An Exact Assessment of the Two-Stage EPI Sampling Method

AN EXACT ASSESSMENT OF THE TWO-STAGE EPI SAMPLING METHOD

BY

ATINDER BHARAJ, H.B.Sc.

A thesis

submitted to the Department of Mathematics & Statistics

and the School of Graduate Studies

of McMaster University

in partial fulfilment of the requirements

for the degree of

Master of Science

Master of Science (2017) McMaster University

(Mathematics and Statistics) Hamilton, Ontario, Canada

TITLE:	An Exact Assessment of the Two-Stage EPI Sampling Method
AUTHOR:	Atinder Bharaj H.B.Sc., (Mathematics and Statistics) McMaster Univeristy, Hamilton, Canada
SUPERVISOR:	Dr. Román Viveros-Aguilera Dr. Harry Shannon
NUMBER OF PAGES:	xix, 109

This thesis is dedicated to mom, Daisy, for being my number one fan and role model. I would also like to thank my family and friends for making my university experience worthwhile. I can only imagine where I would be without them. Thank you.

Abstract

The Expanded Program on Immunization Sampling Method (known simply as EPI sampling) is a two-stage sampling procedure originally intended for quick estimation of disease prevalence in large geographical regions. The method was developed in the 1970s and all the subsequent assessments of its performance have been conducted by simulation. In her master's thesis, Reyes (2016) studied in detail the second-stage sampling of the method by developing formulas for the exact calculation of the household inclusion probabilities when sectors are used to identify the initial household to generate the EPI samples. The inclusion probabilities were used in turn to obtain exact mean, bias, variance and mean square error of any estimator of disease prevalence in the population. Thus, no extensive simulations are required and the results are exact rather than just estimates.

This thesis is an extension of Reyes' (2016) work. The extension is two-fold; (a) employing strips rather than sectors because they narrow the geographic area for field workers and to use strips to select the first household for the EPI sample at the secondary stage, and (b) carrying out an analysis on simulated population and sampling plans, using both stages of the EPI method. Analyzing the simulated populations showed that equal weight estimator that samples primary units with

replacement with probability proportional to size (EW1) should be used when the target characteristic is thought to be spread randomly throughout the population, and the Horvitz-Thompson estimator that samples primary units systematically with replacement (HTSYS) should be used when the disease is believed to spread from a central location or through pocketing. Comparing the strip and sector sampling methods at the secondary stage using their effective areas leads to a comparative basis in which the inclusion probabilities are identical for both methods.

Acknowledgements

I would like to thank my advisors Dr. Román Viveros-Aguilera and Dr. Harry Shannon for their constant support and supervision throughout the duration of my thesis. In my undergraduate studies, Dr. Román Viveros-Aguilera was a statistician and teacher who I aspired to be like. He has played a key role in my academic career and future endeavours. Having him as a professor for my undergraduate and graduate studies, and as an advisor for my thesis was truly an honour.

A special thanks to the Canadian Institutes of Health Research for providing funding for this thesis. I hope this paper can be implemented in their studies on rapid survey techniques. In addition, I would like to thank Dr. Patrick Emond of McMaster University for taking his time in teaching me the fundamental knowledge to use SHARCNet, and Jonathan Earl for assisting me in developing functions to utilize the Google Maps API function.

I would like to thank Dr. Paul McNicholas and Dr. Shui Feng for serving in my thesis defense committee and for their useful suggestions to improve the written part of my thesis.

For my time as a undergraduate and graduate student at McMaster, I would like to thank all of my family, friends, colleagues and classmates for making it enjoyable.

Abbreviations

The number following the entries refers to the section in which the notation is introduced.

General Abbreviations

PSU	Primary sampling unit, 1.1
EPI	Expanded Program on Immunization, 1.1
EW	Equal weight, 4.2
HT	Horvitz-Thompson, 4.2
MSE	Mean square error, 4.1
PPS	Probability proportional to size, 1.1
PSU	Primary sampling unit, 1.1
WHO	World Health Organization, 1.1
SSU	Secondary sampling unit, 1.1
WOR	Without replacement, 2.2
WR	With replacement, 2.2
SRS	Simple random sample, 3
SE	Standard error, 4

HTSYS Horvitz-Thompson estimator that systematically samples PSUs, 4.2

HTWR Horvitz-Thompson estimator that samples PSUs with replacement using probability proportional to size, 4.2

HTSRS Horvitz-Thompson estimator where the constant $a_i = \frac{M}{Nm}$, 4.2

NASG National Administration of Surveying, Mapping and Geoinformation, 5.3.1

PDF Probability density function, 3.2.1

GPS Global Positioning Systems, 5.3

GIS Geographic Information Systems, 5.3

API Application Program Interface, 5.3

Notation

The number following the entries refers to the section in which the notation was introduced.

Primary Sampling

m	Number of PSUs sampled from the population, 2.1.1
M	Number of PSUs in the population, 2.1.1
i	Index used to label PSUs, 2.1.1
N_i	Number of SSUs in the i^{th} PSU, 2.1.1
N	Number of SSUs in the population, 2.1.1
r	Number of primary samples to group at least m PSUs using systematic sampling, 2.1.1
k	Number of PSUs remaining to be systematically sampled into m groups, 2.1.1
q	Maximum number of additional PSUs that can be systematically added onto existing systematic samples of size m, 2.1.1
$s_{(m)}$	Arbitrary sample of m PSUs, 2.1.4
$S_{(m)}$	All possible samples of m PSUs, 2.1.4

α_i Inclusion probability of PSU i being included in a sample of size m, 2.1.4

t Index used to identify the selection stage when sampling m PSUs, 2.2.1

ℓ_t The PSU selected when sampling m PSUs at the t^{th} stage, 2.2.1

p_i Proportion of SSUs from town i relative to the entire population, 2.2.3

α_{ik} Joint inclusion probability of including town i and k in the primary sample, 4.1

Secondary Sampling

$\mathcal{H}_i$ Set of all SSUs in PSU i, 3.1

x_j x-coordinate of the location of household j from some fixed town i, 3.1

y_j y-coordinate of the location of household j from some fixed town i, 3.1

r_j Distance of household j from some fixed town's center, 3.1

θ_j Angle of household j relative to the initial arm from some fixed town's center, 3.1

α The size of a strip's base in some random direction, 3.1

h_j Household j from some fixed town, 3.1

D Some household's distance relative to the town center,3.1.1

$\theta_{j,1}$ Smallest direction of a strip that captures household j from some town,3.1.1

$\theta_{j,2}$ Largest direction of a strip that captures household j from some town,3.1.1

Θ Random variable representing the set of non-empty directions θ can take on for some strip, 3.2.1

E_i Empty sub-interval for some strip in some town, 3.2.1

$L(\cdot)$ Length of a union of sets, 3.2.1

n_i Number of SSUs sampled from PSU i, 3.2.2

s'_{n_i} Some EPI sample of of size n_i from town i, 3.2.2

$S'_{(n_i)}$ All EPI samples of of size n_i from town i, 3.2.2

β_{ij} Inclusion probability of obtaining household j from town i, given town i was already sampled. 3.2.2

$\beta_{i,jl}$ Joint inclusion probability of including household j and l in the secondary sample, given town i is sampled. 4.1

π_{ij} Inclusion probability of household j from town i, 4.2

Parameters and Estimators

δ Proportion of the population who are vaccinated against some disease, 4

$\hat{\delta}$ Estimator of the population proportion who are vaccinated against some disease, 4

$\hat{\delta}$ Estimator of the proportion who are vaccinated against some disease in town i, 4

p Proportion of diseased individuals in the population, 4

μ Proportion of disease affecting the population, 4

τ Total number of individuals who carry the disease within a targeted population, 4

$\hat{\mu}$ Estimator for the proportion of disease affecting the population, 4

$\hat{\tau}$ Estimator for the total number of individuals who carry the disease within a targeted population, 4.1

$\hat{\mu}_{EW}$ Equal weight estimator for the prevalence of disease in the population, 4.2

$\hat{\mu}_{HT}$ Horvitz-Thomspon estimator for the prevalence of disease in the population, 4.2

ϕ Mean parameter for a Pareto distribution, 5.2.1

ψ Dispersion parameter for a Pareto distribution, 5.2.1

Other Notations

$A_{\text{strip},\alpha}$ Effective area of a strip with base α, 5.3.2

γ Span or arc of the sector sampling method, 5.3.2

$A_{\text{sect},\gamma}$ Effective area of a sector with span γ, 5.3.2

Contents

List of Figures

Chapter 1

EPI Sampling

1.1 Background Information

Measles, tetanus and polio are just few of the numerous diseases that have infected and killed many infants around the world. To appropriately measure immunization coverage and fight such diseases, many survey designs have been formulated since the 1970s. One of the most rapid assessments of vaccination coverage was developed by the Expanded Program on Immunization (EPI), a World Health Organization (WHO) program. The EPI developed a cluster sampling design, which is not only simple to use, but became very familiar to developing countries during the 1980s (Brogan et al., 1994).

Originally, the EPI cluster survey design was motivated by the work of Serfling and Sherman (1965), who constructed vaccination surveys to be carried out by local health authorities within the United States. Eradication of all common diseases from infants and children was the prime goal underlying the development and application

of the EPI cluster survey, which will be referred to as the EPI method throughout this paper (Brogan et al., 1994).

Execution of the EPI method is a two-stage process. At the primary stage, the EPI sampling procedure suggests selecting primary sampling units (PSUs), towns or villages, with probability proportional to size (PPS). In this context, the EPI method could be categorized as a PPS cluster sample (Lemeshow and Robinson, 1985). The secondary stage consists of selecting secondary sampling units (SSUs), which are typically households in applications of the EPI method. According to Lemeshow and Robinson (1985) and Henderson and Sundaresan (1982), initiating a sample at the secondary stage for the EPI method begins by first selecting a household at random from a selected PSU.

The EPI coverage survey manual developed by the WHO states in detail how to select a household from a selected town. If a household list is available, then it serves as a sampling frame. After assigning a number to each household in the sampling frame, a number is selected at random and the corresponding household is chosen as the first household to visit. All subsequent households surveyed are selected using physical distances from the previous house's front door to all remaining household's front doors. This process is repeated until the desired sample size is obtained. On the other hand, if no household list is available, the EPI method at the secondary level is initiated by first selecting a central location in the sampled town. A random direction from the centre of the town is chosen, by spinning a pen or bottle, and all households along that path reaching the edge of the town are recorded. To select an initial household using this alternate procedure, a household is chosen at random from those that were identified along the path. Subsequent houses are selected using

the aforementioned procedure. If there are no eligible participants in the household for the survey, the surveyor must simply move onto the next closest household. Using only eligible candidates, the process is repeated until the desired sample size is obtained (WHO, 2008).

Many papers have conducted thorough simulations regarding the required sample sizes of PSUs and SSUs that must be collected using the EPI method to obtain adequate precision in the ensuing inferences. According to Brogan et al. (1994), 30 PSUs and 7 SSUs are typically recommended for quick and cost-efficient results. More information on the EPI method and its history could be found in detail in Reyes' (2016) paper titled, "An Analysis of Equally Weighted and Inverse Probability Weighted Observations in the Expanded Program on Immunization (EPI)."

1.2 Approaches to the EPI Method

Over the past five decades, the EPI method has been vital in estimating vaccination coverage, relief operations, nutritional surveys and health care delivery (Reyes, 2016). However, the EPI method has been used in cases where organizations and governments did not comprehend the statistical and analytical aspects of the method correctly. In consequence, many of the results obtained by such groups resulted in inaccurate conclusions. In an attempt to resolve such issues, the traditional EPI method was modified, creating more complex and expensive methods of assessments (Bostoen and Chalabi, 2006).

One main problem with the EPI method is that it tends to give a higher probability of picking households that are closer to the center of the town (Bostoen and

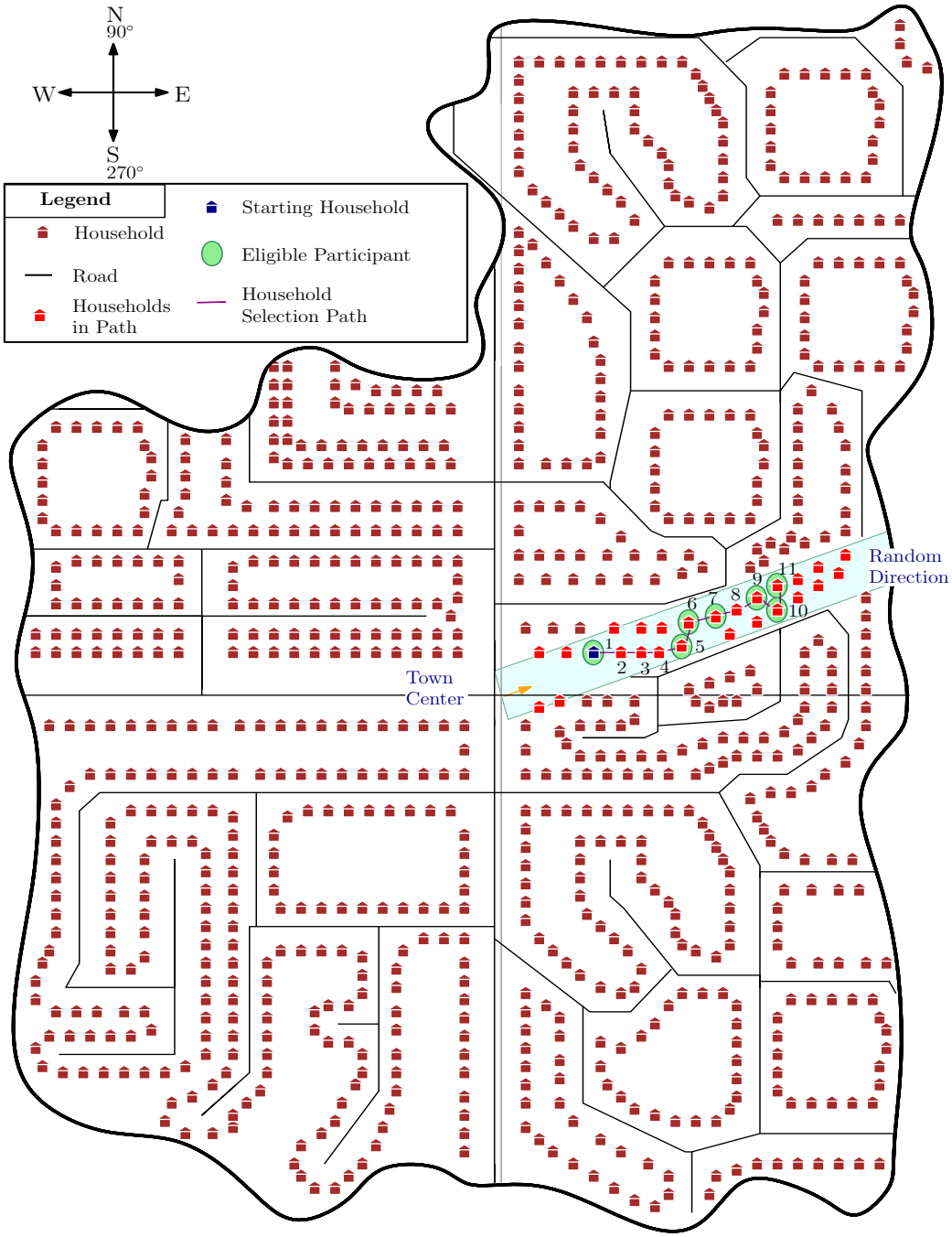


Figure 1.1: Selection of 7 households if a household list is unavailable, according to the EPI method. It can be seen some households are visited, but do not contain eligible participants which results in omission. These households are skipped and the next nearest neighbour is located for the survey, until a sample of 7 eligible candidates is obtained.

Chalabi, 2006). See Chapter 3 for additional details. To increase accuracy and precision of the EPI method, Brogan et al. (1994) suggested a number of modifications: continuing to use PPS sampling to select PSUs, calculating the probability of picking households at the secondary stage and performing a weighted analysis to get an estimate for the vaccination coverage, or any other statistic in general.

In her thesis, Reyes (2016) conducted a detailed analysis of the EPI method at the secondary stage for binary characteristics such as vaccination coverage. Her analysis was two-fold, as it analyzed results from past studies of the EPI method at the secondary stage and made significant advancements at this secondary stage. More specifically, she studied methodological characteristics of the secondary stage of the EPI method by employing the use of an arc or sector to select the initial household, as opposed to the original method devised by the EPI to select households. To continue her work, this paper will explore a two-stage analysis of the EPI method, emphasizing not only the second stage but the primary stage of selecting towns with PPS as well. Simulation of town data will use the work of Reyes (2016), where she constructed a number of towns with varying household spatial and disease prevalence patterns. At the primary stage, this paper focuses on sampling techniques using PPS such as systematic sampling, and sampling with and without replacement. Furthermore, it will discuss the statistical theory behind using a strip or path which has not yet been developed. Additionally, it will explore a method of collecting household positioning on a map when this information is not readily available, through the use of Google Maps. Finally, a number of two-stage weighted assessments are undertaken using the methods and simulated data described above.

Chapter 2

First-Stage Sampling: Selecting the Towns with Probability Proportional to Size

According to the WHO (2008), PSUs have been selected using systematic sampling. To date, not many sampling techniques have been performed at the primary stage. One of the main focuses of this thesis is to explore alternative sampling schemes at the primary stage. Subsequently, several methods have been explored in detail. Since this thesis explores exact inclusion probabilities, many alternative sampling methods, such as Lahiri's method, were not studied and reported because of their complex sampling structures. The consequence of using some of these methods will be explored in the case studies that follow in Chapter 4.

2.1 The Systematic Sampling Method

2.1.1 The Method

To select PSUs with PPS, Henderson et al. (1973), Lemeshow and Robinson (1985) and Bennett et al. (1991) use a systematic sampling procedure. These methods could be extended to produce a general systematic sampling approach. Suppose there are M clusters or towns for some population and a sample of size m is desired, where $0 < m \leq M$. Corresponding to each PSU i, for $i \in \{1, \ldots, M\}$, is the size N_i with $N = \sum_{i=1}^{M} N_i$ being the total size of the population. Proceeding with the systematic sampling method, two lists of size N must be constructed. The first list contains a sequence of integers starting from 1 to N and the second list is composed of PSUs in the population where the PSUs are repeated correspondingly to their respective sizes N_i for all $i \in \{1, \ldots, M\}$. A depiction of the first and second list are represented in Figures 2.1 and 2.2 respectively.

$$\left|1, \ldots, N_1\right|N_1 + 1, \ldots, N_1 + N_2\Big| \ldots \left| \sum_{i=1}^{M-1} N_i + 1, \ldots, N\right|$$

Figure 2.1: List of integers from 1 to N.

$$\underbrace{\overbrace{1, \ \ldots \ , 1,}^{N_1} \ \overbrace{2, \ \ldots \ , 2,}^{N_2} \ \ldots \ \overbrace{, M, \ \ldots \ , M}^{N_M}}_{N}$$

Figure 2.2: List of size N, with M PSUs repeated proportionally to their respective sizes.

