Spatial Statistics 65 (2025) 100877

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

An optimised rabies vaccination schedule for rural settlements

Rian Botes [a], Inger Fabris-Rotelli [a,*], Kabelo Mahloromela [a], Ding-Geng Chen [a,b]

[a] University of Pretoria, Department of Statistics, South Africa
[b] College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA

ARTICLE INFO

Keywords:
Rabies
Sampling
Spatial sampling
Point patterns
Vaccination
Rural settlement
Kernel-weighted spatial sampling

ABSTRACT

The timely and efficient administration of rabies vaccinations to animals in rural villages is necessary to attain a state of herd immunity. Efficient sampling of households in a rural village is of utmost importance in reaching the most animals for vaccination, with the least effort, and in the lowest time. This research seeks to both optimise the spatial sampling scheme used to sample households, as well as the route travelled by persons performing door-to-door vaccinations. The walking time in minutes is regarded as the cost of a vaccination scheme and is minimised in this paper. The distribution of houses in a rural village constitutes a spatial point pattern in $\mathbb{R}^2$, and as such, spatial point pattern analysis techniques as well as some spatial sampling schemes are applied throughout this research. The penultimate aim of this work is to provide policy makers with additional tools to combat rabies, a disease which remains endemic to some countries in West and Central Africa, and Asia.

Contents

* Corresponding author.
E-mail address: inger.fabris-rotelli@up.ac.za (I. Fabris-Rotelli).

https://doi.org/10.1016/j.spasta.2024.100877
Received 1 April 2023; Received in revised form 5 December 2024; Accepted 9 December 2024
Available online 17 December 2024

1. Introduction

A One Health research approach to combatting rabies requires attention paid not only to human health, but also to animal health (Lebov et al., 2017). Not only is it cheaper to vaccinate animals than humans against rabies (Lavan et al., 2017), but the vaccination of animals is the only way to confidently curb the spread of rabies in a rabies endemic region (Mbilo et al., 2021). Since dog transmitted rabies accounts for nearly all rabies cases (World Health Organization, 2018), annual mass dog vaccination campaigns are often carried out in rabies stricken regions (Hampson et al., 2009; Wera et al., 2017). Such sustained mass dog vaccination campaigns have proven effective in reducing instances of rabies in rabies endemic countries (Arief et al., 2017; Lavan et al., 2017; Mbilo et al., 2021; Undurraga et al., 2020). Due to the fact that rabies is most common in developing and resource-limited nations on the African and Asian continents (Masiira et al., 2018), the vaccination schedule developed in this research is optimised for rural areas where rabies is especially prevalent (Mbilo et al., 2021). Rural village data from a household census in Tanzania is therefore used to implement the optimised rabies vaccination schedule.

Two main alternatives to dog vaccination exist. The culling of a dog population under which rabies is prevalent is one option, while the vaccination of humans is another. The culling of a dog population is not as effective as animal vaccination, and is regarded as an 'expensive distraction' from the more important dog vaccination (Arief et al., 2017). From the study performed in Arief et al. (2017), dog culling had no effect, and in certain regions even worsened the condition of rabies. This is most likely due to complacency following what seems to have been an effective wave of dog culling. The vaccination of a human population, and post-exposure treatment is considerably more expensive than vaccinating dogs (Lavan et al., 2017). Furthermore, vaccinating humans instead of dogs does little to prevent the spread of rabies among a dog population. Focusing only on human vaccination leaves the virus unchecked in the animal population. Rabies infections will therefore continue to spread in regions where it is already endemic, posing a threat to bordering regions currently free of rabies. The vaccination of dogs therefore remains the most effective means to control the spread of rabies and in the long term, eradicate the disease from a previously endemic region. This research therefore focuses efforts on optimising a vaccination schedule aimed at a dog population, instead of a human population.

The four main methods for vaccination are static point vaccination stations, door-to-door vaccination, capture-vaccinate-release methods, and oral vaccinations in the form of medically treated dog treats placed around a village upon which, once swallowed, a dog receives a dosage of a rabies vaccination (Undurraga et al., 2020). The most popular method of dog vaccination is the use of static vaccination points (Fabris-Rotelli et al., 2020). This method, however, while cost effective, is dependent on dog owners bringing their dogs to these vaccination points. A door-to-door vaccination scheme ensures greater vaccination coverage by taking the vaccine to the dog, instead of waiting for the dog owner to bring their dog to the vaccine. This research therefore seeks to optimise a door-to-door vaccination schedule.

This paper focuses on a rural village setting for vaccination, specifically in Tanzania, Africa. The aim is to design a door-to-door vaccination scheme that vaccinates all animals found at a residence.

A study performed by Knobel et al. (2008) provides further insights into why rural households in Tanzania own dogs. It was found that dogs are more prominent in households that also own livestock, however only 12.7% of these livestock-owning households indicated cattle herding and protection as their cardinal reason for domesticating dogs. What is interesting is that 23.5% of livestock-owning households indicated that their primary reason for keeping dogs is to deter rodents and other pests from infesting their crops. Perhaps the most reassuring finding for this door-to-door vaccination schedule is that the majority of both livestock and non-livestock-owning households (61.9% and 76.4% respectively) own dogs to protect their homes from unwanted intruders. While this suggests that a large proportion of Tanzanian-owned dogs would be present upon the arrival of a vaccinator, a significant percentage of dogs can almost definitely be expected absent while roaming. Vaccinating all dogs the vaccination encounters will obtain larger coverage. Methods other than a door-to-door vaccination schedule will need to be employed to reach domesticated (owned) dogs who are either herding animals, or protecting crops. The matter of inoculating free-roaming (unowned) dogs can also not be addressed by a door-to-door vaccination technique.

The vaccination of owned dogs, who are out in crops and pastures during the day is arguably more imperative than vaccinating dogs that remain at home, as these travelling dogs are at a greater risk of encountering other carriers of this disease, especially free-roaming dogs or bats.

It may also be more likely for such dogs to contract rabies as they roam around outside the village. Younger dogs are also more susceptible to infection due to the fact that they are less likely to be vaccinated than older dogs, who were alive during previous vaccination campaigns. It is also a belief that puppies should not be vaccinated, in fear that their immune systems are not yet mature enough to seroconvert against a rabies vaccine (Arief et al., 2017; Morters et al., 2015). This idea is however untrue and belief therein may lead to greater risk for the young dog population contracting and transmitting rabies (Morters et al., 2015). This research aims to further equip both private and public health sectors with a novel method to better understand the distribution of potential rabies risk in rural areas, and to then mitigate this risk through effective mass dog vaccination campaigns. This research is also performed to be in line with the Zero by 30 campaign (World Health Organization, 2018, 2019) led by the World Health Organisation (WHO).

This research improves upon the vaccination scheme of Fabris-Rotelli et al. in Fabris-Rotelli et al. (2020) by applying a travelling salesman algorithm instead of a minimum spanning tree algorithm to calculate the walking distance between houses. In addition, a

new sampling scheme is also proposed in this research, named the kernel-weighted spatial sampling scheme. Therein, the stopping point selection process is optimised by means of determining the least amount of stopping points needed to reach the maximum number of houses. While Fabris-Rotelli et al. used a kernel density estimate to determine stopping points, this research considers k-means to select stopping points. Ultimately, this study further optimises the simulation framework proposed by Fabris-Rotelli et al.

The remainder of this document is structured as follows. Section 2 gives a thorough presentation of the methodology and assumptions used to construct the optimised vaccination schedule. Section 3 provides further insight into the dataset used in the paper and the parameter tuning for the k-means algorithm. Section 4 discusses the results as well as future work. This paper is then concluded in Section 5.

