubio, R. Hijmans, M. Sumner, D. MacQueen, J. Lemon, and J. O'Brien. *sp: Classes and Methods for Spatial Data*, 2016. R package version 1.2-4.

[58] M. D. Perlman and M. J. Wichura. Sharpening buffons needle. *The American Statistician*, 29(4):157–163, 1975.

[59] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[60] B. D. Ripley. *Spatial Statistics*, volume 575. John Wiley & Sons, 2005.

[61] J. A. Royle and D. Nychka. An algorithm for the construction of spatial coverage designs with implementation in splus. *Computers & Geosciences*, 24(5):479–488, 1998.

[62] N. Sehar, M. Ahsan, and M. G. Khan. Mixed allocation in stratified sampling. *Aligarh Journal of Statistics*, 25:1–11, 2005.

[63] Y. Sembiko. Rabies in Tanzania. In *Proceedings of the 3rd International Conference of the Southern and Eastern African Rabies Group, Harare, Zimbabwe*, pages 29–32, 1995.

[64] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[65] A. Smit. Interpolation in stationary spatial and spatial-temporal datasets. Master's thesis, University of Pretoria, 2010.

[66] D. L. Stevens Jr and A. R. Olsen. Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465):262–278, 2004.

[67] S. Thompson. *Sampling*. John Wiley & Sons, second edition, 2002.

[68] Y. Tillé and A. Matei. *sampling: Survey Sampling*, 2016. R package version 2.8.

[69] W. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46:234–240, 1970.

[70] A. Tucker. *Applied Combinatorics*. John Wiley & Sons, Inc., 2006.

[71] J. V. Uspensky. *Introduction to Mathematical Probability*. McGraw-Hill, 1937.

[72] A. Van Laar and A. Akça. *Forest Mensuration*, volume 13. Springer Science & Business Media, 2007.

[73] D. Walvoort, D. Brus, and J. De Gruijter. An r package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers & Geosciences*, 36(10):1261–1267, 2010.

[74] J. Wang, A. Stein, B. Gao, and Y. Ge. A review of spatial sampling. *Spatial Statistics*, 2:1–14, 2012.

[75] R. Webster and M. Oliver. *Geostatistics for Environmental Scientists*. John Wiley & Sons, 2007.

[76] N. Wollenhaupt and R. Wolkowski. Grid soil sampling. *Better Crops*, 78(4):6–9, 1994.

[77] E. Yfantis, G. Flatman, and J. Behar. Efficiency of Kriging estimation for square, triangular, and hexagonal grids. *Mathematical Geology*, 19(3):183–205, 1987.

[78] J. Zinsstag, S. Dürr, M. Penny, R. Mindekem, F. Roth, S. M. Gonzalez, S. Naissengar, and J. Hattendorf. Transmission dynamics and economics of rabies control in dogs and humans in an African city. *Proceedings of the National Academy of Sciences*, 106(35):14996–15001, 2009.