When an integer is selected from the first list from 1 to N, it is treated as an identifier used to find the corresponding element in the second list. Let r be the

number of primary samples to group at least m PSUs, such that $r = \max\{t \in \mathbb{Z} | t \leq \frac{N}{m}\}$. This may lead to some PSUs not being grouped into the r samples. Thus, for some non-negative integer k, such that $0 \leq k \leq m-1$, the remaining PSUs can be systematically grouped. Consequently, the following result is obtained

$$N = mr + k. \tag{2.1}$$

The reason there are r groups of at least size m is because when $k > 0$, there are PSUs that have yet to be included into samples. Hence, the maximum size of additional PSUs that can be added systematically onto existing samples of size m are:f

$$q = \min\left\{ s \in \mathbb{Z} | s \geq \frac{k}{r} \right\}$$

This implies that there are:

1. $r \cdot \left(q - \frac{k}{r}\right)$ samples of size $m + q - 1$
2. $r \cdot \left(1 - \left(q - \frac{k}{r}\right)\right)$ samples of size $m + q$

If each of the samples above were summed by their respective sizes, it would lead to the following result

$$\begin{aligned}
&(m+q-1)\left(r\left(q-\frac{k}{r}\right)\right) + (m+q)\left(r\left(1-\left(q-\frac{k}{r}\right)\right)\right) \\
=&(m+q)r\left(q-\frac{k}{r}\right) - r\left(q-\frac{k}{r}\right) + (m+q)r - (m+q)r\left(q-\frac{k}{r}\right) \\
=&-rq + k + mr + qr = mr + k = N.
\end{aligned}$$

The procedure is initiated by constructing the aforementioned lists of size N. Once constructed, a number i between 1 and r is selected randomly from the first list. Next, every r^{th} number is selected after picking i from the first list. If an r^{th} multiple leads to a number greater than N, reject it and stop further sampling. Once this is done and samples have been constructed, assign the corresponding towns from the second list.

2.1.2 Systematic Sampling Example

Suppose $N = 30$, $M = 10$ and samples of size 8 are required. Assume the information required to construct the systematic samples of the population is as in Table 2.1.

Table 2.1: A population with $N = 30$ households divided into $M = 10$ towns.

Town	Index	Town Size
1	1 2	$N_1 = 2$
2	3 4 5 6	$N_2 = 4$
3	7 8 9 10 11	$N_3 = 5$
4	12 13 14 15	$N_4 = 4$
5	16 17	$N_5 = 2$
6	18 19 20	$N_6 = 3$
7	21 22	$N_7 = 2$
8	23 24 25 26	$N_8 = 4$
9	27	$N_9 = 1$
10	28 29 30	$N_{10} = 3$

Using the methodology developed in Section 2.1.1, there can be $r = \max\{t \in \mathbb{Z} | t \leq \frac{N}{m} = 3.75\} = 3$ groups formed. With this, $k = N - mr = 6$ remaining PSUs still need to be systematically grouped into the 3 samples. Systematically sampling the remaining 6 PSUs produces an addition of $q = \min\{s \in \mathbb{Z} | s \geq \frac{k}{r} = \frac{6}{3}\} = 2$ PSUs per sample, at most. Thus, there are $r(q - \frac{k}{r}) = 0$ samples of size $m + q - 1 = 9$

and $r(1-(q-\frac{k}{r}))=2$ samples of size $m+q=10$. Now that information about the samples has been identified, Section 2.1.1 can be used to perform systematic sampling proportional to size. Begin by selecting a number from the first list from 1 to $r=3$. Once selected, add r to each possible starting element and attach the result to its respective sample until an r^{th} multiple leads to a number greater than N. Carrying this procedure through results in the samples of the form

$$r=3;\quad \left\{\begin{array}{cccccccccc} \overbrace{1 & 4 & 7 & 10 & 13 & 16 & 19 & 22}^{m=8} & \overbrace{25 & 28}^{q=2} \\ \hline 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 & 26 & 29 \\ \hline 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 & 27 & 30 \end{array}\right.$$

Clearly, these samples pertain to the first list that was created. To create samples that correspond to the PSUs, simply take each element in the sample which serves as an index to the town it corresponds with. This simple index search results in obtaining the following samples of PSUs:

Table 2.2: Systematic Samples constructed with $N=30$, $M=10$ and $m=8$

Sample 1	1	2	3	3	4	5	6	7	8	10
Sample 2	1	2	3	3	4	5	6	8	8	10
Sample 3	2	2	3	4	4	6	7	8	9	10

2.1.3 Inclusion Probabilities

Clearly, there are r non-overlapping samples formed using the described method. Since the systematic sampling procedure is initiated using a uniform distribution, each sample is equally likely to be selected. Consequently, the probability of selecting

any sample of size m, say $s_{(m)}$, using systematic sampling is $\frac{1}{r}$. Let α_i be probability that town i is included in the primary sample and let $S_{(m)}$ be the set which contains all possible primary samples of size m under the assigned sampling scheme. Hence

$$\alpha_i = \sum_{s_{(m)} \in S_{(m)}} P(s_{(m)}) \mathbb{1}(i \in s_{(m)}) = \sum_{s_{(m)} \in S_{(m)}} \frac{1}{r} \mathbb{1}(i \in s_{(m)}), \tag{2.2}$$

where $\mathbb{1}(i \in s_{(m)}) = \begin{cases} 1, & \text{if PSU } i \text{ is included in the sample } s_{(m)}; \\ 0, & \text{otherwise.} \end{cases}$

If a PSU's size is greater than or equal to r, it must be included in every sample at least once. On the other hand, if a PSU's size is less than r, it appears proportional to its size relative to r. Thus, Equation (2.3) could be generalized to the following result

$$\begin{aligned} \alpha_i &= \begin{cases} 1, & N_i \geq r; \\ \frac{N_i}{r}, & N_i < r \end{cases} \\ &= \frac{\min(N_i, r)}{r}. \end{aligned} \tag{2.3}$$

Alternatively, if $N_i \leq r$ for all $i \in \{1, \ldots, M\}$,

$$\alpha_i = \frac{N_i}{r} = \frac{N}{r}\frac{N_i}{N} = \left(\frac{mr + k}{r}\right)\frac{N_i}{N} = m\frac{N_i}{N} + \frac{k}{r}\frac{N_i}{N}. \tag{2.4}$$

2.2 Sampling WOR using PPS

Perhaps two of the most elementary sampling techniques that are simple to implement are sampling with replacement (WR) and sampling without replacement (WOR). Sampling without replacement refers to a sampling process which does not allow objects or elements to appear more than once within a sample. Similarly, sampling with replacement refers to sampling processes which allow objects or elements to appear more than once in a sample. Sampling WOR produces a subset of samples that are possible under sampling WR, however the sample probabilities generally vary even if they produce the exact same samples.

2.2.1 The Method

Recall, for any PSU or town i, its size is defined by the number of households that are captured in its region, which was denoted by N_i. Thus, the total size of the population is simply the sum of the town's or PSU's sizes which is $\sum_{i=1}^{M} N_i = N$. To sample without replacement with probability proportional to size, begin by constructing a list of size N composed of PSUs being repeated equally to their respective sizes. Refer to Figure 2.2 for an example of the mentioned list. Once this list is acquired, randomly select one element from this list. Once selected, record that particular PSU, and remove all elements in the list that correspond to that PSU. Repeat this process until a sample size of $0 < m \leq M$ is obtained.

Thus, one algorithm to obtain a sample of size m WOR with PPS is the following

1. Construct a list of size N composed of all PSUs replicated to their respective sizes.

2. Let t be an index to identify the element being selected for the sample of PSUs. So the t^{th} PSU selected is recorded as ℓ_t. Randomly select and record a PSU from the list generated in Step 1. After the selected town or PSU ℓ_1, add it to the sample s and remove all PSUs that are equal to i from the list.

3. In general, from the original list, there are $N - \sum_{i=1}^{t} N_{\ell_i}$ PSUs to select in the second stage of sampling PSUs for any given stage t.

4. Repeating steps 2 and 3 produces the required sample $s_{(m)}$.

Example using Sampling WOR with PPS

To illustrate how to sample WOR with PPS, suppose a sample of size $m = 3$ is desired from the information provided in Table 2.1. Begin by constructing the list of size N composed of PSUs replicated proportionally to their sizes. Thus, a list of the following structure is produced:

$$\underbrace{\overbrace{1, \ \ldots \ ,1,}^{N_1=2} \ \overbrace{2, \ \ldots \ ,2,}^{N_2=4} \ \ldots \ \overbrace{,10, \ \ldots \ ,10}^{N_{10}=3}}_{N=30}$$

Figure 2.3: PSU list of size $N = 30$, according to Table 2.1.

According to this list, a possible sample of size $m = 3$ is presented in Table 2.4. Analyzing Table 2.4 shows that using the described algorithm in 2.3.1, a possible sample WOR with PPS is $s_{(3)} = \{1, 3, 9\}$.

Table 2.3: Sampling $m = 3$ towns, according to the information in Table 2.1.

Stage	Selected Town	Town Size	Remaining Towns	Remaining elements in List
$t = 1$	$\ell_1 = 1$	$N_1 = 2$	$M - t = 9$	$N - \sum_{i=1}^{1} N_{\ell_i} = 34$
$t = 2$	$\ell_2 = 3$	$N_3 = 5$	$M - t = 8$	$N - \sum_{i=1}^{2} N_{\ell_i} = 29$
$t = 3$	$\ell_3 = 9$	$N_9 = 1$	$M - t = 7$	$N - \sum_{i=1}^{3} N_{\ell_i} = 28$

2.2.2 Another Approach to Sampling WOR using PPS

An alternative method to generate samples WOR using PPS is to continuously pick PSUs with probability proportional to size, until a sample size of m different towns is obtained. A clear implication of this technique is that the algorithm may pick a number of PSUs that is at least m. However, the algorithm does not stop until m different PSUs are obtained. If there happens to be one very large PSU size relative to all the other PSUs, this algorithm may produce an issue regarding computational time due to the potentially unduly large number of repetitions. Although this method selects PSUs with replacement, the final sample recorded is a sample without replacement.

Thus, the probability that PSU i is selected at any stage of the sampling process is $\frac{N_i}{N}$.

1. Create a list of size N, repeating town i N_i times for $i \in \{1, \ldots, M\}$.

2. Randomly select a PSU from the list generated in step 1 and add it to the sample $s_{(m)}$.

3. Randomly select another PSU from the list in step 1. If the PSU or town is

already in the sample, discard it and sample again. Once a different PSU is obtained, add it to the sample $s_{(m)}$.

4. Repeat the process above until m different PSUs are selected.

2.2.3 Additional Notes: Expected Number of Draws

Throughout this paper, all the samples generated under their assigned sampling scheme have a known sample size. With this technique the desired sample size is known, but the question of how many draws it would take to obtain that particular size m arises naturally. This leads to a famous problem called the *Coupon Collector Problem*. As stated by Ferrante and Frigo (2014) the classic coupon collector problem assumes that there are M coupons, all with an equally likely probability of selection, which give some prize. The coupon collector problem then asks, how many coupons would have to be sampled to obtain all M prizes? In other words, what is the expected number of coupons needed to be purchased to obtain M unique prizes?

In Ferrante and Frigo (2014), they extend this problem by asking how many coupons would be required to obtain a subset of prizes, say m, where the probability of obtaining the coupons are unequally likely. In terms of the alternate method to sample WOR using PPS, this problem could be interpreted as, "How many PSUs must be sampled repeatedly so that $m \leq M$ different PSUs are obtained, given they have unequal probabilities of selection?"

Let $p_i = \frac{N_i}{N}$ be the probability of drawing PSU i in any given draw, such that $\sum_{i=1}^{M} p_i = \frac{N_1+\ldots+N_M}{N} = 1$. Let X_1 denote the random number of draws needed to obtain the first PSU. It is trivial to see that $X_1 = 1$ since any selection in the first

draw results in picking a unique PSU. Let X_2 denote the additional draws needed from the first to obtain a PSU that is not equal to the first draw. In general, let X_i denote the number of draws needed to go from the $(i-1)^{th}$ draw to the i^{th} unique draw in the sample s^*. Thus, the number of random draws needed to obtain $m \leq M$ different PSUs is

$$X_M(m) = X_1 + X_2 + \ldots + X_m = \sum_{i=1}^{m} X_i.$$

If $p_{i_1,i_2,\ldots,i_k} = 1 - p_{i_1} - p_{i_2} - \ldots - p_{i_k}$, then according to Ferrante and Frigo (2014) the expected number of draws needed to obtain the k^{th} unique draw after $(k-1)$ unique draws is

$$\mathbb{E}(X_k) = \sum_{i_1 \neq i_2 \neq \ldots \leq i_{k-1}=1}^{k} \frac{p_{i_1} \cdot p_{i_2} \cdot \ldots \cdot p_{i_{k-1}}}{p_{(i_1)} \cdot p_{(i_1,i_2)} \cdot \ldots \cdot p_{(i_1,\ldots,i_{k-1})}}. \tag{2.5}$$

In general, if there are M PSUs and a sample of $m \leq M$ is desired, the expected draws to obtain a sample of size m with unequal probabilities of selecting each PSU is,

$$\begin{aligned}\mathbb{E}(X_M(m)) &= \mathbb{E}\left(\sum_{i=1}^{m} X_i\right) \\ &= 1 + \sum_{i_1}^{m} \frac{p_{i_1}}{p_{(i_1)}} + \ldots + \sum_{i_1 \neq i_2 \neq \ldots \leq i_{m-1}=1}^{m} \frac{p_{i_1} \cdot p_{i_2} \cdot \ldots \cdot p_{i_{m-1}}}{p_{(i_1)} \cdot p_{(i_1,i_2)} \cdot \ldots \cdot p_{(i_1,\ldots,i_{m-1})}}.\end{aligned} \tag{2.6}$$

In terms of application, this formula has nothing to do with the assessment of the EPI method. However, if there are a few large towns and many small towns, then the

application of this procedure helps researchers estimate the number of PSUs needed to obtain a sample of m unique PSUs.

2.2.4 Inclusion Probabilities

With both methods described in Sections 2.3.1 and 2.3.2, there are inclusion probabilities that could be calculated. As shown in Equation (2.2), the inclusion probability that some town i is included in the sample is the sum of all sample probabilities that contain town i. However, the number of samples that could be produced sampling WOR using PPS is ${}_MP_m$. For a large enough population with many towns and a moderate sample size, the number of samples is simply too large to process. Nonetheless, the calculations for the sample probabilities are still possible under this sampling scheme. Suppose a sample of size m is required and the m PSUs included in the sample in order are $\ell_1, \ell_2, \ldots, \ell_m$. Then, the probability of obtaining an arbitrary sample WOR with PPS $s_{(m)}$ is

$$\begin{aligned} P(s_{(m)}) &= \left(\frac{N_{\ell_1}}{N}\right) \cdot \left(\frac{N_{\ell_2}}{N - N_{\ell_1}}\right) \cdot \left(\frac{N_{\ell_3}}{N - N_{\ell_1} - N_{\ell_2}}\right) \cdot \ldots \cdot \left(\frac{N_{\ell_m}}{N - \sum_{i=1}^{m-1} N_{\ell_i}}\right) \\ &= \frac{N_{\ell_1}}{N} \prod_{i=2}^{m} \frac{N_{\ell_i}}{N - \sum_{j=1}^{i-1} N_{\ell_j}} \end{aligned} \tag{2.7}$$

Although Equation (2.8) is compact, iteratively going through each sample and computing sample probabilities would be an excessive task for any computer for a large enough M and moderate size m. Therefore to reduce this computational strain, another approach must be created to obtain the inclusion probabilities for sampling WOR using PPS. Again, suppose the objective is to compute the inclusion

probability of some town i for samples of size m. Thus,

$$\begin{aligned}
&\alpha_i = P(\text{town } i \text{ is included in } s_{(m)}) \\
&= P(\text{town } i \text{ is selected in } 1^{st} \text{ draw} \cup \ldots \cup \text{town } i \text{ is selected in the } m^{th} \text{ draw}) \\
&= P(\text{town } i \text{ is selected in } 1^{st} \text{ draw}) + \ldots + P(\text{town } i \text{ is selected in the } m^{th} \text{ draw})
\end{aligned} \tag{2.8}$$

$$\begin{aligned}
=&\frac{N_i}{N} + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_{k_1}}{N} \frac{N_i}{N - N_{k_1}} + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \sum_{k_2 \neq \{i, k_1\}}^{M} \frac{N_{k_1}}{N} \frac{N_{k_2}}{N - N_{k_1}} \frac{N_i}{N - N_{k_1} - N_{k_2}} + \cdots \\
&+ \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \sum_{k_2 \neq \{i, k_1\}}^{M} \cdots \sum_{k_{m-1} \neq \{i, k_1, \ldots, k_{m-2}\}}^{M} \frac{N_{k_1}}{N} \frac{N_{k_2}}{N - N_{k_1}} \times \ldots \times \frac{N_i}{N - N_{k_1} - \ldots - N_{k_{m-1}}}.
\end{aligned} \tag{2.9}$$

Note, Equation (2.9) is a result of events being mutually exhaustive. To validate Equation (2.9), it can be used to show that the sum of the inclusion probabilities when sampling WOR using PPS satisfies Equation (2.5) (see Appendix A.1)

To validate this equation, Equation (2.5) can be used. Equation (2.10) is a general formula to directly compute inclusion probabilities when sampling WOR using PPS. However, there are cases which produce compact formulas for inclusion probabilities that are not time consuming to calculate.

2.2.5 Special Cases

$m = 2$

Suppose $m = 2$. This would mean that some town i could appear in the sample in either the first draw or the second draw. Thus,

$$\begin{aligned}
\alpha_i =& P(\text{town } i \text{ is included in } s_{(2)}) \\
=& P(\text{town } i \text{ is selected in the } 1^{st} \text{ draw}) + P(\text{town } i \text{ is selected in the } 2^{nd} \text{ draw}) \\
=& \frac{N_i}{N} + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_{k_1}}{N} \frac{N_i}{N - N_{k_1}} = \frac{N_i}{N} + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_i}{N} \frac{N_{k_1}}{N - N_{k_1}} = \frac{N_i}{N} \left(1 + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_{k_1}}{N - N_{k_1}} \right)
\end{aligned} \tag{2.10}$$

$m = 3$

Much like the case of $m = 2$, a compact form could be found for inclusion probabilities sampling WOR with PPS for samples of size $m = 3$. Thus,

$$\begin{aligned}
\alpha_i =& P(\text{town } i \text{ is included in } s_{(3)}) \\
=& P(\text{town } i \text{ is selected in the } 1^{st} \text{ draw} \cup \ldots \cup \text{town } i \text{ is selected in the } 3^{rd} \text{ draw}) \\
=& P(\text{town } i \text{ is selected in } 1^{st} \text{ draw}) + \ldots + P(\text{town } i \text{ is selected in the } 3^{rd} \text{ draw}) \\
=& \frac{N_i}{N} + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_{k_1}}{N} \frac{N_i}{N - N_{k_1}} + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \sum_{\substack{k_2=1 \\ k_2 \neq \{i, k_1\}}}^{M} \frac{N_{k_1}}{N} \frac{N_{k_2}}{N - N_{k_1}} \frac{N_i}{N - N_{k_1} - N_{k_2}} \\
=& \frac{N_i}{N} + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_i}{N} \frac{N_{k_1}}{N - N_{k_1}} + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \sum_{\substack{k_2=1 \\ k_2 \neq \{i, k_1\}}}^{M} \frac{N_i}{N} \frac{N_{k_1}}{N - N_{k_1}} \frac{N_{k_2}}{N - N_{k_1} - N_{k_2}}
\end{aligned}$$

$$= \frac{N_i}{N}\left[1 + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_{k_1}}{N - N_{k_1}} + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_{k_1}}{N - N_{k_1}} \sum_{\substack{k_2=1 \\ k_2 \neq \{i,k_1\}}}^{M} \frac{N_{k_2}}{N - N_{k_1} - N_{k_2}}\right] \tag{2.11}$$

$$= \frac{N_i}{N}\left[1 + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_{k_1}}{N - N_{k_1}}\left[1 + \sum_{\substack{k_2=1 \\ k_2 \neq \{i,k_1\}}}^{M} \frac{N_{k_2}}{N - N_{k_1} - N_{k_2}}\right]\right] \tag{2.12}$$

Clearly it is seen that Equation (2.13) is not simple; however, it is the most basic expression that is obtainable for $m = 3$. A major piece of information is obtained when computing the inclusion probabilities for $m = 3$. Specifically in Equation (2.13), there is a pattern. When looking for town i to be included in the sample after the j^{th} draw (the $(j+1)^{th}$ draw), it is simply the product of all draws leading up to the j^{th} draw, times the probability of the $(j+1)^{th}$ draw. This piece of information is vital in the proof of the next case. Furthermore if $m = 4$, the last sum of the inclusion probability will compose of a triple sum. For $m = 5$, the last sum will contain a quadruple sum. In general, for any size m, the last sum will be an $(m-1)$-tuple sum. For future works, it may be possible to condense Equation (2.10) using some recursion formula that relates to the idea of the tuple sums.