2. Methodology

The proposed vaccination schedule herein introduces a new sampling scheme, the kernel-weighted spatial sampling scheme. Stopping points are selected to which a vaccinator should travel using a vehicle, and houses within 200 metres of these stopping points are deemed accessible to a door-to-door vaccinator. The distance of 200 metres is deemed reasonable for a vaccinator on foot (see Fabris-Rotelli et al. (2020) for additional explanation). Different sampling schemes are then used to sample at least 70% of the accessible houses for dog vaccination, and the distance required to walk from each stopping point and the sampled houses is calculated. The sampling scheme resulting in the shortest total walking distance to reach 70% of all the dogs in a village is the preferred sampling scheme. This research further seeks to improve the methods proposed by Fabris-Rotelli et al. by employing a travelling salesman algorithm (TST) instead of a minimum spanning tree (MST) algorithm when calculating the route that a vaccinator should walk between sampled houses. It is briefly explained in the methodology of this research that a TST approach always results in a walking distance that is shorter than that calculated by an MST (Michael and Kurt, 2007). A second contribution is that this research optimises the parameter k of the k-means algorithm used to select stopping points. While Fabris-Rotelli et al. used a kernel density estimate as well, the bandwidth parameter of this kernel density estimate is not optimised, and requires human intervention.

We define some notation next to enable to presentation of the methodology that follows. Let N denote the total number of house in the population, let n denote the number of houses in a sample from N, such that $n < N$, and let S denote a sample itself. For a village with K stopping points, the sample S of n houses is segmented into these K groups. Furthermore, let $n_k, k = 1, 2, \ldots, K$ denote the number of sampled houses within a 200 metre radius from the kth stopping point so that $n = \sum_{k=1}^{K} n_k$. This 200 metre radius represents the accessible houses at stopping point k. We denote the total number of accessible houses in a village as $N_a \leq N$. The aim of the sampling herein is to sample from N_a. The vaccination coverage is represented as the percentage of N then sampled. The 70% coverage required by WHO will thus only be reached if the K stopping points are well designed. Our simulation study investigates this within the various possibilities.

2.1. Stopping point selection using k-means

The ideal stopping point algorithm allocates the least number of stopping points while simultaneously maximising the number of accessible houses N_a. The method to select stopping points is k-means (MacKay, 2003).

Let x denote a point pattern, spawned by some unobserved point process X. Furthermore, let the point pattern x be observed in the window W. The k initial values for the k-means algorithm can be chosen from this point pattern x, or as random (or pre-determined) points on the window W. Unless the initial values of the k-means algorithm is fixed, a different clustering solution is produced after each instance. In the current application, this means that the same number of k stopping points can result in different sets of attainable houses N_a, and thus different coverage percentages in terms of houses. The fact that the k-means clustering algorithm produces a different solution depending on random starting values poses a slight challenge. For some fixed number of k stopping points, the location of these points can vary significantly depending on the starting values. Since the placement of the k stopping points vary, so too does the important value of N_a. In a very real sense, some sets of k initial values are better than others, given that they result in a final solution with more accessible houses.

The k-means stopping point selection algorithms require the parameter k to be set as a priori. This parameter determines the number of stopping points in a village. Since only houses within a 200 metre radius of a stopping point are considered accessible to a vaccinator, more stopping points are required in order to increase N_a. While the distance travelled by a vaccination between stopping points (by car) is considered negligible in this research context, it remains necessary for the number of stopping points to be as little as possible to achieve a practical solution, while also ensuring that N_a is as large as possible. The reason why the distance travelled between stopping points is considered negligible is because vaccinators would use vehicles to travel between stopping points. While this will still take time and cost fuel, the cost relative to the vaccinators time and energy spent walking between houses and vaccinating animals is different enough to warrant the driving costs to be excluded from this study. The cost of moving between stopping points should, of course, not be ignored when planning a rabies vaccination campaign. It is therefore this cost that should be minimised.

The bullet point outline below shows the step-by-step process used to generate walking distance distributions in rural villages, in order to determine which sampling scheme yields the optimal solution (shortest walking distance). This process should also be used when the reader would like to run their own simulation using the code provided here, on GitHub.

1. Select a sampling scheme.

2. Select a TST algorithm (The farthest insertion heuristic is recommended).
3. Select a sampling coverage (30%, 40%, 50%, 60% or 70%).
4. Find the value of k needed to generate the right stopping point configuration for the desired coverage (see Table 1 for k-means).
5. Generate the k stopping points using your selected stopping point algorithm (remember to initiate k-means with the right random seed from Table 1.)
6. Generate 10 000 random samples from the set of attainable houses N_a using your chosen sampling scheme.
7. Calculate the total walking distance between each of the k stopping point and its surrounding houses. Do this for every sample using your chosen TST algorithm. The result is a set of 10 000 walking distances.
8. Calculate descriptive statistics on the 10 000 walking distances, and plot the distribution.

2.2. Sampling schemes

A trade-off exists between obtaining maximum vaccination coverage, and keeping the walking distance of a door-to-door vaccination approach as low as possible. Since the household locations in a village constitute spatial point pattern data, the use of spatial sampling schemes seem an obvious approach given the current spatial context. Two traditional (non-spatial) sampling schemes are also explored here for further insight, and for the purpose of comparing the effectiveness of spatial sampling schemes to traditional sampling schemes in addressing this trade-off between vaccination coverage and walking distance. This research looks into seven of the eight sampling schemes used by Fabris-Rotelli et al. in Fabris-Rotelli et al. (2020), while also applying the proposed new spatial sampling scheme.

The two non-spatial sampling schemes applied in this research are the simple random sampling scheme and the traditional stratified sampling. Five spatial sampling schemes are applied in this research. These schemes are the uniform, stratified, systematic regular, systematic non-aligned, and systematic hexagonal spatial sampling schemes

2.2.1. Traditional simple random sampling (SRS)

Traditional simple random sampling (SRS) is performed directly on the list of households in a village without considering their geographic location. The `sample()` function in R (R Core Team, 2022) is used to execute SRS without replacement. Sampling is configured without replacement since it is not necessary to visit the same house for vaccination more than once. The sample size is always specified to be 70% of the available houses, N_a, constituting a sample size of $n = 0.7$ of N_a. The SRS scheme ensures that each household in N_a has an equal chance of being selected, with a probability of selection equal to $\binom{N_a}{n}^{-1}$.

2.2.2. Traditional cluster sampling (TCS)

Having determined K stopping points, a traditional cluster sampling (TCS) scheme samples a subset of the K stopping points, where each stopping point is regarded as a stratum (the clusters). If the kth stopping point is sampled as one of the strata, then the n_k accessible houses around that stopping point form part of the sample S. cluster sampling in this research is performed by first considering all N_a accessible houses in each of the K strata as being sampled, at which point $n = N_a$. One stratum is then removed at random, and the sample S then consists of all houses excluding those in the kth stratum that was removed. The proportion $p = \frac{n}{N_a}$ is then recorded, and another stratum is removed, resulting in a smaller sample. This process is continued until $p = 0.7$. Setting p exactly equal to 0.7 is rarely possible while randomly removing strata, and for this reason, the removal process is stopped one iteration before $p < 0.7$, at which point p is some value slightly above 0.7. In order to set the sample size n exactly equal to 70% of the available houses N_a, the difference between n and $0.7 \times N_a$ is calculated, and points equal to this difference is randomly removed from the sample S. This ensures that the number of points n in S is 70% of N_a at every iteration of the TCS scheme. This adjustment allows for an accurate comparison between the TCS scheme and the results of other sampling schemes that are able to sample 70% of N_a without any adjustment. This process is delineated in Algorithm 2, and may be accessed using this link.[1]

2.2.3. Uniform spatial sampling (USS)

The uniform spatial sampling (USS) scheme performs sampling by first generating p random points on the spatial window W under question. Since we are working in $\mathbb{R}^2$, each random coordinate on W is generated by a set of two random uniform values within the bounds of the window W. The generation of such points are achieved by using the `runifpoint()` function from the `spatstat` library (Baddeley and Turner, 2005) in R. Each set of uniformly generated points in W constitute a randomly placed polar coordinate. The house closest in terms of Euclidean distance to this point is sampled.[2] To ensure that a house is not sampled more than once, the sampled house at the ith iteration is removed from W before sampling the next house at iteration $i + 1$. The number of uniform spatial points generated in W is set to be equal to 70% of the accessible houses N_a.