2.2.6 Equal sized towns

Say across all towns in a population, the sizes were equal. Thus, $N_k = Q$, for $k \in \{1, 2, \ldots, M\}$. Consequently, $\sum_{k=1}^{M} N_k = \sum_{k=1}^{M} Q = MQ$. In this case, it can be proven that the inclusion probability for selecting town i in a sample of size m is $\frac{m}{M}$. In other words, it can be stated that sampling equal sized towns of size m WOR using PPS results in the inclusion probabilities of all PSUs to be $\alpha_i = \frac{m}{M}$. This result

is shown by a proof of induction (see Appendix A.1).

As a matter of fact, it was noticed through simulations that the behavior of inclusion probabilities for large M with relatively equal sized towns is close to the inclusion probabilities obtained in the case of equal sized towns while sampling WOR with PPS.

2.3 Sampling with Replacement using PPS

2.3.1 The Method

Another classic sampling scheme is sampling with replacement. The major advantage of sampling WR using PPS is that it is very tractable. The most impressive feature it brings forth to the EPI sample is how straightforward it is to compute inclusion probabilities using this particular method. Much like sampling WOR using PPS, constructing all possible samples is no easy task, as many computers would fail to generate all possible samples under such a scheme for a substantially large number of towns. To be more precise, the total number of samples that could be constructed of size m for some population with M towns is M^m. Logically, every PSU has a chance to appear at each draw of the sample. This gives chance to each PSU to be selected m times repeatedly for a sample of size m.

Although generating all samples is difficult for a large sample size and number of PSUs, generating a particular sample WR using PPS is however simple. The following steps describe the algorithm:

1. Create a list of size N, repeating each town in the list proportional to its size.

2. Randomly select a PSU from the list constructed in step 1. Once selected, add that particular PSU to the sample $s_{(m)}$.

3. Repeat step 2. until a sample of size m is obtained.

Example using Sampling WR with PPS

To show how to produce a sample WR using PPS, Table 2.1 will be used again. Suppose a sample of size $m = 4$ is desired. As in the example of sampling WOR with PPS, begin by constructing a list of size N composed of PSUs replicated according to their sizes (refer to Figure 2.3 for the generated list). Using the same notation as in Section 2.3.1, a possible sample produced WR with PPS is presented in the following table.

Table 2.4: Sampling $m = 4$ towns WR using PPS, according to information in Table 2.1.

Stage	Selected Town	Town Size
$t = 1$	$\ell_1 = 1$	$N_1 = 2$
$t = 2$	$\ell_2 = 3$	$N_3 = 5$
$t = 3$	$\ell_3 = 9$	$N_9 = 1$
$t = 4$	$\ell_4 = 3$	$N_3 = 5$

So, according to Table 2.5, a possible sample WR using PPS is $s_{(4)} = \{1, 3, 9, 3\}$ for the data described in Table 2.1.

2.3.2 Inclusion Probabilities

Although sample probabilities are difficult to compute because all possible samples are too many to store and process in the average computer, individually they still have a closed form. Suppose a sample $s_{(m)}$ is desired and the m PSUs included in the sample in order are $\ell_1, \ell_2, \ldots, \ell_m$. Then, the probability of obtaining an arbitrary sample $s_{(m)}$ WR using PPS is

$$\begin{aligned} P(s_{(m)}) &= \frac{N_{\ell_1}}{N} \cdot \frac{N_{\ell_2}}{N} \times \ldots \times \frac{N_{\ell_m}}{N} \\ &= \frac{1}{N^m} \times \prod_{i=1}^{m} N_{\ell_i} \end{aligned} \tag{2.13}$$

Again, the number of samples is extremely large for large M and at least moderate sized m. It may not be feasible to spend time generating all possible samples and computing each sample's probability of selection to obtain inclusion probabilities. Alternatively, a more direct method to compute inclusion probabilities is used. Consider a sample of size m collected using sampling WR with PPS. Thus town i has a probability of $\frac{N_i}{N}$ of being included in the sample for any given selection stage. The probability that town i is not chosen is $1 - \frac{N_i}{N} = \frac{N - N_i}{N}$. This implies that the probability that town i is not chosen in a sample of size m is $\left(\frac{N-N_i}{N}\right)^m$. As a result, the probability of town i being included in a sample of size m is

$$\alpha_i = 1 - \left(\frac{N - N_i}{N}\right)^m, \tag{2.14}$$

for all $i \in \{1, \ldots, M\}$

2.3.3 Connections between systematic sampling, sampling WOR and WR with PPS

There are implications when sampling without replacement. Specifically, Lohr (2009) states that when sampling m PSUs without replacement using PPS, the inclusion probabilities must satisfy the following result

$$\sum_{i=1}^{M} \alpha_i = m. \tag{2.15}$$

With this result, a special case could be looked at under systematic sampling where there is a connection between systematic sampling with PPS and sampling WOR with PPS. Consider the case when $k = 0$ and $N_i \leq r$ for all $i \in \{1, \ldots, M\}$. In consequence, $N = mr$ and

$$\begin{aligned} \alpha_i &= \frac{N_i}{r} = \frac{m}{m}\frac{N_i}{r} = m\frac{N_i}{N} \\ \therefore \sum_{i=1}^{M} \alpha_i &= \sum_{i=1}^{M} m\frac{N_i}{N} = \frac{m}{N}\sum_{i=1}^{M} N_i = m. \end{aligned}$$

Intuitively, this statement is sensible because each PSU will either appear in all samples or a subset of them and there will be no PSUs left unaccounted for in the process ($k = 0$). Conclusively, if m PSUs are to be systematically sampled when $k = 0$ and $N_i \leq r$ for all i, then this is the equivalent to sampling WOR using PPS. Regarding the samples themselves when systematically sampling under the mentioned conditions, only a subset of the samples are produced from the samples that would be possible WOR using PPS. To clarify, if all samples of size m for some population $N = mr$ under the systematic sampling procedure, say $S_{\mathrm{m,SS}}$, and under

sampling WOR with PPS, say $S_{m,\text{WOR}}$, are generated then $S_{m,\text{SS}} \subseteq S_{m,\text{WOR}}$.

In the case where $N_i > r$ for some $i \in \{1, \ldots, M\}$, those PSUs must be included in at least one sample more than once. Similarly, when sampling WR using PPS, PSUs could appear in a sample more than once. Consequently, when systematically sampling m PSUs and some PSU i has size $N_i > r$, it produces at least one sample which contains those particular PSUs more than once. These particular samples which contain PSUs included more than once would then be a subset of the possible samples that could be constructed by sampling WR using PPS. Again, if all possible samples of size m for some population $N = mr + k$ under systematic sampling and under sampling with replacement using PPS, say $S_{m,\text{WR}}$, are considered, then $S_{m,\text{SS}} \subset S_{m,\text{WR}}$. Note, it is not possible for $S_{m,\text{SS}} = S_{m,\text{WR}}$ because the case where some PSU i is included in a systematic sample more than once, it eliminates the possibility of there being some other sample(s) that do not include that PSU i. Under a different variation of the systematic sampling procedure, it might be possible.

Although systematic sampling could produce samples which are the same as sampling WOR and WR using PPS, the inclusion probabilities are not the same in general. For example, when PSU i has size $N_i > r$, its inclusion probability is $\alpha_i = 1$. Compared to sampling with replacement, this is not the same because the inclusion probabilities when sampling WR with PPS can never be 1 exactly. This will be shown in greater detail in later sections.

Chapter 3

Second-Stage Sampling: Selecting Households Using the EPI Method

Chapter 2 explored a variety of techniques to sample with probability proportional to size. The core data needed to compose EPI samples at the primary stage using PPS sampling techniques require PSU sizes, which can be constructed using the number of SSUs within specific PSUs. In context of the EPI method, households with eligible candidates for a particular survey are typically the SSUs.

As such, household data plays an equally vital role in the construction of samples at the secondary stage of sampling within the PSUs. In her paper, Reyes (2016) explored in great detail the various implications of the methods used to sample at the secondary stage in previous studies of the EPI method. Depending on the method used to collect samples at the secondary stage, a variety of samples and sample probabilities are possible. According to Reyes (2016), a simple random sample (SRS) of households from a particular town or village gives equal weight to all household(s) in

the town. On the other hand, the EPI method selects the nearest unsampled household(s) from the last sampled household, producing unequal sample probabilities in general.

3.1 Strip Sampling

As previously mentioned, the EPI method tends to give higher probabilities of picking the first household to those that are near the center of the town or village. To address this problem, Reyes (2016) explored a method known as arc or sector sampling. Aside from the detailed analysis of this alternate method, there has yet to be any released literature exploring the statistical aspects of the original sampling method at the secondary stage for the EPI method, as proposed by the World Health Organization (2008). The work that follows will expand and modify this traditional approach.

For simplicity, consider an arbitrary town i, for $i \in \{1, \ldots, M\}$, which is composed of households $\mathcal{H}_i \in \{1, \ldots, N_i\}$. Since each household location could be mapped onto a two-dimensional Cartesian plane, they could all be represented by their Cartesian coordinates (x_j, y_j). Each household's Cartesian coordinate could also be transformed to produce a polar coordinate, such that

$$\begin{aligned}(x_j, y_j) &\rightarrow (r_j, \theta_j) \\ r_j &= \sqrt{x_j^2 + y_j^2} \\ \theta_j &= \tan^{-1}\left(\frac{y_j}{x_j}\right),\end{aligned}$$

where r_j and θ_j represent the radius from the origin and the angle from the initial

arm (positive x-axis), respectively.

According to the WHO (2008), after a direction has been randomly selected from the center of the town by either software or simply spinning a bottle, the surveyor must walk in that direction and record all households until the edge of the town has been reached. This description is vague because it does not clearly indicate what households in proximity to the surveyor should be recorded. More specifically, if 2 or more surveyors were asked to walk along a particular path from the center of the town, it is likely that at least one would have a different set of households recorded in their collection compared to the other surveyors.

To develop a more specific method of recording households along a certain path or direction of a town, strips will be used. These particular strips would be composed of 4 sides, of which 3 are essential, and the strips will rotate from directions in $[0, 2\pi)$ about the midpoint of the width of the strip. More precisely, the width of the strip would take on some value $\alpha \in (0, 2 \max_{i \in \{1,\ldots,N_i\}} r_i]$ and be perpendicular to the initial spin or direction. The lengths of the strip would both be oriented in the direction of the initial spin, go towards the edge of the town and be perpendicular to the width of the strip. In application, the algorithm needed to perform strip sampling is:

1. Specify the width of the strip, α.

2. Randomly select a direction θ from $[0, 2\pi)$ using a uniform distribution. Produce the corresponding strip with width α from the origin in the direction θ. If the strip is empty (contains no households), repeat this step until a non-empty strip is obtained.

3. Record all households in the non-empty strip and randomly select one household. Execute the nearest neighbor algorithm as mentioned by the WHO (2008) and collect a sample of n households.

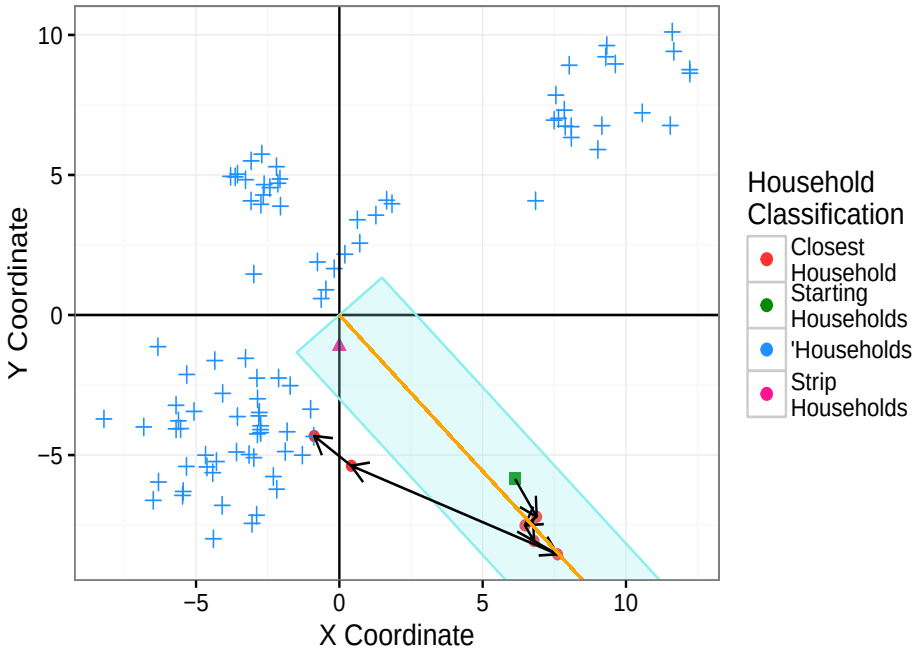


Figure 3.1: Illustration of an EPI sample of size 3 at the secondary stage using a strip of width $\alpha = 4$ and $\theta = 5.473$ for a town with 100 households. Arrows were used to show the sequence of households selected in the nearest neighbour process.

Strip sampling is geometrically complex when finding the directions of θ that bound some household j, say h_j. The key concept to grasp is that there are three lines that construct the strip and utilizing them leads to finding the directions that capture h_j. When analyzed correctly, the process is easier to comprehend and constructing any strip sample probabilities at the secondary stage is facilitated.

3.1.1 Household directions with radii greater than $\frac{\alpha}{2}$

Within the width of every strip, a circle with radius $\frac{\alpha}{2}$ can be drawn about the origin or center of the town. In this case, the two sides of the strip are perpendicular to this arbitrary circle at the two vertices. This would imply that any household h_j can be found lying on the side of two strips with unique directions of θ and form two right angle triangles to the origin of the town and each of the vertices independently. Since this spatial arrangement creates a right angle triangle, the sides $\alpha/2$ and D, distance from the origin to the household form a third side, say L, using the Pythagorean theorem. A visual depiction of this method is provided in Figure 3.2.

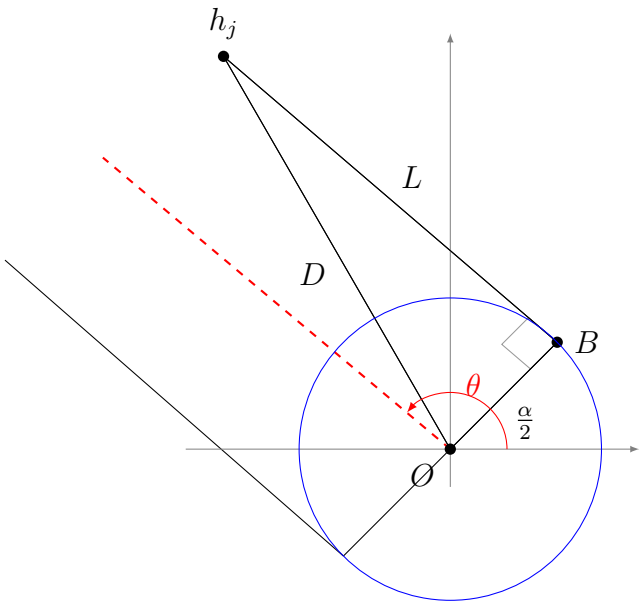


Figure 3.2: Constructing a right angle triangle with household h_j found on the edge of some strip with direction θ, intersecting at one of the vertices, B.

The coordinates of the origin are fixed at $(0, 0)$ and the coordinates of household

h_j are also fixed at some point say (x_j, y_j). Thus,

$$D = \sqrt{(x_j - 0)^2 + (y_j - 0)^2}.$$

Using the Pythagorean Theorem,

$$L = \sqrt{D^2 - \left(\frac{\alpha}{2}\right)^2}.$$

The aim of this procedure is to find the coordinates associated with the vertices, say B_1 and B_2, so that the corresponding direction to these coordinates, say θ_{B_1} and θ_{B_2} respectively, could be used to find the directions that some household h_j is found in (see Figure 3.3). The relationship between the direction of these coordinates and the bounds is that they are perpendicular to the bounds. Once the directions θ_{B_1} and θ_{B_2} are obtained, simply adding or subtracting $\frac{\pi}{2}$ would give the bounds for the strip that capture some household h_j.

Begin by letting the unknown coordinates associated with either points B_1 or B_2 be (x, y). Using the fact that the squared distance from the origin to point B_1 or B_2 must be $(\alpha/2)^2$, one equation is constructed to obtain x and y. Another equation is formed using the fact that the squared distance between points h_j and B_1 or B_2 must be L as shown in Figure 3.3. With this, a system of two equations is obtained with two unknowns x and y.

$$x^2 + y^2 = \left(\frac{\alpha}{2}\right)^2 \tag{3.1}$$

$$(x - x_j)^2 + (y - y_j)^2 = L^2 \tag{3.2}$$

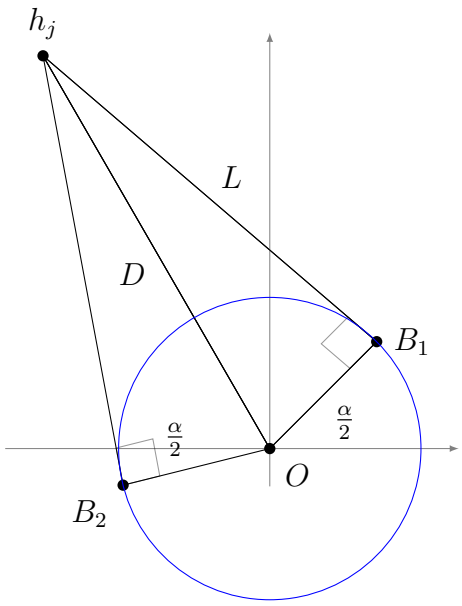


Figure 3.3: An illustration showing the relationship with some household h_j having $r_j > \alpha/2$, and the vertices B_1 and B_2.

Upon expanding Equation (3.1) and subtracting it from Equation (3.2), the following results are obtained:

$$\begin{aligned}
x^2 - 2x_j x + x_j^2 + y^2 - 2y_j y + y_j^2 - x^2 - y^2 &= L^2 - \left(\frac{\alpha}{2}\right)^2 \\
-2x_j x + x_j^2 - 2y_j y + y_j^2 &= D^2 - 2\left(\frac{\alpha}{2}\right)^2 \\
2y_j y &= -2x_j x + \left(x_j^2 + y_j^2 - D^2 + 2\left(\frac{\alpha}{2}\right)^2\right) \\
y &= -\frac{x_j}{y_j}x + \frac{x_j^2 + y_j^2 - D^2 + 2\left(\frac{\alpha}{2}\right)^2}{2y_j} \\
y &= c_2 x + c_1,
\end{aligned}$$

where $c_1 = \frac{x_j^2+y_j^2-D^2+2\left(\frac{\alpha}{2}\right)^2}{2y_j}$ and $c_2 = -\frac{x_j}{y_j}$.