[1] The reason why this sampling scheme is described in an algorithm is because no R package exists that can be used to directly apply the TCS sampling scheme as described in this research. The algorithm at this link is therefore meant to clarify this sampling scheme for the reader, making it easier to replicate it.

[2] A fast and slow animation of the uniform spatial sampling scheme applied to the Machochwe village may be viewed here.

2.2.4. Spatial stratified sampling (SSS)

The spatial stratified sampling (SSS) scheme considers each stopping point to be the centre of a stratum, and all the accessible houses around each stratum are available for sampling. This is a similar approach to the traditional stratified samplings (TCS) scheme. What is different, however, is that the SSS scheme samples 70% of the houses over every stratum. The TCS scheme on the other hand only selects some of the strata, and considers all of the houses accessible around each strata as part of the sample. The SSS scheme generates n_k random points within a disc of radius 200 metres around the kth stopping point, and then samples the n_k houses which are closest to the n_k randomly generated points. In order to achieve the desired household coverage of 70% of N_a, the number of sampling points n_k in the kth stratum is calculated to be 70% of the houses around the kth stopping point. The SSS algorithm[3] is delineated in Algorithm 3 for further clarity.

2.2.5. Systematic regular spatial sampling (SRSS)

For a sample size of n, the window W is broken into q equally sized squares and a point is generated in the centre of each square. This results in a regular grid of points on W. The `rsyst` function from the `spatstat` package is used to generate such a regular point pattern in W. The house closest to the ith generated point is then sampled, and removed to prevent that point from being sampled again. The `rsyst` function, however, does not generate the same number of sampling points each iteration.

A form of manipulation to the generated pattern is therefore required in order to ensure that exactly 70% of the available houses N_a are being sampled for every iteration. This manipulation is similar to what is done for the traditional stratified sampling scheme in Section 2.2.2, and for the spatial stratified sampling scheme from the previous section. If the number of points p in the systematic regular spatial point pattern is greater than $0.7 \times N_a$, the points equal to the difference between p and $0.7 \times N_a$ is randomly removed from the sampling pattern, ensuring that $n = 0.7 \times N_a$.

2.2.6. Systematic non-aligned spatial sampling (SnaSS)

This fourth spatial sampling scheme follows a similar approach to the previous SRSS scheme. That is, the window W is also partitioned by a grid of q squares, however instead of generating points in the centre of each square, points are generated in a random location within each square. The resulting set of generated points are therefore still regularly spread out across W, but with some more variation as opposed to the SRSS scheme.

The `spsample` function from the `spatstat` R package (Baddeley and Turner, 2005) is used to sample $0.7 \times N_a$ points in W. To sample these spatial locations in a non-aligned fashion, the option `type = 'nonaligned'` is specified for the `spsample` function. Despite being able to specify exactly how many points should be sampled in the window W, the actual number of realised points often differs from the goal of $0.7 \times N_a$. The point pattern manipulation approach used for the SRSS scheme is therefore also used here to ensure that exactly 70% of N_a is sampled.

2.2.7. Systematic hexagonal spatial sampling (SHSS)

This final sampling scheme partitions the window W with a grid of h hexagons instead of squares, and generates a point at the centre of each hexagon. Similar to the previous four spatial sampling schemes, a house is sampled if it is the closest house to one of the generated points. To generate points for the SHSS scheme, the `spsample` function from the `spatstat` (Baddeley and Turner, 2005) is used again, but here with the option `type = 'hexagonal'` specified. The SHSS scheme also requires manipulation of its samples, when the sample size n is in excess of the $0.7 \times N_a$ goal. Any iterations of the SHSS scheme resulting in a sample size less than the desired goal is rejected, and subsequent iterations are run until n is equal to, or greater than $0.7 \times N_a$.

2.2.8. Kernel-weighted spatial sampling scheme

A new sampling scheme is introduced and applied here. This sampling scheme is named the kernel-weighted spatial sampling (KWSS) scheme. This probability-based spatial scheme directly exploits the spatial distribution of each rural village in order to obtain a more representative sample of houses. As a probability-based sampling scheme, the KWSS scheme assigns a probability of being sampled to each spatial location in the window W. The KWSS scheme is therefore partial towards certain spatial locations, and will generate more samples in certain areas than in others.[4] Spatial locations in W with a higher density of houses is preferred over lower-density regions.

The sampling mechanics of this newly proposed scheme works in a similar fashion to the five previously discussed spatial sampling schemes. That is, after generating a set of spatial sampling points on W, the house nearest to each spatial sampling point is chosen to form part of the sample of houses to be visited by a vaccinator. The manner in which these sampling points are generated, however, is different from the previous five schemes, and no function within the `spatstat` package (Baddeley and Turner, 2005) exists to generate a set of points for this scheme. Before generating the set of sampling points, a Gaussian kernel density estimate $\hat{\lambda}(u)$ of the village point pattern under consideration is computed for each spatial location u. This density estimate is then used as a map to dictate the likelihood of a spatial sampling point being generated at a given point in space. Dense areas with a larger $\hat{\lambda}(u)$ receive a higher likelihood of being sampled than areas where housing is sparse. This results in more houses being sampled in regions where the point pattern is more dense.

[3] This algorithm may be viewed using this link.

[4] A comparison between animations of the uniform spatial sampling (USS) scheme and the KWSS scheme here shows how the KWSS scheme is able to secure a much more representative sample from the high-density areas of Machochwe than the USS scheme.

The intensity $\lambda(u)$ may be estimated using either corrected or uncorrected kernel estimators (Baddeley et al., 2015), where the corrected estimator adjusts for edge effects. An uncorrected kernel estimator is used here, and the edge effects are accounted for by adding a buffer to the spatial window W. This buffer extends the border of W with 4 times the bandwidth value h, or, four standard deviations.

The uncorrected kernel estimator used for the KWSS scheme is given by

$$\hat{\lambda}(u) = \sum_{i=1}^{n} \kappa_h(u - x_i), \tag{1}$$

where $\kappa_h(u) = h^{-1}\kappa(h^{-1}u)$ denotes some probability density function used as a kernel to estimate $\hat{\lambda}(u)$ (Diggle, 1985). The kernel used in this research is the isotropic Gaussian kernel, resulting in

$$\hat{\lambda}(u) = \frac{1}{\sqrt{2\pi h^2}} \sum_{i=1}^{n} \exp\left\{ -\frac{1}{2}\left(\frac{(u - x_i)^2}{h} \right) \right\}. \tag{2}$$

This kernel is used because it is simply a function of distance and does not account for the direction as well (Baddeley et al., 2015). This is ideal for the current use case, as the direction is not of importance. The parameter h from Eq. (2) is the standard deviation, or scale parameter of the Gaussian kernel, which is referred to as the bandwidth of the kernel (Baddeley et al., 2015). Smaller values of the bandwidth h result in a coarser estimate of $\lambda(u)$ while large value of h yield a smoother estimate. In this research, an h parameter of 50 is used, as it yielded results better able to attain the 70% (see Botes (2023) for details). The function `density.ppp` from the package `spatstat` (Baddeley and Turner, 2005) is used to compute estimates for $\hat{\lambda}(u)$, and the probability p_i of generating a sampling point at the spatial location u_i is equal to

$$p_i = \frac{\hat{\lambda}(u_i)}{\sum_{i=1}^{n} \hat{\lambda}(u_i)}. \tag{3}$$

The kernel weighted spatial sampling scheme is summarised in Algorithm 4, which may be viewed using this link.

2.3. Algorithms to calculate walking distance

Different methods can be used to compute the route followed by a vaccinator from the kth stopping point to each sampled house n_k. The method employed by Fabris-Rotelli et al. in Fabris-Rotelli et al. (2020) is a minimum spanning tree (MST), as it is less copmputationally expensive than a TST algorithm. In this research, the travelling salesman tour (TST) (Skiena, 1990) method is used. The reason why the MST algorithm is not used, is because it generates a tree. It is not practical for a vaccinator to walk on a path generated by a tree, since the vaccinator would need to traverse each edge of the MST twice to visit each house at least once and then return back to the stopping point. This would result in unnecessary walking distance. It is therefore more intuitive and optimal to treat this problem as a travelling salesman problem. When constructing a TST from an MST graph, it can be shown that the worst case solution for a TST is the MST (Laporte, 1992). Most heuristic algorithms will always provide a solution that is less than two times the total distance along an MST graph (Michael and Kurt, 2007).

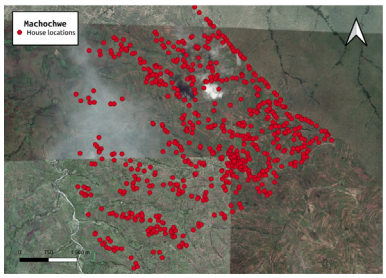
Several algorithms to aproximate the TST problem exist, such as the Lin–Kernighan (Lin and Kernighan, 1973) heuristic algorithm, and variation of the nearest neighbour and insertion algorithms (Rosenkrantz et al., 1977). The algorithm used in this research is the farthest insertion algorithm (Rosenkrantz et al., 1977), which is implemented using the TSP package (Michael and Kurt, 2022) in R. An in-depth simulation comparison between eight different TST algorithms are drawn in Botes (2023), and the farthest insertion algorithm was found to yield the fastest computation time, given its accuracy.

3. Results

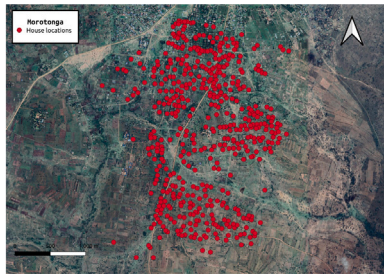
3.1. The data

The Tanzanian village dataset[5] was compiled from a census of 90 rural Tanzanian villages conducted between August 2014 and October 2016. Relevant fields in the dataset include the name of the village where each house is situated, the GPS coordinates of each house, and the number of dogs above and below 3 months of age. The dataset also makes a distinction between vaccinated and unvaccinated dogs. In this application, however, all dogs are considered to be unvaccinated. Whether or not vaccinated dogs are included in results will not affect the findings in terms of which TST algorithm, stopping point algorithm and sampling scheme is the most optimal. If the methods of this research were to be applied to design a rabies door-to-door vaccination schedule based from recent census data, all dogs that have already been vaccinated should be excluded before sampling the houses. While the dataset distinguishes between puppies (those dogs in the dataset which are younger than 3 months of age) and adults dogs, these two fields were aggregated and no distinction is drawn between dogs of different ages. It must be noted that this dataset also includes houses that do not own any dogs. These houses were not considered when calculating the optimised rabies vaccination schedule.

[5] The use of this dataset was approved by the Faculty of Natural and Agricultural Science Research Ethics committee at the University of Pretoria under the reference NAS339/2019. The data used in the paper was obtained from Katie Hampson (https://www.gla.ac.uk/schools/bohvm/staff/katiehampson/).



(a) Satellite image of Machochwe with house locations. [7]



(b) Satellite image of Morotonga with house locations. [8]

Fig. 1. Satellite images of Machochwe and Morotonga.

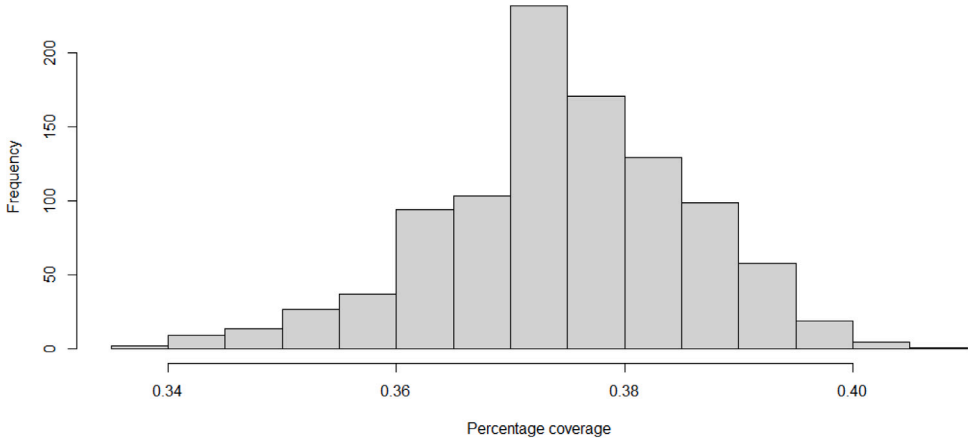


Fig. 2. The distribution of coverage obtained from 1000 instances of k-means with $k = 50$, and random initial values, in Machochwe.

From the 90 villages in the census dataset, two villages are chosen for the purpose of this research. The village of Machochwe and the village of Morotonga. These two villages are displayed in Fig. 1. Morotonga is a smaller village, with only 246 dog-owning households, while Machochwe is a relatively large village, containing 632 dog-owning households. The houses in Morotonga are also more regularly spread out over space, while the village of Machochwe is more clustered. On average there is more than one dog at each house in the villages (Fabris-Rotelli et al., 2020).

3.2. k-means parameter tuning

Finding the k which achieves the maximum coverage with the smallest number of stopping points is achieved by means of a simulation study. This simulation study comprises of calculating the minimum number of stopping points required to reach different coverage goals. The different coverage goals are $30\%, 40\%, 50\%, 60\%$ and 70%. Although the aim is of course 70%, the simulation study investigates also lower values to understand the sampling algorithms better. Calculating the optimal k for different coverages aids in determining with greater certainty which of the sampling schemes are the most optimal in minimising walking distance, and maximising vaccination coverage. It would, for example be interesting to verify if a sampling scheme which is optimal at 70% coverage, is also optimal at 30%, 40%, 50% and 60% coverage. Since each sampling scheme strictly samples 70% of N_a, the optimal number of stopping points for each coverage is the same for each of the eight sampling schemes.

To optimise the parameter k for each of the five coverage thresholds, the lowest possible k yielding each of the five thresholds must to be found. Table 1 shows the minimum value of k necessary in order to reach each of the five coverage thresholds for Morotonga and Machochwe respectively. Finding these values of k is done using an iterative approach. Fig. 2 shows the distribution of sampling coverages possible when running k-means 1000 times with different initial values at each iteration. After each iteration, 70% of the attainable houses around the 50 stopping points are sampled, and the resulting coverage percentage is noted. Fig. 2 illustrates that this coverage can vary between a 34% and 40% in the village of Machochwe. The approach to optimise k therefore requires several iterations of k-means to be run for different values of k, to determine which k is the minimum required to reach a desired coverage threshold at least once.

To illustrate the iterative approach used to optimise the k in k-means, say that one would like to minimise (optimise) the value of k needed to reach the 30% coverage threshold in the village of Morotonga. The process begins by selecting a reasonable value

Table 1
The minimum k-means value of k necessary to reach each coverage goal in Morotonga and Machochwe. The probability of reaching a coverage with some k, and the random seed for the k-means starting points are also tabulated.