Substituting $y = c_2x + c_1$ into Equation (3.2) yields

$$\begin{aligned} x^2 + (c_2x + c_1)^2 &= \left(\frac{\alpha}{2}\right)^2 \\ x^2 + c_2^2x^2 + 2c_1c_2x + c_1^2 - \left(\frac{\alpha}{2}\right)^2 &= 0 \\ (1 + c_2^2)x^2 + (2c_1c_2)x + \left(c_1^2\left(\frac{\alpha}{2}\right)^2\right) &= 0 \\ \therefore x = \frac{-2c_1c_2 \pm \sqrt{(2c_1c_2)^2 - 4(1 + c_2^2)\left(c_1^2\left(\frac{\alpha}{2}\right)^2\right)}}{2(1 + c_2^2)} & \end{aligned} \tag{3.3}$$

It is clear why two different coordinates are produced mathematically; one set of coordinates corresponding to B_1 and the other to B_2. However, there is a resulting complication from this procedure. The bounds for those households that lie on the x-axis cannot be evaluated because both c_1 and c_2 contain denominators of y_j, which would equal 0 on the x-axis. To avoid this issue, a vertical shift either up or down by some small value could be continuously applied to all households, until no household is found on the x-axis.

The coordinates found, say (x_1, y_1) and (x_2, y_2), are utilized to compute θ values that form a set capturing h_j. Spatially, the bounds can be systematically calculated depending on whether households have y-coordinates that lie either above or below the x-axis. Let $\theta_{j,1}$ and $\theta_{j,2}$ represent the lower and upper bound of the directions for the strip of width α that capture some household j. If the y-coordinate of household j is greater than 0, $\theta_{j,1}$ and $\theta_{j,2}$ is found by letting:

1. $x_j^{(1)} = \min(x_1, x_2)$ and $y_j^{(1)}$ be the corresponding coordinate.
   Thus, $\theta_{j,1} = \tan^{-1}\left(\frac{y_j^{(1)}}{x_j^{(1)}}\right) - \frac{\pi}{2}$

2. $x_j^{(2)} = \max(x_1, x_2)$ and $y_j^{(2)}$ be the corresponding coordinate.

   Thus, $\theta_{j,2} = \tan^{-1}\left(\frac{y_j^{(2)}}{x_j^{(2)}}\right) + \frac{\pi}{2}$

If the y-coordinate of household j is less than 0, $\theta_{j,1}$ and $\theta_{j,2}$ are found by letting:

1. $x_j^{(1)} = \max(x_1, x_2)$ and $y_j^{(1)}$ be the corresponding coordinate.

   Thus, $\theta_{j,1} = \tan^{-1}\left(\frac{y_j^{(1)}}{x_j^{(1)}}\right) - \frac{\pi}{2}$

2. $x_j^{(2)} = \min(x_1, x_2)$ and $y_j^{(2)}$ be the corresponding coordinate.

   Thus, $\theta_{j,2} = \tan^{-1}\left(\frac{y_j^{(2)}}{x_j^{(2)}}\right) + \frac{\pi}{2}$

So for any household h_j that has radius $r_j > \frac{\alpha}{2}$, the method above can be used to give the range of directions that capture it, which are $\theta \in [\theta_{j,1}, \theta_{j,2}]$ using a strip with width α.

3.1.2 Households with radii less than or equal to $\frac{\alpha}{2}$

It is important to distinguish households that have radii exceeding $\alpha/2$ relative to the origin because the method used to compute the directions θ that bound households with radii less than or equal to $\frac{\alpha}{2}$ varies. Specifically, for any h_j with a radius $r_j \leq \frac{\alpha}{2}$ from the origin, it is not the sides of the strip that will be used to find the bounds, but rather the base of the strip. Suppose the corresponding direction of h_j relative to the positive x-axis is θ_{h_j}. From θ_j, the strip can rotate about the origin up to 90° counterclockwise and clockwise, while still containing household h_j. Letting $\theta_{j,1} = \theta_{h_j} - \frac{\pi}{2}$ and $\theta_{j,2} = \theta_{h_j} + \frac{\pi}{2}$, the bounds capturing some household h_j are $\theta \in [\theta_{j,1}, \theta_{j,2}]$. Consequently, the range of the directions that capture household h_j with $r_j \leq \frac{\alpha}{2}$ is π. Spatially, this implies that households with radii exceeding $\frac{\alpha}{2}$

from the origin have a range of directions that capture them being strictly less than π. In general, the farther the household is from the origin, the smaller the range of directions that capture it.

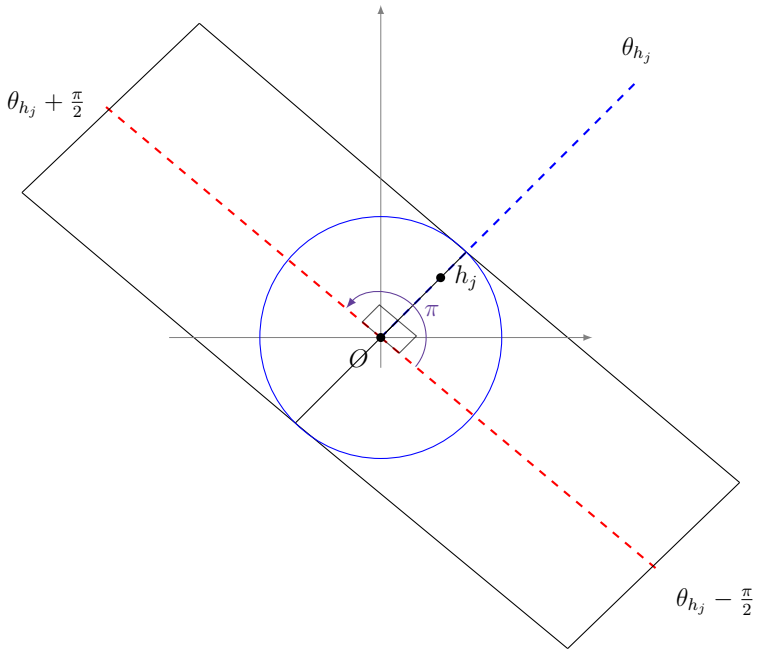


Figure 3.4: An illustration showing the relationship with some household h_j having $r_j \leq \alpha/2$. It is clearly observed that the strip with base α can rotate from the direction $\theta_{h_j} - \frac{\pi}{2}$ to $\theta_{h_j} + \frac{\pi}{2}$ while ensuring that h_j is contained in that particular strip.

3.2 Computation of Household Inclusion Probabilities using Strip Sampling

Naturally from a sampling point of view, a quantity of general interest is the inclusion probability of every household h_j using strip sampling. It is important to note that since the EPI method is a two-stage sampling method, the households selected at

the secondary stage depend on which towns were selected at the primary stage. To compute the household inclusion probabilities, two probabilities are needed: the probability of selecting the first household and the probabilities of obtaining the rest of the sample members at the secondary stage.

3.2.1 Probability of Selecting the First Household.

One of the intricate components needed to calculate the probability of selecting some household j first, $P(h_j)$, using strip sampling for a fixed base α, is the process in finding all empty strips. An empty strip could be defined as a strip of base α which contains no household and a non-empty strip could be defined as a strip of base α which contains at least one household. Let Θ be a random variable representing the set of directions θ for a strip with fixed base α, which are non-empty.

After computing the lower and upper bounds that capture each household, they could be stored in two vectors v_1 and v_2 respectively. Specifically, $v_1 = \{\theta_{1,1}, \theta_{2,1}, \ldots, \theta_{N_i,1}\}$ and $v_2 = \{\theta_{1,2}, \theta_{2,2}, \ldots, \theta_{N_i,2}\}$. Under the assumption that two or more households cannot be assigned the same Cartesian coordinates, vectors v_1 and v_2 must contain unique pairs of values for those households with radii greater than $\alpha/2$. In contrast, households with radii less than $\alpha/2$ could have the exact same bounds only if their directions θ from the origin of the town are identical.

The main purpose for these vectors, v_1 and v_2, is that they are used to produce the sets that contain empty strips. However, an issue that arises is that some households may have polar coordinates that lie in quadrants 1 and 4. Consequently, their lower and upper bounds may be beneath 0 radians and exceed 2π radians, respectively, due to the procedure for how bounds are computed as seen in 3.1.1 and 3.1.2. For

this reason, 2π should be added to each element in v_1 and each element in v_2 should be subtracted by 2π. Two new vectors that account for this could be defined as $v_1^* = \{v_1, v_1 + 2\pi\}$ and $v_2^* = \{v_2, v_2 - 2\pi\}$. A method to compute the sets which contain empty strips, for some fixed base α, is given in the following algorithm:

1. For some household h_j, generate a strip with direction $\theta = \theta_{j,2}$ so that the strip's directions is the upper bound which captures household j.

2. Identify all households in the strip, excluding those that have an upper bound of $\theta_{j,2}$. If there are no other households, an empty strip is produced in the sets of theta values from

$$\theta \in [\theta_{j,2}, \min(v_1^* > \theta_{j,2})]\,.$$

   Otherwise, no empty strips are produced.

3. Repeat steps 1-2 for each household in the town.

These empty sub-intervals (say K of them) would be non-overlapping and can be denoted by $E_1, \ldots, E_K$. Therefore, the set of directions that produce non-empty strips can be defined as

$$\begin{aligned} \Theta &= [0, 2\pi) - \bigcup_{k=1}^{K} E_k \\ &= [0, 2\pi) \cap \left(\bigcup_{k=1}^{K} E_k \right)^C . \end{aligned} \tag{3.4}$$

Since empty directions are discarded, the effective sampling set for θ which produce

non-empty strips for some width α is Θ and will have length

$$L(D) = 2\pi - \sum_{k=1}^{K} L(E_k), \tag{3.5}$$

where $L(\cdot)$ is a function that computes the length of some set(s) of intervals. For dense towns or cities like Toronto, the town center is mostly found at the center of a populated area. Thus, it is very unlikely to observe empty strips and $L(D)$ for such a town will most likely be 2π. On the other hand, towns that are populated in a specific area away from the sea or large power plants have a greater chance in displaying empty strips. In a real world setting, each town has an associated town center. On the other hand for simulated towns, we can define the town center to be the point given by the average of the households coordinates.

Since the initial direction for θ is randomly uniformly selected, its appropriate probability density function (PDF) for sampling θ is

$$f(\theta) = \begin{cases} \frac{1}{L(\Theta)}, & \text{if } \theta \in D; \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

For any household h_j, $P(h_j)$ can be expressed as

$$\begin{aligned} P(h_j) &= \mathbb{E}(g(\theta, h_j)) && (3.7) \\ &= \int_{\Theta} g(\theta, h_j) f(\theta) d\theta \\ &= \int_{\Theta} \frac{1}{L(\Theta)} g(\theta, h_j) d\theta, && (3.8) \end{aligned}$$

where

$$g(\theta, h_j) = \begin{cases} \frac{1}{n(\theta,\alpha)}, & \text{if strip}(\theta,\alpha) \text{ contains household } h_j \text{ ;} \\ 0, & \text{otherwise;} \end{cases} \tag{3.9}$$

and $n(\theta, \alpha)$ represents the number of households contained in a strip of width α, in the direction θ. Clearly, only directions that contain h_j are needed to be considered. This reduces the integral to compute $P(h_j)$ in Equation (3.7) from the entire set Θ to the interval $[\theta_{1,j}, \theta_{2,j}]$.

Much like in the sector sampling method discussed by Reyes (2016), there will be a subinterval $(\theta^* - b, \theta^* + b)$, for some $b \in \mathbb{R}_0^+$, so that the group of households contained in the strips with width α in this subinterval are all the exact same. This allows strips with directions in the interval $[\theta_{j,1}, \theta_{j,2}]$ to be observed as the union of non-overlapping subintervals $C_1, C_2, \ldots, C_A$, such that $[\theta_{j,1}, \theta_{j,2}] = \bigcup_{a=1}^{A} C_a$, for which the households contained in the strip(θ, α) will be the same for every direction in C_a. In particular, $n(\theta, \alpha)$ will be constant in C_a. Accordingly, Equation (3.7) could be thought of as a finite sum of non-overlapping subintervals $C_1, C_2, \ldots, C_A$. Letting θ_{C_a} be any direction in the subinterval C_a,

$$\begin{aligned} P(h_j) &= \int_{\Theta} \frac{1}{L(\Theta)} g(\theta, h_j) d\theta \\ &= \int_{\Theta} \frac{1}{L(\Theta)} \frac{1}{n(\theta,\alpha)} d\theta \\ &= \int_{\theta_{j,1}}^{\theta_{j,2}} \frac{1}{L(\Theta)} \frac{1}{n(\theta,\alpha)} d\theta \\ &= \frac{1}{L(D)} \sum_{a=1}^{A} \frac{L(C_a)}{n(\theta_{C_a},\alpha)}. \end{aligned} \tag{3.10}$$

3.2.2 Generating EPI Samples

Using the strip method at the second stage of the EPI method, a non-empty strip of width α from the town center is generated, and then the first household is randomly selected within the non-empty strip. From the remaining households, the second household selected is the one which is physically closest to the first household. In general, for $q \geq 2$, the q^{th} household is sampled from the $N_i - q + 1$ remaining households from town i, which is physically closest to the $(q-1)^{th}$ household that was previously sampled. As mentioned before, if there is more than one household that is physically closest from the previously sampled household, then one household is randomly selected. This process is repeated in town i until a sample of size n_i is obtained. For simplicity, define the process of selecting the initial household using a non-empty strip of width α and selecting subsequent nearest neighbors as an EPI path.

If all households were distributed such that only one household is the nearest neighbor to any other household, the number of samples that could be generated along all EPI paths must be equal to N_i. In general, let $S'_{(n_i)}$ denote all possible EPI samples of size n_i for town i and let s'_{n_i} denote one sample of households along a particular EPI path in the order the households were selected. Let the vector $\boldsymbol{h_{i,q}} = (h_{(1)}, h_{(2)}, \ldots, h_{(q)})$ denote a sample of q households along some EPI path and $h_{(j)}$ be the j^{th} household included in the sample. Note, $h_{(j)}$ is usually not the same as h_j. The probability that $\boldsymbol{h_{i,q}}$ is selected is

$$\begin{aligned} P(\boldsymbol{h_{i,q}}) =& P((h_{(1)}, h_{(2)}, \ldots, h_{(q)}) \\ =& P(h_{(q)} | (h_{(1)}, h_{(2)}, \ldots, h_{(q-1)})) \times P((h_{(1)}, h_{(2)}, \ldots, h_{(q-1)})) \end{aligned}$$

$$
\begin{aligned}
&= P(h_{(q)}|(h_{(1)}, h_{(2)}, \ldots, h_{(q-1)})) \times P(h_{(q-1)}|(h_{(1)}, h_{(2)}, \ldots, h_{(q-2)})) \times \\
&\quad \times P((h_{(1)}, h_{(2)}, \ldots, h_{(q-2)})) \\
&\quad \vdots \\
&= P(h_{(q)}|(h_{(1)}, h_{(2)}, \ldots, h_{(q-1)})) \times P(h_{(q-1)}|(h_{(1)}, h_{(2)}, \ldots, h_{(q-2)})) \times \\
&\quad \times \ldots \times P(h_{(2)}|h_{(1)}) \times P(h_{(1)}) \\
&= P((h_{(2)}, h_{(3)}, \ldots, h_{(q)})|h_{(1)}) \times P(h_{(1)}). \qquad (3.11)
\end{aligned}
$$

Let $\boldsymbol{g_{i,q}} = \{k \in \mathcal{H}_i | k \notin \boldsymbol{h_{i,q}}\}$ be the remaining households in town i, after the first q have been selected. After $q-1$ households have been sampled, a function to find the q^{th} household using the nearest neighbor approach is,

$$
\underset{u \in \boldsymbol{g_{i,q-1}}}{\arg\min}\, d_{u,h_{(q-1)}} = \underset{u \in \boldsymbol{g_{i,q-1}}}{\arg\min} \sqrt{(x_u - x_{h_{(q-1)}})^2 + (y_u - y_{h_{(q-1)}})^2}. \qquad (3.12)
$$

Recall, there are cases where households are equidistant from each other, as observed in many residential areas. Performing the nearest neighbor algorithm, the number of households closest to $h_{(q-1)}$ from the set of households $\boldsymbol{g_{i,q-1}}$ is denoted by $n(h_{(q-1)}, \boldsymbol{g_{i,q-1}})$. Thus, the probability that household $h_{(q)}$ is included in the sample, given that households $h_{(1)}, h_{(2)}, \ldots, h_{(q-1)}$ are included in the sample is given by

$$
P(h_{(q)}|(h_{(1)}, h_{(2)}, \ldots, h_{(q-1)})) = \begin{cases} \frac{1}{n(h_{(q-1)}, \boldsymbol{g_{i,q-1}})}, & \text{if } h_q \in \boldsymbol{g_{i,q-1}}; \\ 0, & \text{otherwise.} \end{cases} \qquad (3.13)
$$

After finding all EPI paths of size n_i for N_i starting households, $s'_{(n_i)}$, the probability of obtaining $s'_{(n_i)}$, $P(s'_{(n_i)})$, can be computed using Equation (3.10). The resulting inclusion probability of h_j, using a strip of width α within some town i, is simply the sum of all EPI samples in town i that contain h_j. More generally, denote β_{ij} as the inclusion probability of obtaining household j from town i in the secondary sample, given that town i was included in the primary sample. This can be expressed by

$$\beta_{ij} = \sum_{s'_{(n_i)} \in S'_{(n_i)}} P(s'_{(n_i)}) \mathbb{1}(j \in s'_{(n_i)}), \tag{3.14}$$

where

$$\mathbb{1}(j \in s'_{(n_i)}) = \begin{cases} 1, & \text{if household } j \text{ from town } i \text{ is included in the secondary} \\ & \text{sample, given that town } i \text{ is included in the primary sample;} \\ 0, & \text{, otherwise.} \end{cases}$$

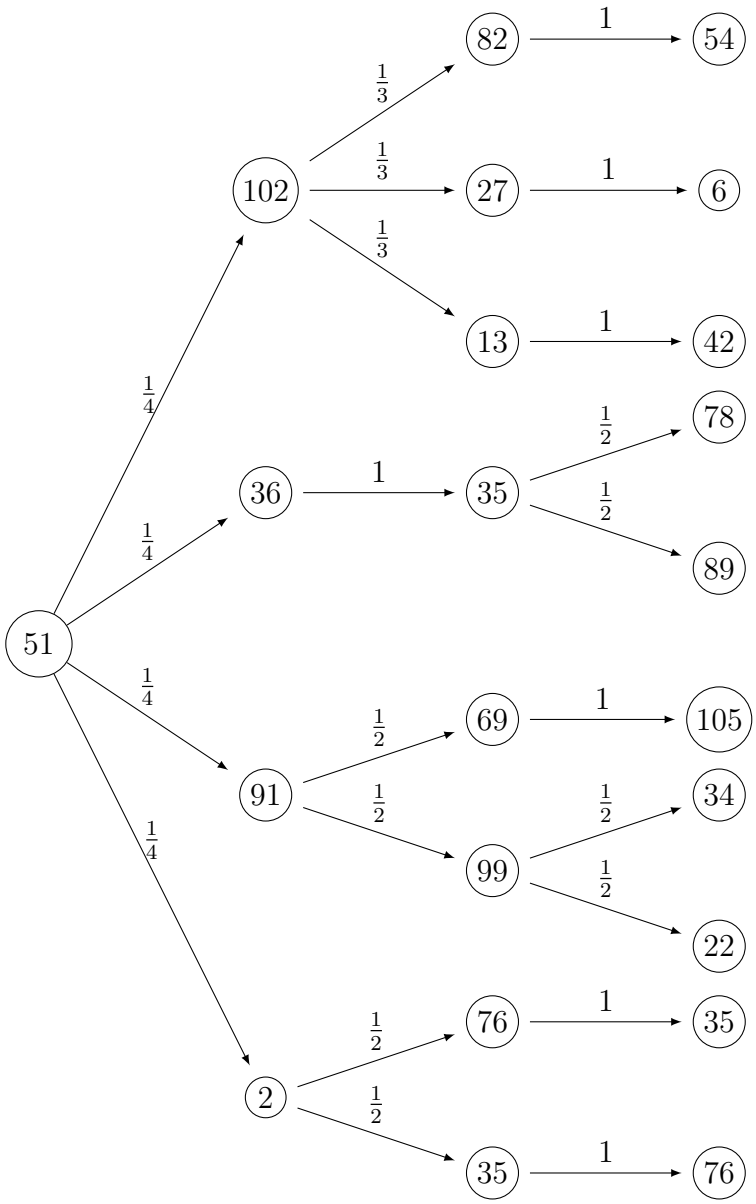


Figure 3.5: Tree diagrams showing all possible EPI samples of size $n = 4$ that can be created at the secondary stage for some town i with $N_i = 113$, given that household 51 is selected first. As an example, the probability of obtaining households 91, 69 and 105, given that household 51 was selected first, is $P((91, 69, 105)|51) = \frac{1}{4} \cdot \frac{1}{2} \cdot 1 = \frac{1}{8}$.

Chapter 4

Two-Stage Sampling and Statistical Analysis

There are various problems associated with the EPI method, particularly within the secondary stage as mentioned in Chapter 1. For this reason, many articles exploring the EPI method do not examine both the first and second stage. However, a two-stage analysis may lead to tractable results.

In the work of Lemeshow and Robinson (1985), an estimator for the proportion of the population vaccinated was developed, along with the standard error of the estimator. Let δ be the true proportion of the population vaccinated and n be the number of children sampled from each of the sampled towns. Then according to Lemeshow and Robinson (1985),

$$\hat{\delta} = \sum_{i \in s_m} \sum_{j \in s'_n} \frac{y_{ij}}{mn}$$

$$SE(\hat{\delta}) = \left[\frac{1}{m(m-1)} \sum_{i \in s_m} (\hat{\delta}_i - \hat{\delta})^2 \right]^{\frac{1}{2}},$$

where $\hat{\delta}_i$ is the estimator for the proportion of the population vaccinated in town i and

$$y_{ij} = \begin{cases} 1, & \text{if household } j \text{ from town } i \text{ has a vaccinated eligble participant;} \\ 0, & \text{if household } j \text{ from town } i \text{ has a non-vaccinated eligble participant.} \end{cases} \tag{4.1}$$

The advantage of using this expression is that it is easily calculable compared to traditional cluster sampling formulas. However, this estimator could be generalized so that it accounts for the probability of including specific households from towns and even towns themselves.

For the studies that follow, assume that every household contains an eligible participant. Further assume that only binary observations, y_{ij}, with outcomes 0 or 1 are collected and studied from the eligible participants. Let the prevalence, which is the proportion of diseased individuals in the population, and the population total of diseased individuals be given respectively by

$$\mu = \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij} \tag{4.2}$$

$$\begin{aligned} \tau &= \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij} \qquad (4.3) \\ &= N\mu. \end{aligned}$$

Due to the fact that the population parameters are proportional to one another by a constant N, once a group of statistics have been derived for one of the parameter estimates, the statistics for the other parameter's estimates are simple to produce. That is, once a group of statistics are derived for estimators for prevalence, another group of statistics can be found for estimators of the population total of diseased individuals. The next set of derivations focus on finding the expected value and variance for any linear estimator of μ. To begin, let the general linear estimator for the population proportion of diseased individuals be defined as

$$\hat{\mu} = \sum_{i \in s_{(m)}} \sum_{j \in s'_{(n_i)}} a_i b_{ij} y_{ij} \tag{4.4}$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{N_i} a_i b_{ij} y_{ij} U_i V_{ij}, \tag{4.5}$$

where a_i and b_{ij} are constants and

$$U_i = \begin{cases} 1, & \text{if town } i \text{ is included in the primary sample;} \\ 0, & \text{otherwise;} \end{cases} \tag{4.6}$$

$$V_{ij} = \begin{cases} 1, & \text{if household } j \text{ from town } i \text{ is included in the secondary sample, given} \\ & \text{that town } i \text{ is included in the primary sample;} \\ 0, & \text{otherwise.} \end{cases} \tag{4.7}$$

Note, Equation (4.4) implies that the number of households sampled from each sampled town can vary and does not need to be of constant size n.

4.1 Expected Value and Variance

Two statistics of general interest using the EPI method are the bias and the mean square error of estimators. However, the bias is a function of the expected value of the estimator and the mean square error is a function of the bias and variance of the estimator. For this reason, formulas to calculate the expected value and variance of the estimator in Equation (4.5) are critical.

Suppose that X and Y are two random variables with the joint probability density function, PDF, $f(x, y)$. Let $h(x, y)$ denote another joint PDF of X and Y. Then,

$$\begin{aligned}
\mathbb{E}(h(X,Y)) &= \int\int h(x,y)f(x,y)dxdy \\
&= \int \left(\int h(x,y)f(x,y)dy \right) dx \\
&= \int \left(\int h(x,y)\frac{f(x,y)}{f(x)}f(x)dy \right) dx \\
&= \int \left(\int h(x,y)f(y|x)dy \right) f(x)dx \\
&= \int \mathbb{E}(h(X,Y)|X)f(x)dx \\
&= \mathbb{E}[\mathbb{E}(h(X,Y)|X)]. \qquad (4.8)
\end{aligned}$$

In the computation for the formulas involving the expected value and the variance of the two-stage estimator found in Equation (4.5), Equation (4.8) is very important. This leads to a simpler way of computing the expected value of $\hat{\mu}$ in Equation (4.5) by taking the double expectation which is conditional on the random variable U_i.

Thus,

$$\begin{aligned}
\mathbb{E}(\hat{\mu}) &= \mathbb{E}\left(\mathbb{E}(\hat{\mu}|U_i)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\sum_{i=1}^{M}\sum_{j=1}^{N_i} a_i b_{ij} y_{ij} U_i V_{ij} | U_i\right)\right) \\
&= \mathbb{E}\left(\sum_{i=1}^{M} a_i U_i \sum_{j=1}^{N_i} \mathbb{E}(b_{ij} y_{ij} V_{ij} | U_i)\right) \\
&= \mathbb{E}\left(\sum_{i=1}^{M} a_i U_i \sum_{j=1}^{N_i} b_{ij} y_{ij} \mathbb{E}(V_{ij} | U_i)\right) \\
&= \sum_{i=1}^{M} a_i \mathbb{E}(U_i) \sum_{j=1}^{N_i} b_{ij} y_{ij} \mathbb{E}(\mathbb{E}(V_{ij} | U_i)) \\
&= \sum_{i=1}^{M} a_i \mathbb{E}(U_i) \sum_{j=1}^{N_i} b_{ij} y_{ij} \mathbb{E}(V_{ij}). \qquad (4.9)
\end{aligned}$$

Above, the two expectations needed to be calculated are $\mathbb{E}(U_i)$ and $\mathbb{E}(V_{ij})$. However, both equations contain binary random variables. Let u_i be a realization of the random variable U_i, such that $u_i = \{0, 1\}$. Thus,

$$\begin{aligned}
\mathbb{E}(U_i) &= \sum_{u_i=0}^{1} u_i P(U_i = u_i) \\
&= 1 \cdot P(\text{town } i \text{ is included in the primary sample}) + 0 \\
&= \alpha_i \qquad (4.10)
\end{aligned}$$

In a similar manner, let v_{ij} be the realization of the random variable V_{ij}, such

that $v_{ij} = \{0, 1\}$. So,

$$\begin{aligned}\mathbb{E}(V_{ij}) &= \sum_{v_{ij}=0}^{1} v_{ij} P(V_{ij} = v_{ij}) \\ &= \beta_{ij} \end{aligned} \tag{4.11}$$

Using Equations (4.10) and (4.11), Equation (4.9) reduces to

$$\begin{aligned}\mathbb{E}(\hat{\mu}) &= \sum_{i=1}^{M} a_i \mathbb{E}(U_i) \sum_{j=1}^{N_i} b_{ij} y_{ij} \mathbb{E}(V_{ij}) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{N_i} \alpha_i \beta_{ij} a_i b_{ij} y_{ij} \end{aligned} \tag{4.12}$$

A useful intermediate quantity that will be used to calculate the variance is

$$\mathbb{E}(\hat{\mu}|U_i) = \sum_{i=1}^{M} \sum_{j=1}^{N_i} a_i b_{ij} y_{ij} U_i \beta_{ij}.$$

From the standard definition of the variance,

$$\begin{aligned} Var(\hat{\mu}) &= \mathbb{E}\left[(\hat{\mu} - \mathbb{E}(\hat{\mu}))^2\right] \\ &= \mathbb{E}\left[((\hat{\mu} - \mathbb{E}(\hat{\mu}|U_i)) + (\mathbb{E}(\hat{\mu}|U_i) - \mathbb{E}(\hat{\mu})))^2\right] \\ &= \mathbb{E}\left[(B_1 + B_2)^2\right] \\ &= \mathbb{E}(B_1^2) + 2\mathbb{E}(B_1 B_2) + \mathbb{E}(B_2^2), \end{aligned} \tag{4.13}$$

where

$$\begin{aligned} B_1 &= \hat{\mu} - \mathbb{E}(\hat{\mu}|U_i) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{N_i} a_i b_{ij} y_{ij} U_i (V_{ij} - \beta_{ij}) \end{aligned} \tag{4.14}$$

$$\begin{aligned} B_2 &= \mathbb{E}(\hat{\mu}|U_i) - \mathbb{E}(\hat{\mu}) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{N_i} a_i b_{ij} y_{ij} \beta_{ij} (U_i - \alpha_i). \end{aligned} \tag{4.15}$$

Analyzing each term in Equation (4.13), a simpler expression could be produced. Let $D_i = \sum_{j=1}^{N_i} b_{ij} y_{ij} (V_{ij} - \beta_{ij})$ and note $U_i^2 = U_i$. Analyzing the first term yields

$$\begin{aligned} \mathbb{E}(B_1^2) &= \mathbb{E}\left(\left[\sum_{i=1}^{M}\sum_{j=1}^{N_i} a_i U_i D_i\right]^2\right) \\ &= \mathbb{E}\left(\sum_{i=1}^{M} a_i^2 U_i D_i^2 + 2\sum\sum_{i<k} a_i U_i D_i a_k U_k D_k\right) \\ &= \sum_{i=1}^{M} a_i^2 \mathbb{E}(U_i D_i^2) + 2\sum\sum_{i<k} a_i a_k \mathbb{E}(U_i U_k D_i D_k). \end{aligned} \tag{4.16}$$

Computing the expectations of Equation (4.16) separately,

$$\begin{aligned} \mathbb{E}(U_i D_i^2) =& \mathbb{E}[\mathbb{E}(U_i D_i^2|U_i)] \\ =& \mathbb{E}[\mathbb{E}(U_i D_i^2|U_i)] \\ =& \mathbb{E}[U_i \mathbb{E}(D_i^2|U_i)] \\ =& \mathbb{E}\left[U_i\left(\mathbb{E}\left(\sum_{j=1}^{N_i}(b_{ij}^2 y_{ij}^2 (V_{ij} - \beta_{ij})^2)|U_i\right)\right.\right. \end{aligned}$$

$$
\begin{aligned}
&+2\sum\sum_{j<l}\mathbb{E}(b_{ij}y_{ij}b_{il}y_{il}(V_{ij}-\beta_{ij})(V_{il}-\beta_{il})|U_i)\Bigg)\Bigg]\\
=&\mathbb{E}\left[U_i\left(\sum_{j=1}^{N_i}(b_{ij}^2y_{ij}^2\mathbb{E}((V_{ij}-\beta_{ij})^2)|U_i)\right.\right.\\
&\left.\left.+2\sum\sum_{j<l}b_{ij}y_{ij}b_{il}y_{il}\mathbb{E}((V_{ij}-\beta_{ij})(V_{il}-\beta_{il})|U_i)\right)\right]\\
=&\mathbb{E}\left[U_i\left(\sum_{j=1}^{N_i}b_{ij}^2y_{ij}^2\beta_{ij}(1-\beta_{ij})+2\sum\sum_{j<l}b_{ij}y_{ij}b_{il}y_{il}(\beta_{i,jl}-\beta_{ij}\beta_{il})\right)\right]\\
=&\alpha_i\left(\sum_{j=1}^{N_i}b_{ij}^2y_{ij}^2\beta_{ij}(1-\beta_{ij})+2\sum\sum_{j<l}b_{ij}y_{ij}b_{il}y_{il}(\beta_{i,jl}-\beta_{ij}\beta_{il})\right), \quad (4.17)
\end{aligned}
$$

where $\beta_{i,jl}$ is defined as the joint inclusion probability of including h_j and h_l from town i in the secondary sample, given that town i was selected in the primary sample. It can be expressed by

$$\beta_{i,jl}=\sum_{s'_{(n_i)}\in S'_{(n_i)}}P(s'_{(n_i)})\mathbb{1}(\{j,l\}\in s'_{(n_i)}), \quad (4.18)$$

where

$$\mathbb{1}(\{j,l\}\in s'_{(n_i)})=\begin{cases}1, & \text{if household's } j \text{ and } l \text{ from town } i \text{ are both included in the secondary sample, given that town } i \text{ is in the primary sample;}\\ 0, & \text{otherwise.}\end{cases}$$

Similarly, the expectation for the second term of Equation (4.16) could be reduced

to

$$\begin{aligned}\mathbb{E}(U_iU_kD_iD_k) &= \mathbb{E}(\mathbb{E}(U_iU_kD_iD_k|U_i)) \\ &= \mathbb{E}(U_iU_k\mathbb{E}(D_iD_k|U_i)) \\ &= \mathbb{E}[U_iU_k\mathbb{E}(D_i|U_i)\mathbb{E}(D_k|U_i)] \\ &= 0, \end{aligned} \tag{4.19}$$

since D_i and D_k are independent, and $\mathbb{E}(D_k|U_i) = 0$. Substituting Equations (4.17) and (4.19) into Equation (4.16),

$$\mathbb{E}(B_1^2) = \sum_{i=1}^{M} a_i^2\alpha_i \sum_{j=1}^{N_i} b_{ij}^2 y_{ij}^2 \beta_{ij}(1-\beta_{ij}) + 2\sum_{i=1}^{M} a_i^2\alpha_i \sum\sum_{j<l} b_{ij}y_{ij}b_{il}y_{il}(\beta_{i,jl} - \beta_{ij}\beta_{il}). \tag{4.20}$$

The second term in Equation (4.13) is the simplest of the three to analyze. Disregarding the constant and finding the expected value of this term with respect to U_i yields

$$\begin{aligned}\mathbb{E}(B_1B_2) &= \mathbb{E}[\mathbb{E}(B_1B_2|U_i)] \\ &= \mathbb{E}\left(\mathbb{E}\left[(\hat{\mu} - \mathbb{E}(\hat{\mu}|U_i)) \cdot (\mathbb{E}(\hat{\mu}|U_i) - \mathbb{E}(\hat{\mu}))|U_i\right]\right) \\ &= \mathbb{E}(\mathbb{E}\left[(\hat{\mu} - \mathbb{E}(\hat{\mu}|U_i))||U_i\right]) \cdot (\mathbb{E}(\hat{\mu}|U_i) - \mathbb{E}(\hat{\mu})) \\ &= [\mathbb{E}(\hat{\mu}||U_i) - \mathbb{E}(\hat{\mu}|U_i)] \cdot (\mathbb{E}(\hat{\mu}|U_i) - \mathbb{E}(\hat{\mu})) \\ &= 0 \end{aligned} \tag{4.21}$$

The third and final term in Equation (4.13) only contains one random variable which is U_i. It allows for a simpler interpretation because there is no need to take a conditional expected value, as was required in Equation (4.18). Nonetheless, following a similar analysis,

$$\begin{aligned}
\mathbb{E}(B_2^2) =& \mathbb{E}\left[\left(\sum_{i=1}^{M} a_i(U_i-\alpha_i)\sum_{j=1}^{N_i} b_{ij}y_{ij}\beta_{ij}\right)^2\right] \\
=& \mathbb{E}\left[\left(\sum_{i=1}^{M} a_i C_i(U_i-\alpha_i)\right)^2\right] \\
=& \mathbb{E}\left[\sum_{i=1}^{M} a_i^2(C_i)^2(U_i-\alpha_i)^2 \right. \\
& \left. +2\sum\sum_{i<k} a_i a_k C_i C_k (U_i-\alpha_i)(U_k-\alpha_k)\right] \\
=& \sum_{i=1}^{M} a_i^2(C_i)^2\mathbb{E}[(U_i-\alpha_i)^2] \\
& + 2\sum\sum_{i<k} a_i a_k C_i C_k \mathbb{E}[(U_i-\alpha_i)(U_k-\alpha_k)] \\
=& \sum_{i=1}^{M} a_i^2(C_i)^2\alpha_i(1-\alpha_i) \\
& + 2\sum\sum_{i<k} a_i a_k C_i C_k(\alpha_{ik}-\alpha_i\alpha_k),
\end{aligned} \tag{4.22}$$

where α_{ik} is the joint inclusion probability of including town i and town k in the primary sample, and $C_i = \sum_{j=1}^{N_i} b_{ij}y_{ij}\beta_{ij}$. Much like Equation (4.18), α_{ik} can be expressed by

$$\alpha_{ik} = \sum_{s_{(m)}\in S_{(m)}} P(s_{(m)})\mathbb{1}(\{i,k\}\in s_{(m)}), \tag{4.23}$$

where

$$\mathbb{1}(\{i,k\} \in s_{(m)}) \begin{cases} 1 & \text{, if both towns } i \text{ and } k \text{ are included in the primary sample} \\ 0 & \text{, otherwise} \end{cases}.$$

Using the reduced expressions for the three terms in Equation (4.13), a formula for the variance of the two-stage estimator is

$$\begin{aligned} Var(\hat{\mu}) = &\sum_{i=1}^{M} a_i^2 \alpha_i \sum_{j=1}^{N_i} b_{ij}^2 y_{ij}^2 \beta_{ij}(1-\beta_{ij}) + 2\sum_{i=1}^{M} a_i^2 \alpha_i \sum\sum_{j<l} b_{ij} y_{ij} b_{il} y_{il} (\beta_{i,jl} - \beta_{ij}\beta_{il}) \\ &+ \sum_{i=1}^{M} a_i^2 (C_i)^2 \alpha_i (1-\alpha_i) + 2\sum\sum_{i<k} a_i a_k C_i C_k (\alpha_{ik} - \alpha_i \alpha_k). \end{aligned} \tag{4.24}$$

Recall, the expected value and variance for the estimator of τ is simply going to be a function directly proportional to Equations (4.12) and (4.24) by the constant N. Specifically,

$$\hat{\tau} = N\hat{\mu}$$

$$\mathbb{E}(\hat{\tau}) = N\mathbb{E}(\hat{\mu})$$

$$Var(\hat{\tau}) = N^2 Var(\hat{\mu})$$

Under the EPI sampling method, a constant number of households are selected from each sampled town. This implies there will be n households sampled from each selected town, where $1 \leq n \leq \min_{i \in \{1,\ldots,M\}} N_i$. Traditionally, the EPI method samples 7 or 30 households from each sampled town, as suggested by the WHO (2008). Commonly, the parameter of interest in rapid assessment surveys using the EPI is

vaccination coverage and prevalence. For this exact reason, this thesis will emphasize assessing the accuracy and precision of several estimators for μ. The statistical associations of these assessments are the bias and mean square error (MSE), respectively. The equations for the bias and MSE are

$$Bias(\hat{\mu}) = \mathbb{E}(\hat{\mu}) - \mu; \tag{4.25}$$

$$MSE(\hat{\mu}) = Var(\hat{\mu}) + Bias(\hat{\mu})^2. \tag{4.26}$$

4.2 Types of estimators for μ

There are a number of estimators that could be produced for μ depending on the constants a_i and b_{ij}, as well as the sampling methods used at both the primary and secondary stages. This paper will explore two general estimators, which can produce several other estimators depending on the sampling methods performed at the primary and secondary stages respectively. As proposed by Lemeshow and Robinson (1985), one possible estimator for μ is to give each household and town an equal weight (EW) at their respective levels of sampling. To be more precise,

$$\begin{aligned}\hat{\mu}_{EW} &= \sum_{i \in s_{(m)}} \frac{1}{m} \sum_{j \in s'_{(n)}} \frac{1}{n} y_{ij} \\ &= \sum_{i=1}^{M} \sum_{j=1}^{N_i} \frac{1}{m} \frac{1}{n} U_i V_{ij} y_{ij}\end{aligned} \tag{4.27}$$

Equation (4.27) shows that an EW estimator for μ is a case where the constants $a_i = \frac{1}{m}$ and $b_{ij} = \frac{1}{n}$ in Equation (4.5). Although the constants for the EW have been

defined, the inclusion probabilities at the town and household levels will depend on the method of sampling. Consequently, there are still many possible variations for the EW estimator $\hat{\mu}_{EW}$ leading to various estimates for the expected value and variance.