Minimum k for Morotonga				
Coverage goal	k	Coverage	Probability	Random seed
30%	11	30.08%	5.7%	4
40%	15	40.24%	0.6%	26
50%	20	50.00%	7.3%	15
60%	26	60.16%	0.1%	697
70%	50	69.92%	0.2%	326
Minimum k for Machochwe				
Coverage goal	k	Coverage	Probability	Random seed
30%	36	30.22%	1.8%	193
40%	50	40.03%	0.5%	999
50%	68	50.00%	0.7%	174
60%	94	60.28%	0.2%	84
70%	205	69.94%	0.1%	13

of k, which we will set as 8. Next, 1000 iterations of k-means is executed. A random seed of i is set using the `set.seed` function in R (R Core Team, 2022) at each iteration, where i corresponds to the iteration number, $i = 1, 2, \ldots, 1000$. Doing so allows us to fix the random starting points of the ith k-means iteration with a seed of i, making the results reproducible. If, after 1000 iterations, no stopping point solution allowed the SRS scheme to attain to the sampling threshold of 30%, another 1000 iterations of k-means is executed, using a $k = 9$. This process is repeated j times for $j = 0, 1, 2, \ldots$ until the parameter $k + j$ is found for which the SRS scheme can reach the 30% coverage threshold at least once over the 1000 iterations. The optimal k in this example is the first k for which $\frac{0.7 \times N_a}{N}$ is greater than 0.3 for at least one of the 1000 iterations. In Table 1, the value of $j = 3$, since three different k values were tested before coverage of at least 30% was reached.

The column titled 'Probability' in Table 1 indicates that 57 of the 1000 k-means iterations achieved coverage of at least 30%, hence the probability of reaching the 30% coverage threshold with $k = 11$ is 5.7%. Note that while the column is titled 'Probability', the probabilities in this column should be seen as estimates of the true, unknown probability. Ascertaining a more accurate probability requires larger simulation studies to be performed. The probabilities displayed in Table 1 are however deemed accurate enough for the purposes of this research. The 'Random seed' column from Table 1 indicates that a random seed of 4 may be used to generate a k-means stopping point solution with $k = 11$, that will also result in a sampling coverage of 30.08%. This random seed of 4 is used in the upcoming Section 3.3 to fix the sampling coverage when calculating the walking distance distributions for each sampling scheme. It is not desirable for each iteration of k-means to produce a different household coverage. Walking distance is undeniably a positive function of household coverage, and it is therefore important to keep household coverage fixed at some threshold when comparing the walking distance between sampled houses for each sampling scheme.

Drawing a comparison between the SRS scheme which, for arguments sake, could require 13 km walking distance to reach 40% of houses and the USS scheme requiring 15 km to reach 46% is no comparison at all, since the coverage percentages vary. It is therefore necessary to fix the random initial values of the k-means algorithm to ensure that the value of N_a remains constant for each iteration of the k-means algorithm. It was found, for example, that setting a random seed at 26 yields a k-means solution where 40.24% of houses in Morotonga are accessible for sampling. By setting this determined seed value seed at each iteration, it is possible to compare the performance of the eight sampling schemes with a constant house coverage.

Having optimised the parameter k for the k-means algorithm in the previous section, this section performs a simulation study to determine which of the eight sampling schemes is the most optimal for a door-to-door vaccination schedule. That is, which sampling scheme is able to sample 70% of the attainable houses provided by the stopping point algorithms, such that the walking distance between each stopping point and the sampled houses is a minimum. The same five coverage thresholds from the previous sections are used to ascertain which sampling scheme is most optimal for the 30%, 40%, 50%, 60% and 70% coverage thresholds. Remember that each of the sampling schemes introduced in Section 2.2 is set up to sample 70% of the attainable houses N_a. By varying the value of k for a stopping point algorithm (and therefore changing the value of N_a), it is possible to increase the sample size n, as a proportion of the total number of houses N.

3.3. Walking distance distributions

To illustrate the method used to generate a walking distance distribution, consider the following example. To calculate the distance that a vaccinator needs to walk to vaccinate 30% of the dog-owning households in Morotonga, the k-means algorithm with a k of 11 must first be executed with a random seed of 4. By sampling 70% of the attainable houses around these stopping points, the sample size n is exactly[6] 30.08% of the total number of houses N, in Morotonga. Since every sampling schemes randomly samples

[6] See Table 1.

Table 2
Descriptive statistics on the 10 000 instances of each sampling scheme executed for the five different coverage thresholds in the village of Morotonga, after using the k-means algorithm to select stopping points. The lowest statistics for each row are highlighted in **bold**, and the highest in *italics*.

Walking distance statistics per sampling scheme: Morotonga K-means								
30% coverage								
	SRS	SHSS	SnaSS	SSS	SRSS	TCS	USS	KWSS (50)
Mean	9.251	*9.9352*	9.8867	9.289	9.8583	**7.7153**	9.8508	8.8559
Median	9.2604	*9.9454*	9.8882	9.2953	9.8555	**7.7852**	9.8583	8.8658
Std. Dev.	0.3048	**0.1293**	0.1463	0.2639	0.154	*0.4305*	0.2205	0.3689
Min.	8.0519	*9.3719*	9.323	8.3329	9.1202	**6.6639**	8.9965	7.4288
Max.	10.1954	10.2984	10.363	10.1685	10.2566	**8.4373**	*10.5657*	10.1541
40% coverage								
	SRS	SHSS	SnaSS	SSS	SRSS	TCS	USS	KWSS (50)
Mean	12.687	*13.1961*	13.1024	12.4969	13.0704	**10.6235**	12.9317	12.2532
Median	12.6995	*13.1855*	13.1054	12.5039	13.0937	**10.634**	12.9383	12.2618
Std. Dev.	0.3483	0.2423	0.2337	0.321	**0.2323**	*0.4376*	0.3275	0.3973
Min.	10.9004	*12.2121*	12.1474	11.2119	12.1129	**9.321**	11.5781	10.6746
Max.	13.8769	14.0131	13.9253	13.6573	13.7562	**11.6625**	*14.0569*	13.5412
50% coverage								
	SRS	SHSS	SnaSS	SSS	SRSS	TCS	USS	KWSS (50)
Mean	15.269	*16.1874*	16.0916	14.5923	16.0298	**12.8667**	15.6523	14.6833
Median	15.2763	*16.2002*	16.0967	14.5962	16.038	**12.8749**	15.6568	14.6894
Std. Dev.	0.4117	0.2555	**0.2497**	0.3666	0.2686	0.4359	0.3687	*0.4647*
Min.	13.9033	*15.246*	15.1539	13.2649	15.0655	**11.3018**	14.3006	12.7787
Max.	16.7057	16.9079	*16.9875*	15.8401	16.9839	**14.0726**	16.8071	16.2195
60% coverage								
	SRS	SHSS	SnaSS	SSS	SRSS	TCS	USS	KWSS (50)
Mean	19.1808	*20.138*	20.0353	18.1423	20.0623	**16.3266**	19.3408	18.5249
Median	19.185	*20.1544*	20.0454	18.1496	20.0668	**16.3325**	19.3397	18.5323
Std. Dev.	0.4558	**0.2891**	0.3243	0.3813	0.3034	0.5004	0.3954	*0.5132*
Min.	17.5815	*19.0627*	18.6145	16.8456	18.7766	**14.7873**	17.6704	16.8027
Max.	20.762	21.048	21.1422	19.6069	*21.2276*	**17.8668**	20.7982	20.2418
70% coverage								
	SRS	SHSS	SnaSS	SSS	SRSS	TCS	USS	KWSS (50)
Mean	23.8847	*25.3028*	25.0343	22.3611	25.2285	**21.2471**	24.2909	23.3259
Median	23.8989	*25.2949*	25.0338	22.3614	25.2314	**21.2544**	24.2933	23.3289
Std. Dev.	0.6228	**0.3791**	0.4253	0.5486	0.3839	0.5744	0.5117	*0.6478*
Min.	21.5485	*24.0229*	23.384	20.088	23.7605	**19.1815**	22.2524	20.902
Max.	26.2086	*26.7306*	26.5497	24.3045	26.6738	**23.1013**	26.5843	26.0154

points, not every sample consists of the same houses. As such, the route a vaccinator should walk to visit the sampled houses is different every iteration. By executing a sampling scheme several times on the same set of attainable houses N_a, it is possible to generate a distribution for the walking distance required to visit 30% of the houses for each sampling scheme. The method used to generate a walking distribution is delineated earlier. The nine step process in this list is implemented for k-means, for every coverage threshold, and for each sampling scheme in both the Morotonga and Machochwe villages.