The second general estimator that will be studied is the Horvitz-Thompson (HT) estimator. It is a weighted average that uses the reciprocal of the inclusion probabilities of elements in the sample. It can be represented by

$$\begin{aligned}\hat{\mu}_{HT} &= \frac{1}{N}\sum_{i\in s_{(m)}}\sum_{j\in s'_{(n)}}\frac{y_{ij}}{\pi_{ij}} \\ &= \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\frac{1}{\alpha_i}\frac{1}{\beta_{ij}}U_iV_{ij}y_{ij}, \end{aligned} \tag{4.28}$$

where $a_i = \frac{1}{N\alpha_i}$, $b_{ij} = \frac{1}{\beta_{ij}}$ and

π_{ij} =P(Household j from town i is included in the sample)

=P(Household j from town i is included in the sample, given town i is included)

$\times$ P(Town i is included in the sample)

$= \beta_{ij}\alpha_i$.

For details on the HT estimator, see Lohr (2009). A unique characteristic of the HT estimator is that it is always unbiased when the method of sampling used to define the constants a_i and b_{ij} are the same as the method of collecting samples at the primary and secondary stage. As an example, say the constants a_i and b_{ij} are the reciprocal of the inclusion probabilities using systematic sampling and strip

sampling, respectively. If systematic sampling and strip sampling were used to collect the samples at the primary and secondary stage, then it is trivial to see that this particular HT estimator is unbiased.

Sampling WOR using PPS will not be used when developing estimators for μ due to computational constraints. Specifically, when looking at $Var(\hat{\mu})$, a_{ik} are needed and calculating a_{ik} is not trivial when sampling WOR with PPS. Moreover, the number of samples of size $m \geq 2$ containing PSUs i and k could be generalized and given by the following equation

$$ {}_{M}P_{m} - \left({}_{(M-2)}P_{m} + 2 \times \left({}_{(M-2)}P_{(m-1)} \times \binom{m}{1} \right) \right). \tag{4.29} $$

In lay terms, Equation (4.29) represents the number of ways that all samples of size m could be produced WOR, excluding those samples that do not contain both PSUs i and k or contain only PSU i or only PSU k. For illustration, say there were 30 towns and 15 were meant to be sampled WOR with PPS. Then, there are approximately 4.8961×10^{19} samples of size 15 that contain PSUs i and k, for $i \neq k$. Since there are no elegant closed form equations to compute a_{ik} while sampling WOR with PPS, Equation (4.23) must be used. This brings forth a computational disadvantage because computing these large number of sample probabilities is no simple task, even for computers. For reasons mentioned, sampling WOR using PPS will not be considered when constructing estimators for μ.

Accordingly, the only primary sampling methods that can be used from Chapter 2 to create estimators for μ are systematic sampling and sampling WR using PPS. Systematic sampling is used because there is an exact method to compute inclusion

probabilities of each PSU. Accordingly, computing joint inclusion probabilities is easy and efficient to implement using computer software. Sampling WR using PPS is used for the exact same reasons; however, it has another advantage of having an equation for joint inclusion probabilities. Specifically, when sampling WR with PPS, the joint inclusion probability of obtaining towns i and k in a sample of size m using the inclusion-exclusion principle is

$$\alpha_{ik} = \alpha_i + \alpha_k - \left[1 - \left(1 - \frac{N_i + N_k}{N}\right)^m\right], \tag{4.30}$$

for all $i \neq k$. Furthermore, only the strip method will be used at the secondary stage of sampling. If other sampling methods were introduced at the secondary stage, which could occur, it would only make the study harder to process and analyze.

To distinguish the inclusion probabilities that could be produced, let $\alpha_{i,SYS}$ and $\alpha_{i,WR}$ be the inclusion probabilities generated for town $i \in \{1, \ldots, M\}$ using systematic sampling and sampling WR with PPS respectively. Clearly, there are numerous estimators that can be explored. To organize and present the possible estimators, Table 4.1 has been constructed:

Table 4.1: Estimators for μ, depending on constants a_i and b_{ij} and the primary sampling method.

(a) Sampling PSUs systematically.

Estimator	a_i	b_{ij}
$\hat{\mu}_{EW_1}$	$\frac{1}{m}$	$\frac{1}{n}$
$\hat{\mu}_{HTSRS_1}$	$\frac{M}{N \cdot m}$	$\frac{1}{\beta_{ij}}$
$\hat{\mu}_{HTSYS}$	$\frac{1}{N \cdot \alpha_{i,SYS}}$	$\frac{1}{\beta_{ij}}$

(b) Sampling PSUs WR with PPS.

Estimator	a_i	b_{ij}
$\hat{\mu}_{EW_2}$	$\frac{1}{m}$	$\frac{1}{n}$
$\hat{\mu}_{HTSRS_2}$	$\frac{M}{N \cdot m}$	$\frac{1}{\beta_{ij}}$
$\hat{\mu}_{HTWR}$	$\frac{1}{N \cdot \alpha_{i,WR}}$	$\frac{1}{\beta_{ij}}$

Chapter 5

Estimations on Simulated Populations

In the previous chapters, inclusion probabilities have been developed for both primary and secondary stages of the EPI method so that they can be utilized to evaluate estimators for the population prevalence μ. Another component needed to study the modified EPI method proposed in this paper and the estimators in Table 4.1 is population data.

Past studies on the EPI method obtained data for their analyses through a variety of ways. Some researchers performed the EPI method in actual field work and were able to collect real population data. For example, Bennett et al. (1994) and Katz et al. (1997) conducted studies on the EPI method in 30 selected communities in Uganda and 40 randomly selected communities in Nepal, respectively. More recent studies performed by Reyes (2016) generated populations using computer simulations to study the secondary stage of the EPI method.

To do something less traditional to previous works, a new method to simulate populations using the Google Maps API will be introduced. The aim is to appeal and promote future researchers of the EPI method to focus on this simulation method and study the EPI in a more realistic setting, when real world data is not readily accessible.

5.1 Simulating Populations

5.1.1 Generating Households

One particular way of studying the EPI method is to generate a variety of populations with unique spatial characteristics. In her thesis, Reyes (2016) generated a variety of two dimensional towns with different spatial structures using the work of Bolker (2008). The origin of the Cartesian plane, $(0, 0)$, was assigned as the town's center and the simulated towns were created so that no household could be randomly placed at this particular point. The simulated towns are all contained in some rectangular region which helps control the density of towns. For some arbitrary town i, N_i households were generated with the following four general spatial patterns:

1. Random Pattern - (*runi*). Households were generated using a bivariate uniform distribution and mapped to a two dimensional plane with respect to the dimensions of the town. A special case arises when the length of the dimensions of the town are the same, which produces a square town with randomly uniformly placed households.

2. Grid Pattern - (*grid*). Households were divided into regularly spaced rows and

columns, such that roughly N_i households could be mapped onto the town. Subsequently, adjacent houses were all equidistant from any other adjacent houses.

3. Aggregated Pattern - (*aggr*). Households were distributed into k groups, such that the sum of the number of households across all groups was N_i. Each group had some focal point, and households were dispersed around it.

4. Circular Gradient Pattern - (*circ*). Much like the aggregated pattern, the circular gradient pattern contains houses which are dispersed around a single focal point. However, the concentration of households could be controlled by a rate parameter λ. This allowed the production of towns which were densely packed towards the center, and others that were more spread.

Each of the patterns mentioned above are spatial characteristics observed in real world distributions of houses. In many countries, houses are typically built in rows to reduce costs allocated to building them and to optimize the area used for housing. For this reason, the grid pattern is a reasonable representation for this particular dispersal of houses. The random pattern was used to mimic households found on the countryside, which look randomly scattered with no particular pattern. To represent houses in suburban areas, which are typically composed of communities within a community, the aggregated pattern was used. More exactly, this pattern is used to represent groups of households that are generally clustered together and physical barriers are usually responsible for dividing the particular communities; like roads, highways, large industrial facilities, etc. Finally, to study communities that typically concentrate more around the center of the town, like a downtown geographical region,

the circular gradient pattern was used. For visual aid, Figure 5.1 displays the spatial distributions described above. Note, a package from *R* called *ggplot2* will be used to produce the visualizations of each spatial and disease patterns, as well as assessments.

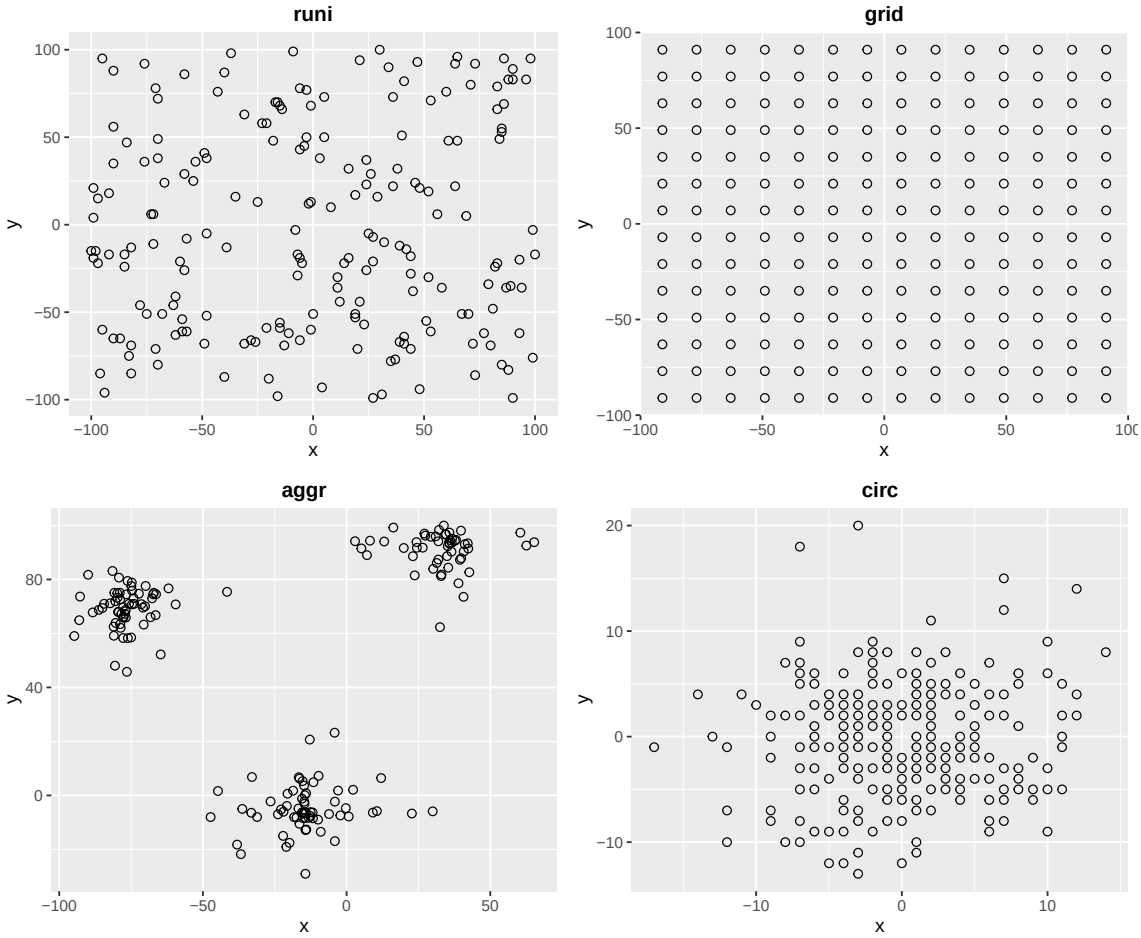


Figure 5.1: Illustration of the four simulated spatial patterns that 196 households in an arbitrary town can take on, projected onto a Cartesian plane.

From Figure 5.1, it can be seen that *runi* and *aggr* spatial patterns are dispersed, such that there is, on average, one nearest neighbour to every household in the town. When looking at the *grid* and *circ* patterns, there are clearly many households that have more than one nearest neighbour. The effect of these spatial patterns on the

calculations of variance for the estimators displayed on Table 4.1 are quite significant. To be more precise, the computational time it takes to calculate these estimators for spatial patterns that are composed of many households with more than one nearest neighbour are typically high.

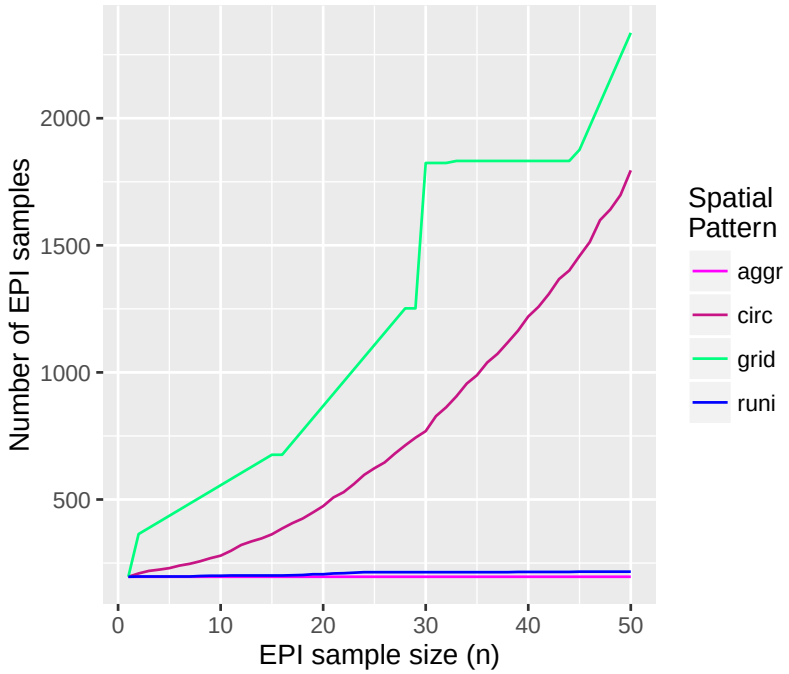


Figure 5.2: A visualization of 200 households generated using a variety of spatial patterns, illustrating the relationship between the EPI sample size n and the number of EPI samples at the secondary stage.

Evidently, it can be seen in Figure 5.2 that when the spatial patterns take on the *grid* and *circ* patterns, the number of EPI samples increases rapdily in a roughly proportional fashion. A much slower proportional increase (with slope equal to 1) is observed when the spatial patterns take on a *aggr* and *runi* pattern. For this reason, the sample of households at the secondary stage should not exceed 30 because for larger sample sizes, the *grid* and *aggr* spatial patterns present computational

difficulties.

5.1.2 Generating Disease Status

Once households have been generated, the next step to complete the simulation of populations for the study of disease prevalence is to assign disease statuses to each of the N households in the M simulated towns. For the study on disease prevalence, each household has four variables: the town the household belongs to, the x_{ij}-coordinate, y_{ij}-coordinate and z_{ij} which is a binary disease status. If a particular member of household j from town i is eligible for the survey and has the disease $z_{ij} = 1$, and if they do not carry the disease $z_{ij} = 0$. For simplicity, it is assumed that every household has exactly one eligible candidate for the survey.

Let the prevalence in town i be denoted by p_i and the prevalence for the entire population be denoted by p. It will be assumed that the prevalence across all towns will be equal and so the population prevalence is simply p. To assign disease status to each household, a Bernoulli random variable Z_{ij} was used and the probability that a household had a candidate with the disease characteristic depends on the disease dispersion pattern (Reyes, 2016). For this reason, Reyes (2016) developed ten disease dispersion patterns that could be implemented to the households in each town, again using the work of Bolker (2008). Only three of these patterns will be studied in this thesis and the mathematical details motivating these patterns are found in Reyes (2016). These three patterns are:

1. Random - *rand*. Given some prevalence level p, each household is equally likely, with probability p, to have the disease characteristic. Larger towns had prevalence closer to p and it was unlikely to generate a town with prevalence exactly

equal to p.

2. Pocketing - *pckt*. Pocketing is a particular disease pattern when groups of households in close proximity in a town exhibit the characteristic of interest. After selecting the number of pockets k, k households are selected using SRS as focal points. Roughly $\frac{N_i \times p}{k} - 1$ households nearest to each focal point are selected such that the prevalence for a particular town is approximately p.

3. Circular Gradient - *cigr*. Much like the circular gradient spatial pattern, households closer to the town center were given higher probabilities of carrying the disease with respect to some prevalence rate p.

Although there were many disease patterns to choose from, they were narrowed down to the three general patterns described above because of how easy they are to observe in real world disease patterns. For instance, many diseases have no particular pattern and appear randomly dispersed throughout a specific population in a geographical region. To mimic this, the random pattern was used. On the other hand, other diseases are spread through physical contact and more often than not, contact is frequently made between neighboring households. Consequently, pocketing was used to display this phenomenon. In addition, when focal points for the pocketing pattern are in close proximity to one another, it creates one very large pocket. For generalization, large pocketing was not specifically analyzed as a pattern on its own. The circular gradient diseased pattern was used to model diseases that generally start at some focal point, say the center of the town, and tend to decrease moving towards the boundaries of the town. A graphical representation of these three disease patterns implemented within the four spatial patterns is displayed in Figure 5.3.

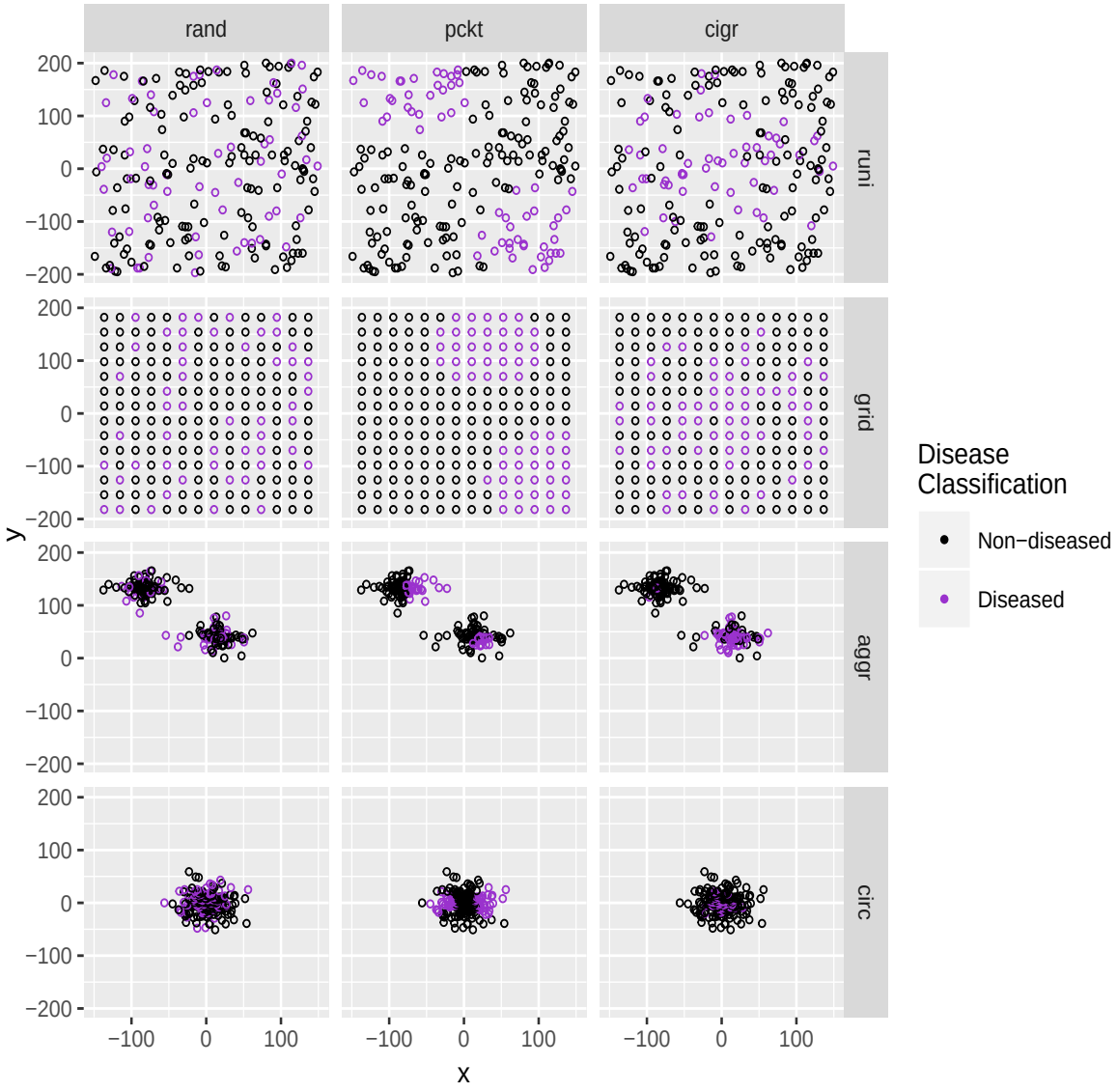


Figure 5.3: A visual representation of the 3 disease patterns that could occur in each of the 4 spatial patterns. Each town had prevalence fixed at $p = 0.3$ and 200 households.

5.2 Evaluating Estimators

5.2.1 Sampling Plans

Aside from spatial and disease patterns used to construct a population, there are a handful of other parameters: the dimensions and prevalence associated to each town, total number of towns, number of towns sampled, number of households sampled and the base of the strip used when executing strip sampling at the secondary stage.

Going from the primary to the secondary level, the first parameter of interest is generating town sizes. To generate town sizes, a Pareto distribution was used and according to the Pareto distribution mentioned in the package of R called *rmutil*, its distribution is of the form:

$$f(u) = \frac{\psi \left(1 + \frac{u}{\phi(\psi-1)}\right)^{-1(\psi+1)}}{\phi(\psi-1)}, \tag{5.1}$$

where $u > 0$, $\phi > 0$ is the mean parameter and $\psi > 0$ is the dispersion (Lindsey, 2017). Since it is a heavy lower tailed distribution, it allows random sampling of values that are more concentrated to the lower tail and less on the upper tail. This feature in application allows the simulation of many small towns and few large towns, as is observed in the real world. Typically, town sizes of 150–500 were desired and to generate towns of this size a Pareto distribution with $\phi = 70$ and $\psi = 100$ were used. After sampling M random numbers using the Pareto distribution with the defined parameters, 150 was added onto each sampled value to ensure the minimum town size was 150 and sizes were truncated to a maximum of 500.

Using the Pareto distribution, $M = 50$ towns were simulated and $m = 30$ are

sampled. Thirty towns were sampled because past studies on the EPI method carried out by Bennett et al. (1994) and Lemeshow and Robinson (1985) have also studied samples of this size, and thirty towns are a recommended sample size in the EPI Manual made by the World Health Organization (2008). The dimensions of the towns also have an impact in the study of the estimators, particularly the inclusion probabilities in the secondary stage. As an example, say each of the town coordinates were scaled by a factor of two and the width of the strip is correspondingly adjusted, the statistical analysis will produce identical results. For this reason, it is assumed that each simulated town will be on a 200×200 unit grid, regardless of town size, spatial and disease patterns or prevalence. Regarding prevalence, there are three levels of interest to study which are $p = \{0.05, 0.2, 0.5\}$. Using prevalences aside from the previously mentioned set is redundant because evaluation of the estimators showed that choosing a prevalence of p and $1 - p$, for $p \in [0, 1]$, displayed similar estimator properties. Three levels of sample sizes at the secondary level were chosen for the study, which are $n_i = \{7, 15, 30\}$ for $i \in \{1, \ldots, M\}$.

To reduce the complexity surrounding simulation plans, it is assumed that each town will have the same spatial and disease patterns as well as prevalence. Overall, there are $4 \times 3 \times 3 \times 3 = 108$ population and sampling settings to explore for each of the six estimators from Table 4.1. If any more factors were considered in the sampling or population plans, the time it would take to process and evaluate these additional simulations would increase proportionally, as shown in Figure 5.2. This meant that the strip with base of size α had to be fixed at one size across the 108 population and sampling plans, which was $\alpha = 5$. A separate study was conducted which analyzed the effect the strip's base size had on the six estimators when the

sampling plan was fixed; selecting $n = 7$ households at the secondary stage and fixing the prevalence in each town to $p = 0.5$. The spatial and disease patterns were still explored at their individual levels in this additional study of the estimators.

5.2.2 Simulation Results

Looking at the distributions of the six proposed estimators under the 108 population and sampling plans, a variety of explorations were carried out. To get a general sense of the assessments of each estimator; summary statistics composed of the minimum, mean and maximum for the bias, variance and MSE were composed in Table 5.1.

Table 5.1: Selected summary statistics for the six proposed estimators from Table 4.1 across 108 population and sampling plans.

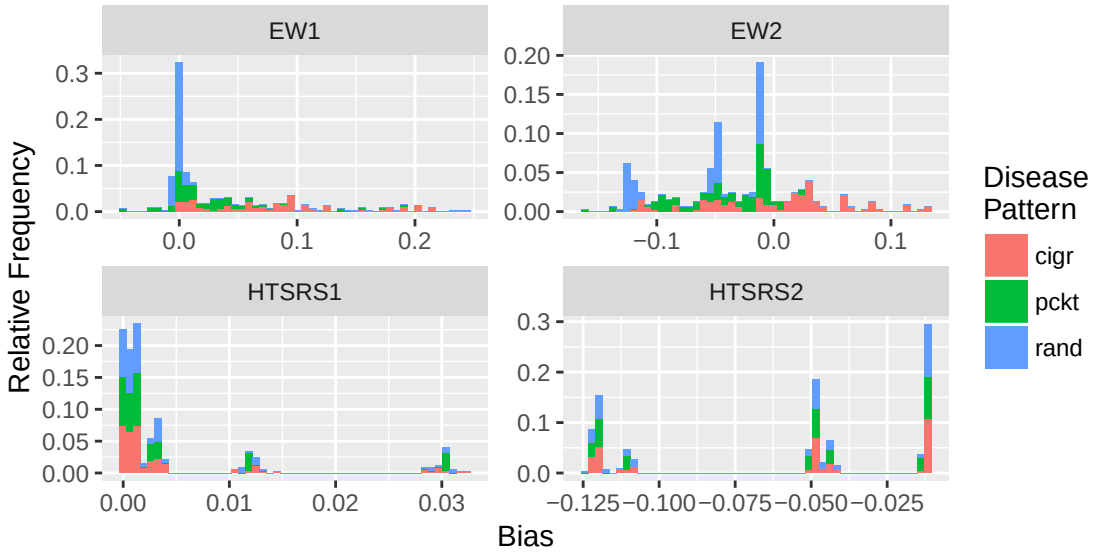
		Bias	Variance	MSE
EW1	Min	-0.05037024	3.8962×10^{-5}	4.03767×10^{-5}
	Mean	0.03768258	0.001375174	0.006226359
	Max	0.241144	0.00681462	0.06161801
EW2	Min	-0.1600543	3.91008×10^{-5}	0.0001570143
	Mean	-0.03321236	0.001519019	0.005698995
	Max	0.1333843	0.005975984	0.03127708
HTSRS1	Min	-0.0001479867	1.82212×10^{-5}	1.823091×10^{-5}
	Mean	0.004488671	0.004045496	0.004135083
	Max	0.03219728	0.02346345	0.02346543
HTSRS2	Min	-0.1236864	2.44383×10^{-5}	0.0001528813
	Mean	-0.05918547	0.003397898	0.008861097
	Max	-0.01067905	0.01865071	0.03326457
HTSYS	Min	0	1.32055×10^{-5}	1.32055×10^{-5}
	Mean	0	0.00383001	0.00383001
	Max	0	0.02321568	0.02321568
HTWR	Min	0	4.0787×10^{-5}	4.0787×10^{-5}
	Mean	0	0.005717246	0.005717246
	Max	0	0.03231401	0.03231401

It can be gathered from Table 5.1 that estimators EW1 and EW2 have the most

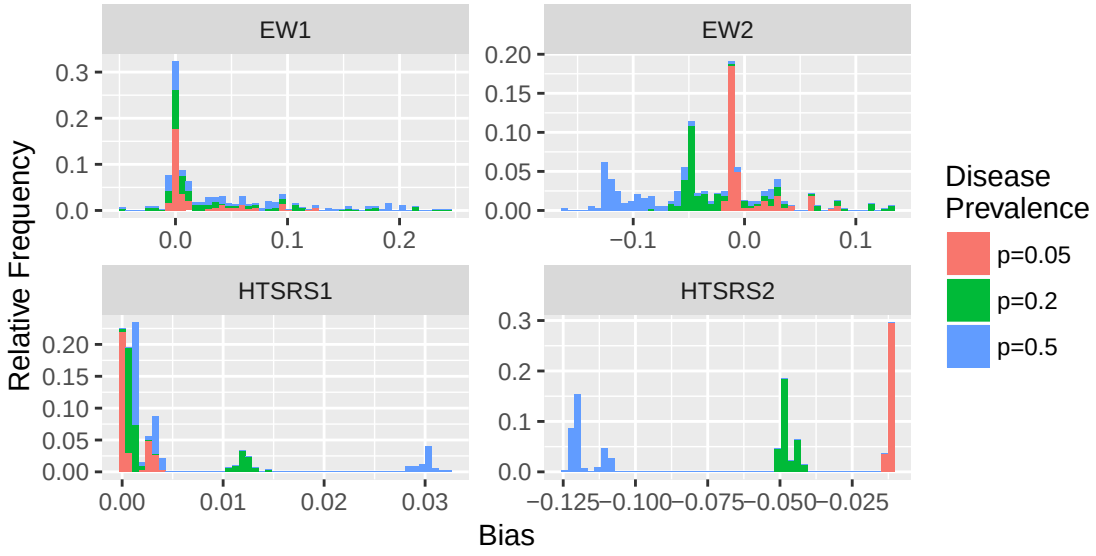
spread in bias. Furthermore, the largest spread in the MSE and largest bias based off the 108 population and sampling plans comes from the EW1 estimator. However, the smallest average and smallest MSE comes from the HTSYS estimator. Observing the HT estimators, it can be seen that under the correct sampling plans the HTSYS and HTWR estimators are indeed unbiased estimators. Consequently, the MSE of these estimators is simply their variance estimates.

To get a visualization of these estimators' spread, relative frequency histograms were constructed for the bias and MSE with respect to the disease patterns and prevalence values as shown in Figures 5.4 and 5.5. The heavy dispersion observed in the EW1 and EW2 estimators is mainly attributed by the populations that follow a central gradient spatial pattern. Comparatively looking at the spread in bias with respect to the prevalence levels shows that the EW2 and HTSRS2 estimators are divided into three groups corresponding to disease prevalence. The general trend within these groups is that the lower the prevalence, the smaller the average bias produced by the EW2 and HTSRS2 estimators. In general, as the prevalence values get closer to 0.5, the spread in bias generally increases across all estimators. Moreover, as the prevalence diverges from $p = 0.5$ the estimators appear to be more unbiased.

Another assessment that could be visually diagnosed is the MSE. Several histograms were studied with respect to a variety of factors from the population and sampling plans, but the prevalence depicted more insightful information regarding the MSE for each estimator. Much like in Figure 5.4b), Figure 5.5 shows three groups formed in the histograms of the EW2 and HTSRS2 estimators with respect to the prevalence values. A recurring pattern observed in Figure 5.5 is that as the prevalence approaches 0.5, the average MSE corresponding with each estimator increases.



(a) Histograms for bias across 4 of the 6 estimators grouped by disease patterns under the remaining 36 population and sampling plans.



(b) Histograms for bias across 4 of the 6 estimators grouped by prevalence values under the remaining 36 population and sampling plans.

Figure 5.4: Bias for the biased estimators proposed in Table 4.1 with respect to Disease Pattern and Prevalence. Note, the bias for the HTSYS and HTWR estimators were trivial to display since they are unbiased under their sampling methods.

However, the EW1 estimator tends to have larger MSE measurements associated with prevalence values farther away from $p = 0.5$.

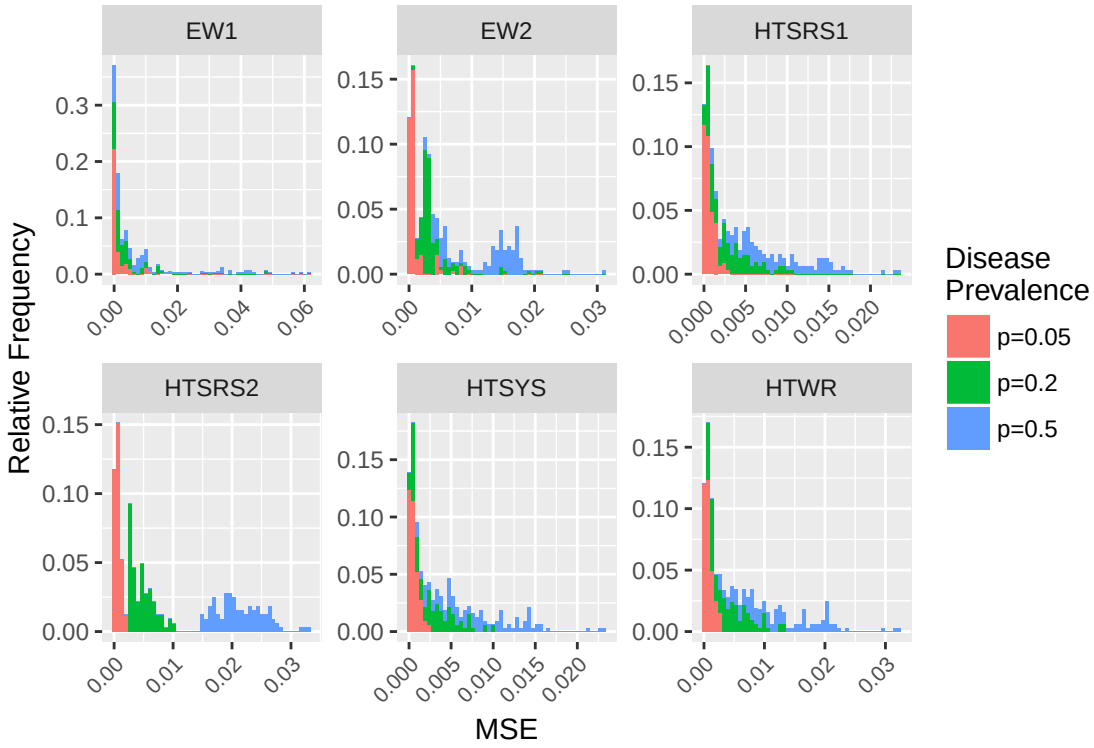


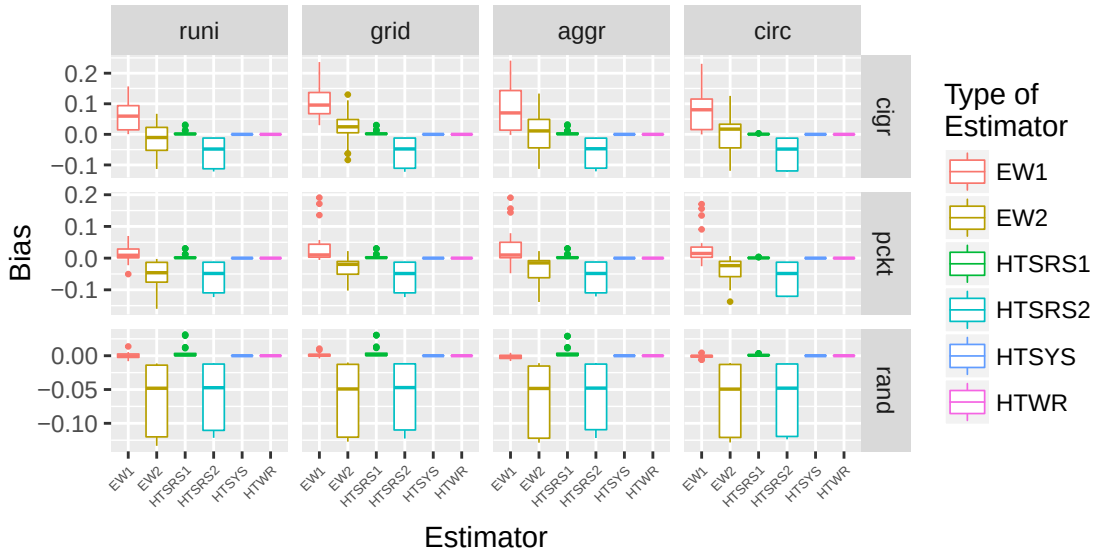
Figure 5.5: A histogram displaying the spread of the MSE across the 6 estimators with respect to the disease prevalence.