3.4. Comparison of the sampling schemes

The descriptive statistics from Tables 2 and 3 are used to determine the optimal sampling scheme for the Morotonga and Machochwe villages for each of the five coverage thresholds. The most important descriptive statistic is the minimum walking distance for each threshold. Considering these results, the TCS algorithm is the best in both the Morotonga and Machochwe villages, since only 19.1815 and 44.9074 kilometres has to be walked by a vaccinator to vaccinate 70% of the dog-owning houses in Morotonga and Machochwe respectively. The standard deviation information is also displayed, however the most important statistic considered is the minimum walking distance. Even though a sampling scheme may have a high mean and a high standard deviation, it could still be the best sampling scheme if it has a minimum walking distance lower than the minima of all the other sampling schemes. Fig. 3a and b plots the walking distributions for the eight sampling schemes, to reach a 70% vaccination coverage in Morotonga and Machochwe respectively. Note that this is 70% of all houses in these villages, as enough stopping points are placed such that Na=N. Plots for the 30%, 40%, 50% and 60% coverages can be viewed in Botes (2023).

Fabris-Rotelli et al. (2020) found the systematic spatial regular sampling (SRSS) scheme to be optimal, while in this research, the SRSS scheme ranged between being the 6th, 7th and 8th best sampling scheme depending on the coverage threshold and village used. The difference between the walking distance distributions for the sampling schemes in this research is also much more spread

Table 3
Descriptive statistics on the 10 000 instances of each sampling scheme executed for the five different coverage thresholds in the village of Machochwe, after using the k-means algorithm to select stopping points. The lowest statistics for each row are highlighted in **bold**, and the highest in *italics*.

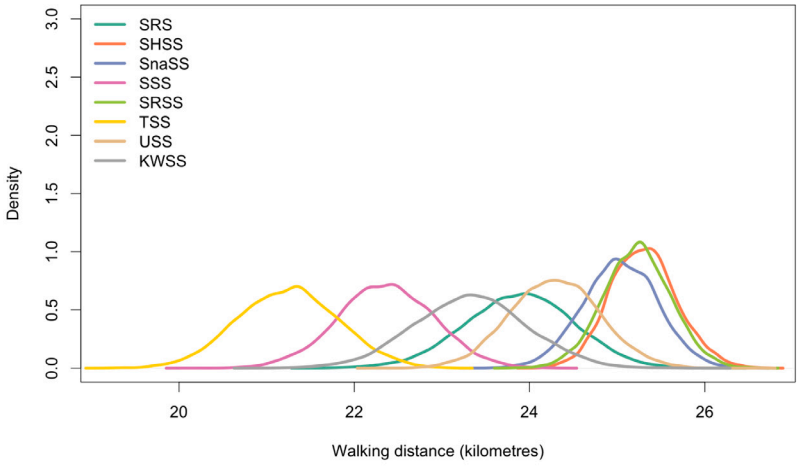
Walking distance statistics per sampling scheme: Machochwe K-means								
30% coverage								
	SRS	SHSS	SnaSS	SSS	SRSS	TCS	USS	KWSS (50)
Mean	22.2486	*23.3703*	23.2253	20.9042	23.2757	**18.5524**	22.1089	21.6728
Median	22.2652	*23.3754*	23.2305	20.909	23.2871	**18.4693**	22.1236	21.6902
Std. Dev.	0.5606	**0.2806**	0.3093	0.4196	0.3194	*1.162*	0.5368	0.6323
Min.	19.8266	*22.3295*	21.8627	19.3224	21.9035	**15.7794**	19.8318	19.1805
Max.	24.1266	24.1377	*24.2411*	22.4782	24.1623	**22.3473**	24.1958	23.7563
40% coverage								
	SRS	SHSS	SnaSS	SSS	SRSS	TCS	USS	KWSS (50)
Mean	32.1471	*34.8445*	34.7432	30.7569	34.8392	**26.9042**	33.5445	31.5313
Median	32.1511	*34.8696*	34.7575	30.7616	34.8502	**26.3498**	33.559	31.5241
Std. Dev.	0.6931	0.3129	0.3197	0.5398	**0.3096**	*1.7699*	0.5357	0.7583
Min.	29.3331	*33.642*	33.3204	28.8152	33.1873	**22.8386**	31.6467	28.6559
Max.	34.5431	35.6196	*35.9277*	32.8183	35.7636	**32.0994**	35.3104	34.1056
50% coverage								
	SRS	SHSS	SnaSS	SSS	SRSS	TCS	USS	KWSS (50)
Mean	40.0961	*42.1906*	42.079	38.0369	42.0806	**34.2486**	41.0959	39.3462
Median	40.1056	*42.2365*	42.082	38.0369	42.101	**34.1781**	41.1024	39.3585
Std. Dev.	0.8056	0.5579	**0.4716**	0.7179	0.506	*1.4524*	0.7058	*0.8625*
Min.	37.0149	*40.4612*	40.0256	35.1633	40.0765	**30.1026**	38.3048	36.0929
Max.	43.0429	43.4334	*43.8248*	40.8769	43.5177	**39.3013**	43.6167	42.4446
60% coverage								
	SRS	SHSS	SnaSS	SSS	SRSS	TCS	USS	KWSS (50)
Mean	50.0226	53.365	53.3758	47.096	*53.4469*	**43.27**	52.4195	49.2092
Median	50.0281	53.3992	53.3749	47.0896	*53.4422*	**42.8202**	52.4236	49.2134
Std. Dev.	0.9057	**0.4901**	0.5343	0.7909	0.5928	*1.823*	0.709	0.9658
Min.	46.5855	51.259	51.1912	44.0944	*51.3641*	**38.914**	49.5094	44.8886
Max.	53.2552	54.9132	55.2321	50.3539	*55.5442*	**49.5117**	55.0782	52.6159
70% coverage								
	SRS	SHSS	SnaSS	SSS	SRSS	TCS	USS	KWSS (50)
Mean	54.9476	58.5623	58.6836	52.2732	*58.7214*	**49.7439**	56.8576	54.9477
Median	54.9385	58.5849	58.6877	52.2722	*58.7692*	**49.6859**	56.8491	54.943
Std. Dev.	1.055	**0.5709**	0.6561	0.8993	0.6861	*1.3319*	0.8297	1.0541
Min.	49.9622	*56.4116*	56.2828	48.9114	56.086	**44.9074**	53.9622	50.845
Max.	59.1519	60.4399	*61.15*	55.6466	60.7331	**54.1996**	60.3476	58.7396

out than in the work of Fabris-Rotelli et al. For example, we see in Table 3 that there is nearly a 10 km difference between the average walking distance of the TCS and SnaSS schemes in the Machochwe village. Key differences between the methodology used by Fabris-Rotelli et al. (2020) have already been highlighted in Section 2.3. While Fabris-Rotelli et al. used a minimum spanning tree (MST) to calculate the route a door-to-door vaccinator needs to walk between houses, this research uses a travelling salesman tour (TST). This was shown to result in at least half the walking distance that an MST would. This code may be used to replicate the simulation study, using the random seeds provided.

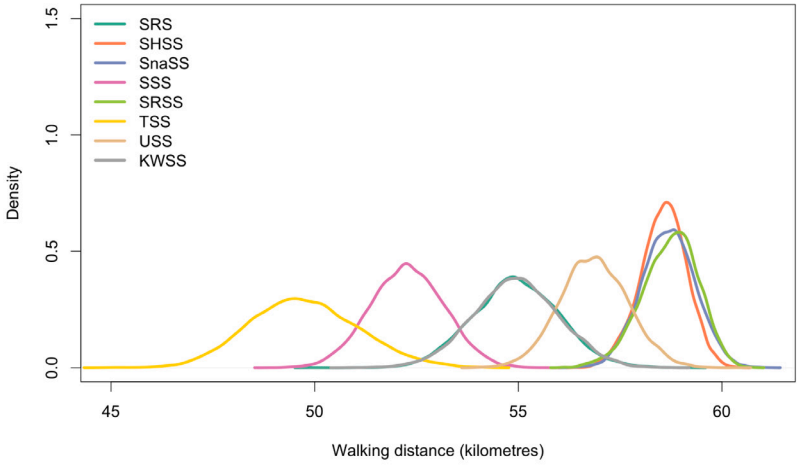
4. Discussion and future work

While this research has successfully optimised a vaccination schedule by comparing eight samplings schemes in two villages, some limitations of this research should be highlighted. The first limitation is that only two villages were used in this simulation study. One of the villages was clustered, and the other regular; one was large, and the other was small (link). Nevertheless, it is recommended to test the methods from this research on more villages of varying sizes and second-order natures. It is possible that a relationship exists between the optimal sampling scheme and the second-order nature of a point pattern that is not discernible from the limited results of this research. It would also be a good idea to compare the k-means algorithm to other clustering algorithms for the selection of stopping points. Some candidates could be more hierarchical-based clustering algorithms or even clustering from a Gaussian Mixture model approach. However, k-means is simple and computationally easy thus providing a good solution.

The distance walked by a vaccinator between houses is assumed to be the shortest Euclidean distance. A more accurate walking distance can be determined by calculating the distance walked along actual roads that may exist between house. It will not always be possible for a vaccinator to walk the shorted Euclidean route. It would also be interesting to take into account the topography of the landscape along which a vaccinator needs to walk when calculating distance, and determining the effort exerted by a vaccinator.



(a) Walking distance distributions in Morotonga



(b) Walking distance distributions in Machochwe

Fig. 3. Walking distance bootstrap distributions for each sampling scheme with k-means stopping points. Only results for the 70% coverage threshold are shown here.

Another point to mention is that the vaccination of 70% of household pets does not guarantee that 70% of the dog population in a village is vaccinated. The methods described in this research should therefore, ideally, be implemented together with a plan to vaccinate 70% of the free-roaming dogs in a village in order to truly attain herd immunity. Future research may look at how to best perform vaccination on free-roaming dogs. The study on dog demographics in Tanzania from Knobel et al. (2008) mentioned in Section 4 also show that it will not be possible to vaccinate 70% of household dogs by visiting 70% of the houses in a village, since around 30% of these dogs can be expected to be away during the day. Depending on the true number of home-bound dogs in a village, alternative means will need to be employed in order to reach dogs that are not at home during the day. An important point that is not discussed in this research is the number of annual vaccination campaigns that should be performed in order for herd immunity to be attained. In order to gauge this, the methods from this research will first need to be applied for multiple years. By monitoring the rabies vaccination over these years, it would be possible to say for certain how many annual campaigns are required to attain true herd immunity.

Another study that may be valuable to consider is to vary the radius of attainable houses around each stopping point. In this research, only houses within a 200 metre radius of each stopping point is considered to be attainable for a vaccinator. An increase in the accessible houses radius would allow for a decrease in the value of k, however it may also result in more kilometres walked,

and less driven. This is something that should be further explored, however, to find the optimal radius. A more accurate approach could also determine radius based on walking distance on roads and paths, namely not the Euclidean distance. It would also be interesting to see what difference it would make, were one to first sample and then select stopping points based on the sampled houses.

The traditional stratified sampling scheme proved to be optimal among the eight schemes analysed in this research. Using this scheme consistently resulted in the lowest average and minimum walking distance. Even though a traditional sampling scheme outperforms the spatial schemes considered, this is an indication that the spatial information of house locations is taken into account with the stopping point determination. This indicates that the use of TSP caters for the spatial autocorrelation present. In further research, more focus should be placed on stratified sampling schemes, to see if it is possible to further minimise the walking distance for a door-to-door vaccinator. It may also be valuable to measure spatially stratified heterogeneity in this research context, and to understand how knowledge of this can assist in further optimising the simulation framework.

It may also be valuable to combine a static vaccination approach with a door-to-door vaccination scheme. A simulation study of such a hybrid approach would also be useful for policy makers in setting up effect rabies vaccination campaigns. Static vaccination stations are most frequently used when vaccinating dogs in rural settlements, as it requires less effort, and is therefore more cost effective. It may be unreasonable to expect policy makers to adopt a vaccination schedule that is completely built upon a door-to-door vaccination approach. However, after having performed this research, a wholly door-to-door approach may be more attractive for policy makers, since a larger body of literature exists proving that such an approach can be optimised and the reach can be measured, and performed in as little as a one or two days, depending on the size of the rural village under consideration. Lastly, future research efforts could go into developing a model based sampling technique, wherein it would be possible to quantify uncertainty regarding whether or not a specific vaccination coverage target will be reached, or not, given dog demographics and other census data, where and when it is available.

5. Conclusion

This research set out to develop an optimised door-to-door rabies vaccination schedule in rural villages. Two rural Tanzanian villages are considered in an application study to compare eight different sampling schemes, to see which one provides the most optimal sample of houses to visit. Optimality throughout this research is measured in terms of walking distance. Sampling schemes resulting in a sample requiring shorter walking distances to cover is preferred to sampling schemes requiring longer walking distances. It is found that the traditional stratified sampling scheme provides the most optimal results when taking into account spatial autocorrelation in the stopping point selection. The k-means algorithm is used to place stopping points throughout a village, from which a vaccinator performs vaccination. This research provides policy makers with a further tool to combat the spread of rabies in rural villages. This research furthermore provides a framework wherein any number of sampling schemes can be compared on any given village. Having optimised this framework, and providing the code on GitHub, any researcher planner of a vaccination campaign is able to input any village, and any alternative sampling scheme, and determine the optimal stopping point configuration and minimum walking distance to reach 70% of the total dog population. This research could therefore ultimately help in construction and planning of future rabies vaccination campaigns in Tanzania, and other endemic regions. With the year 2030 drawing nearer, and with it the WHO goal of zero human rabies deaths, much work is still required. This research, together with previous and other current efforts in the domain of rabies vaccination schedules will go far in assisting decision makers to apply the most time optimal and cost effective methods to combat this deadly disease.

Acknowledgements

This work is based on research supported in part by the National Research Foundation of South Africa (South Africa DST-NRF-SAMRC SARChI Research Chair in Biostatistics, Grant Number 114613 and Grant Number 137785). Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

Appendix. Algorithms

See Algorithms 1, 2, 3, 4 and 5.

Algorithm 1 The k-means algorithm for selecting stopping points.

Input: List of locations of all accessible houses N_a, a value for k, and a random seed to select the initial k-mean values.

Assignment step

· Assign each accessible house to its nearest k-means initial value. Houses closest to the k^{th} stopping point are said to belong to that stopping point.

Update step

· Recalculate the value of each stopping point as the mean of all the houses for which it is responsible after the previous assignment step.

Convergence

· Convergence is reached once the assignments remain unchanged.

Algorithm 2 Traditional stratified sampling algorithm.

Input: List of the location of all accessible houses N_a together with the stopping point (strata) index k that each house belongs to.

· Set $\tilde{p} = 0.7$.
· Initiate a list `strata` with integers 1 to K.
· Set p equal to 1.
while $p \geq \tilde{p}$ **do**
  Randomly sample an integer from `strata`.
  Remove all houses from the stratum corresponding the sampled integer.
  Recompute the value of p
  Remove the sampled integer from `strata` such that it is not sampled again.
end while
· If $p > \tilde{p}$, calculate the difference between n and the required sample size, $0.7 \times N_a$.
· Randomly remove $n - (0.7 \times N_a)$ points from n, such that $n = 0.7 \times N_a$.

Algorithm 3 Spatial stratified sampling algorithm.

· `redo = True`
· Determine the required sample size as $0.7 \times N_a$
while `redo == True` **do**
  · Set $p = 0.7$
  · Sample $(100 \times p)$% of the houses in each of the K strata
  · Round the sample size of each strata to the nearest integer, n_k
  · Calculate the difference between $\sum_{k=1}^{K} n_k$ and $0.7 \times N_a$
  if $\sum_{k=1}^{K} n_k > (0.7 \times N_a)$ **then**
    · Randomly remove sampled points equal to the difference between $\sum_{k=1}^{K} n_k$ and $0.7 \times N_a$
    · Set `redo = False`
  else if $\sum_{k=1}^{K} n_k < (0.7 \times N_a)$ **then**
    · $p = p + 0.01$
    · Continue with `while` loop.
  else
    · `redo = False`
  end if
end while

Algorithm 4 Kernel weighted spatial sampling scheme.

Input: A value for the bandwidth h and a village point pattern $\boldsymbol{x}$ as well as the number of houses to sample, $n = 0.7 \times N_a$.

Perform the following three initial steps:

1. Perform edge correction by extending the spatial window W with a buffer, such that the border of W is expanded by $4 \times h$.

2. Use the `density.ppp` function from the `spatstat` package to calculate a kernel density estimate of a 128×128 grid across the window W and the buffer, resulting 16 384 kernel density estimates. Some points in this square grid lie outside the convex hull window W and the buffer, and are not assigned a density estimate.

3. Use probability sampling to generate n spatial sampling points in W. Denote the i^{th} spatial sampling point as n_i, and the vector of all sampling points as $\boldsymbol{n}$.

for n_i in $\boldsymbol{n}$ **do**

- Determine which house in the point pattern $\boldsymbol{x}$ is nearest to n_i in terms of Euclidean distance.
- Sample the house nearest to n_i.
- Remove the nearest house to avoid sampling the same house more than once.

end for

Algorithm 5 Step by step breakdown of generating a walking distance distribution.

1: Select a stopping point algorithm.
2: Select a sampling scheme.
3: Select a TST algorithm (The farthest insertion heuristic is recommended).
4: Select a sampling coverage (30%, 40%, 50%, 60% or 70%).
5: Find the value of k needed to generate the right stopping point configuration for the desired coverage.
6: Generate the k stopping points using your selected stopping point algorithm (remember to initiate k-means with the right random seed from table 5 for Morotonga or table 6 for Machochwe from the article)
7: Generate 10000 random samples from the set of attainable houses N_a using your chosen sampling scheme.
8: Calculate the total walking distance between each of the k stopping point and its surrounding houses. Do this for every sample using your chosen TST algorithm. The result is a set of 10000 walking distances.
9: Calculate descriptive statistics on the 10000 walking distances, and plot the distribution.

References

Arief, R., Hampson, K., Jatikusumah, A., Widyastuti, M., Basri, C., Putra, A., Willyanto, I., Estoepangestie, A., Mardiana, I., Kesuma, I., et al., 2017. Determinants of vaccination coverage and consequences for rabies control in Bali, Indonesia. Front. Vet. Sci. 3, 123.

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC.

Baddeley, A., Turner, R., 2005. spatstat: an R package for analyzing spatial point patterns. J. Statist. Softw. 12, 1–42.

Botes, R., 2023. An Optimised Rabies Vaccination Schedule for Rural Settlements (Master's thesis). University of Pretoria.

Diggle, P., 1985. A kernel method for smoothing point process data. J. R. Statist. Soc. Ser. C 34 (2), 138–147.

Fabris-Rotelli, I., Reynolds, H., Stein, A., Loots, T., 2020. Spatial sampling for a rabies vaccination schedule in rural villages. Environ. Ecol. Statist. 27 (4), 827–845.

Hampson, K., Dushoff, J., Cleaveland, S., Haydon, D., Kaare, M., Packer, C., Dobson, A., 2009. Transmission dynamics and prospects for the elimination of canine rabies. PLoS Biol. 7 (3), e1000053.

Knobel, D., Laurenson, M., Kazwala, R., Boden, L.A., Cleaveland, S., 2008. A cross-sectional study of factors associated with dog ownership in tanzania. BMC Vet. Res. 4 (1), 1–10.

Laporte, G., 1992. The traveling salesman problem: An overview of exact and approximate algorithms. European J. Oper. Res. 59 (2), 231–247.

Lavan, R., King, A., Sutton, D., Tunceli, K., 2017. Rationale and support for a one health program for canine vaccination as the most cost-effective means of controlling zoonotic rabies in endemic settings. Vaccine 35 (13), 1668–1674.

Lebov, J., Grieger, K., Womack, D., Zaccaro, D., Whitehead, N., Kowalcyk, B., MacDonald, P., 2017. A framework for one health research. One Health 3, 44–50.

Lin, S., Kernighan, B., 1973. An effective heuristic algorithm for the traveling-salesman problem. Oper. Res. 21 (2), 498–516.

MacKay, D., 2003. Information Theory, Inference and Learning Algorithms. Cambridge University Press.

Masiira, B., Makumbi, I., Matovu, J., Ario, A., Nabukenya, I., Kihembo, C., Kaharuza, F., Musenero, M., Mbonye, A., 2018. Long term trends and spatial distribution of animal bite injuries and deaths due to human rabies infection in Uganda, 2001–2015. PLoS One 13 (8), e0198568.

Mbilo, C., Coetzer, A., Bonfoh, B., Angot, A., Bebay, C., Cassamá, B., De Benedictis, P., Ebou, M., Gnanvi, C., Kallo, V., Lokossou, R., Manjuba, C., Mokondjimobe, E., Mouillé, B., Mounkaila, M., Ndour, A., Nel, L., Olugasa, B., Pato, P., Pyana, P., Rerambyath, G., Roamba, R., Sadeuh-Mba, S., Suluku, R., Suu-Ire, R., Tejiokem, M., Tetchi, M., Tiembre, I., Traoré, A., Voupawoe, G., Zinsstag, J., 2021. Dog rabies control in West and Central Africa: A review. Acta Trop. 224, 105459.

Michael, H., Kurt, H., 2007. TSP – infrastructure for the traveling salesperson problem. J. Stat. Softw. (ISSN: 1548-7660) 23 (2), 1–21. http://dx.doi.org/10.18637/jss.v023.i02.

Michael, H., Kurt, H., 2022. TSP: Traveling salesperson problem (TSP). URL: https://CRAN.R-project.org/package=TSP. R package version 1.2-0.

Morters, M., McNabb, S., Horton, D., Fooks, A., Schoeman, J., Whay, H., Wood, J., Cleaveland, S., 2015. Effective vaccination against rabies in puppies in rabies endemic regions. Vet. Rec. 177 (6), 150–150.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org/.

Rosenkrantz, D.J., Stearns, R.E., Lewis, P.M., 1977. An analysis of several heuristics for the traveling salesman problem. SIAM J. Comput. 6 (3), 563–581.

Skiena, S., 1990. Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica. Addison-Wesley, pp. 196–198.

Undurraga, E., Millien, M., Allel, K., Etheart, M., Cleaton, J., Ross, Y., Crowdis, K., Medley, A., Vos, A., Maciel, E., Team, V.E., Wallace, R., 2020. Costs and effectiveness of alternative dog vaccination strategies to improve dog population coverage in rural and urban settings during a rabies outbreak. Vaccine 38 (39), 6162–6173.

Wera, E., Mourits, M., Hogeveen, H., 2017. Cost-effectiveness of mass dog rabies vaccination strategies to reduce human health burden in Flores Island, Indonesia. Vaccine 35 (48), 6727–6736.

World Health Organization, 2018. Zero by 30: The global strategic plan to end human deaths from dog-mediated rabies by 2030. URL: https://apps.who.int/iris/handle/10665/272756.

World Health Organization, 2019. Zero by 30: The global strategic plan to end human deaths from dog-mediated rabies by 2030, United Against Rabies Collaboration First annual progress report; Global strategic plan to end human deaths from dog-mediated rabies by 2030. URL: https://apps.who.int/iris/handle/10665/328053.