Although Table 5.1, Figure 5.4 and 5.5 conclude a variety of results, boxplots of each estimator were produced with respect to two factors in Figures 5.7-5.8 to extract more detailed results regarding the simulations. Studying disease patterns that were simulated using the *rand* pattern, it is clear that the EW1 estimator produces very little bias and the EW2 estimator produces a large spread in bias that underestimated the true prevalence of disease. Interestingly, the HTSRS2 estimator produced the most consistent spread in bias measurements when spatial and disease patterns

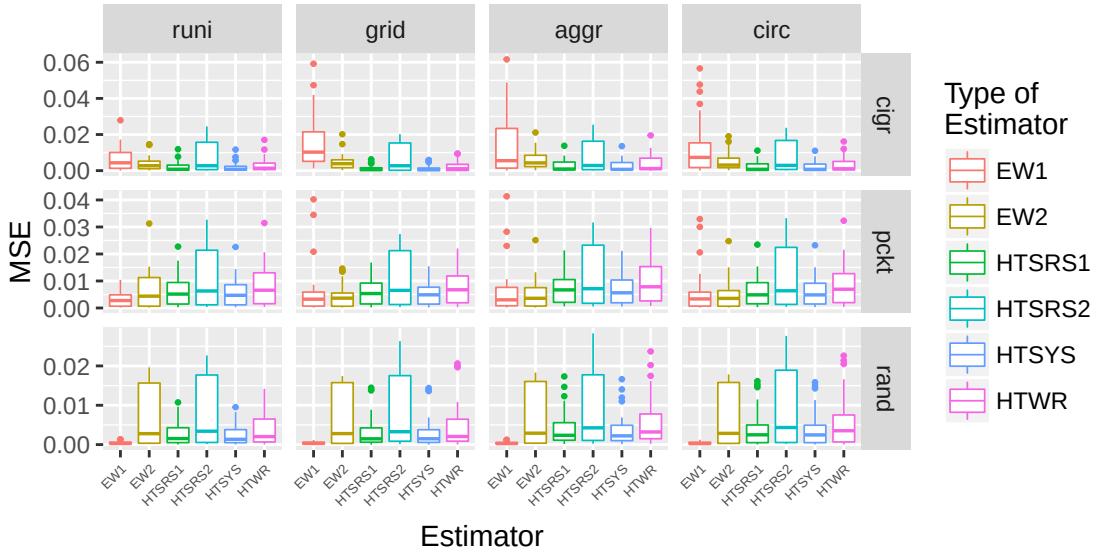
varied, and it always underestimated the true prevalence. A possible justification for this characteristic of the HTSRS2 estimator is its inclusion probabilities' relationship with its defined constant a_i terms. Fixing the spatial patterns and looking at the disease patterns individually, the bias measurements of the HTSRS1 estimator was almost identical across the various disease patterns and deviated very little from the true prevalence.

Conclusions drawn from the plots displaying the spread of the MSE and bias measurements across the six estimators mimicked one another when the spatial and disease patterns were observed simultaneously. When the disease was spread randomly, the EW1 estimator produced the smallest spread in MSE values even though the HTSYS estimator was unbiased. The HTSRS1 and HYSYS estimators produced almost identical spread in MSE measurements, although the HTSYS estimator was unbiased. This may have been attributed to the fact that the primary sampling method used in both estimators was systematic sampling. Even though the HTSRS2 and HTWR estimators both sampled towns WR using PPS, they did not have a similar spread in their bias or MSE measurements.

When studying the estimators with respect to simulated spatial patterns and prevalence levels, the EW estimators' measurements for the bias and MSE were spread throughout each of the individual factor pairings with many extreme measurements. However, the biased HT estimators showed extremely little spread. The HTSRS1 and HTSYS estimators again produce an almost identical spread of MSE measurements. For fixed prevalence values, a change in spatial patterns caused a very minor change in spread of MSE measurements across each estimator respectively.



(a) Boxplots for the bias of the 6 estimators grouped by spatial and disease patterns.



(b) Boxplots for the MSE of the 6 estimators grouped by spatial and disease patterns.

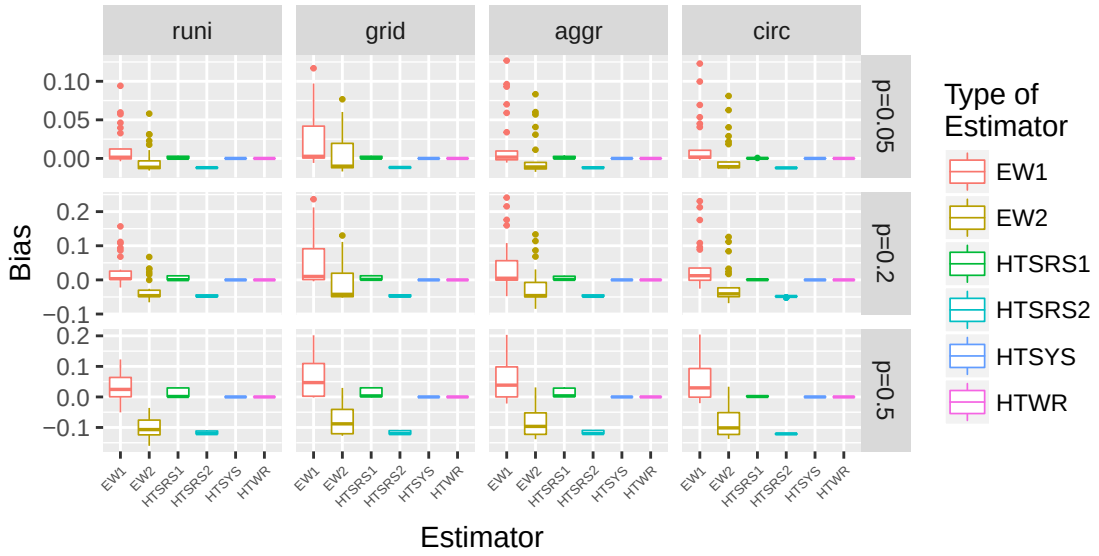
Figure 5.6: Bias and MSE for the estimators proposed in Table 4.1 with respect to spatial and disease patterns.

When the population was simulated using $p = 0.5$ and all spatial patterns excluding the *grid* pattern, the EW1 estimator's spread in MSE measurements produced a median of approximately 0.005. This would imply that around 50% of the data in this particular setting had MSE measurements that were less than or equal to 0.005, indicating the dispersion from the true prevalence was extremely small. A possible justification for this finding is that some measurements in this setting had disease patterns that followed the *rand* pattern, in which case the EW1 has been found to record small MSE measurements.

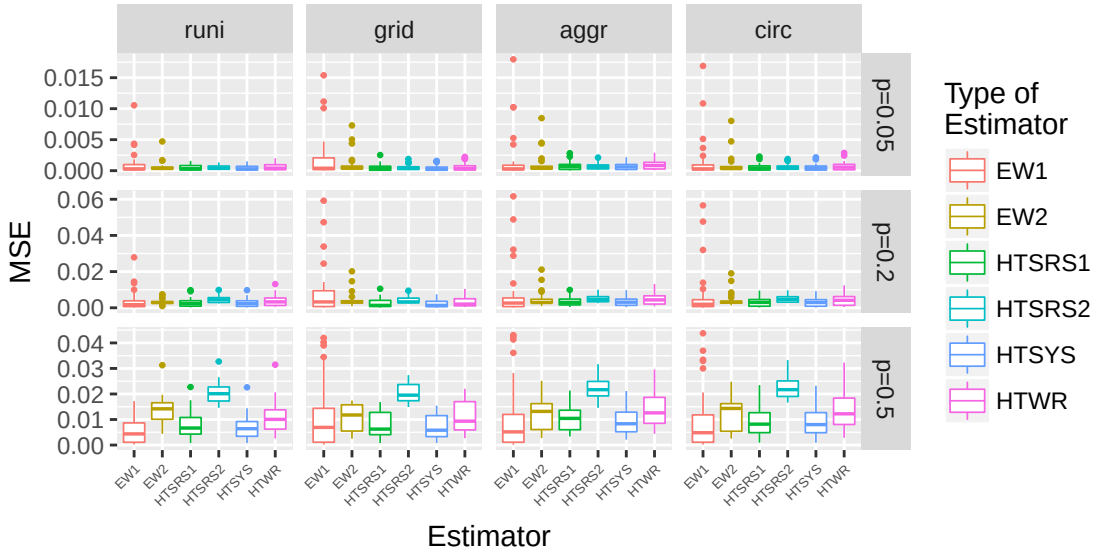
The spread in bias measurements does not change significantly when the prevalence level is fixed and the secondary sample size n increases. However, when the prevalence level is fixed and n increases, the spread in MSE measurements for all estimators decreases considerably. Intuitively, this makes sense because as the sample size increases, a more precise estimate for prevalence can be obtained.

A study on the effect of the strip's base of size α was analyzed across the four spatial and three disease patterns, after fixing the prevalence at $p = 0.5$ and secondary sample size at $n = 7$. The sizes of the base studied were $\alpha = \{0.01, 0.1, 0.5, 1, 2, 5, 10, 25\}$ and a visualization of the study is shown in Figure 5.9. When households are distributed using the *runi* spatial pattern across all simulated towns, the HTSRS1 estimator is shown to produce consistent biased measurements regardless of the size α. Under any other spatial pattern, the HTSRS1 estimator's bias measurements are all very close to 0 for any size α.

When the households are distributed in a circular gradient pattern and the spread of disease is either pocketed or in a central gradient pattern, the EW2 estimator produced smaller bias measurements relative to the EW1 estimator for all size α

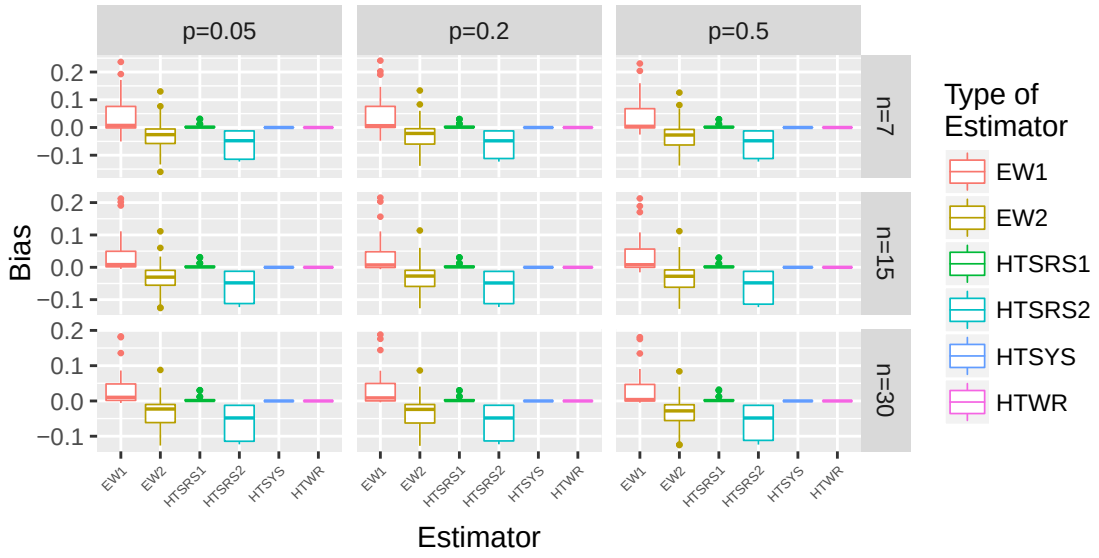


(a) Boxplots for the bias of the 6 estimators grouped by spatial and disease patterns.

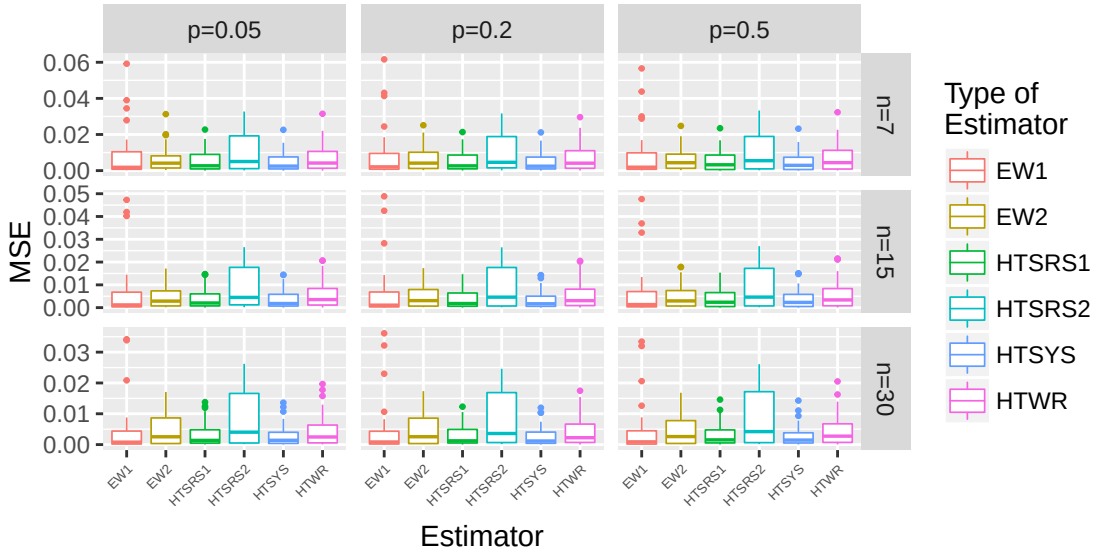


(b) Boxplots for the MSE of the 6 estimators grouped by spatial and disease patterns.

Figure 5.7: Bias and MSE for the estimators proposed in Table 4.1 with respect to spatial patterns and prevalence levels.

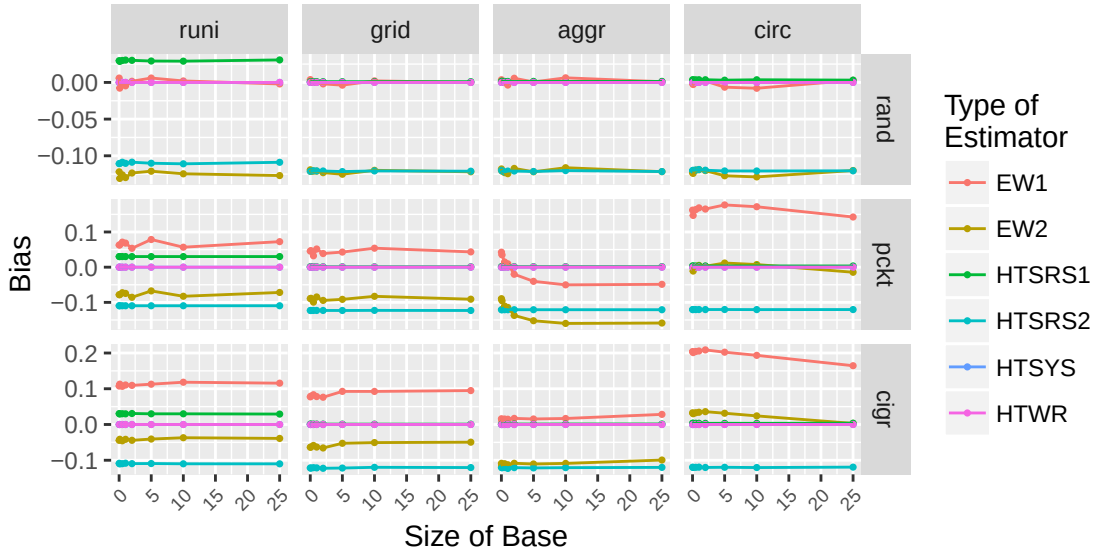


(a) Boxplots for the bias of the 6 estimators grouped by spatial and disease patterns.

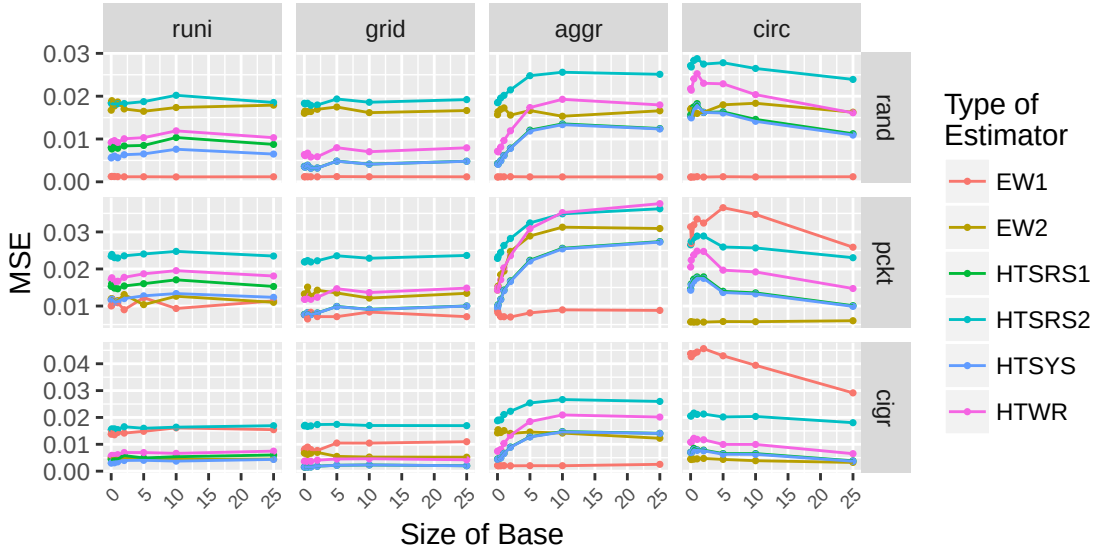


(b) Boxplots for the MSE of the 6 estimators grouped by spatial and disease patterns.

Figure 5.8: Bias and MSE for the estimators proposed in Table 4.1 with respect to household sample size and prevalence levels.



(a) Boxplots for the average bias of the 6 estimators grouped by spatial and disease patterns.



(b) Boxplots for the average MSE of the 6 estimators grouped by spatial and disease patterns.

Figure 5.9: Average bias and MSE for the estimators proposed in Table 4.1 with respect to household sample size and prevalence levels.

studied. In these exact same settings, the EW2 estimator is observed to produce the smallest MSE measurements compared to the other five estimators across all base sizes. Much like previous conclusions drawn when the disease is spread randomly, the EW1 estimator produces accurate estimates (small bias) which are very precise (very small MSE). Households that are distributed under the aggregated gradient pattern with pocketed disease patterns show that the EW estimators displayed a monotonic decrease in bias when α increased; the EW1 estimator producing less biased measurements compared to the EW2 estimator.

Despite increasing α, the HT estimators did not display significant change in their bias measurements across any spatial or disease patterns. The MSE of the HT estimators monotonically increased with respect to α when the households were spread using the *aggr* pattern and pocketing was used to simulate the disease pattern. Conversely, the MSE peaked at a base length of 10 under the random and central gradient disease patterns when households were spread using the *aggr* pattern. At a base of size $\alpha = 0.5$, the MSE measurements for the HT estimators were maximized under the circular gradient pattern and then steadily decreased as α increased. Finally, the HTSYS estimator is seen to produce the very small MSE measurements when the disease was spread using the *cigr* pattern and the households were dispersed using the *runi* and *grid* patterns.

5.3 Internet Resources: Google Maps Geocoding API

In the previous section, an analysis on the six estimators from Table 4.1 was performed under a variety of spatial patterns. However, the towns were created based off simulations and are arguably not perfectly representative of household patterns in the real world. Census unit population information or household inventories is generally not available for public use. However, it is beneficial to assess the EPI method. As expressed by Reyes (2016), a number of surveys were executed in challenging conditions that used Global Positions Systems (GPS) and Geographic Information Systems (GIS) to support the selection of households at the secondary stage.

The shift to using GIS and GPS tools to create a sampling plan has been carried out in Iraq and sub-Sahara Africa (Galway et al., 2012; Pearson et al., 2015). In these studies, household coordinates were recorded manually by embedding placemarkers on tools like Google Earth Pro and used afterwards as a sampling frame to perform surveys. Other surveys in Haiti, for example, had census unit population information provided by government organizations to study malnutrition rates in remote areas Wampler et al. (2013).

Two immediate concerns with the mentioned methods are that: (1) government census information is not generally available for the public and usually requires a lengthy process to access, and (2) it would often take a fair amount of time to record households by hand on tools like Google Earth Pro. An alternative method is to extract real household coordinates by using the Google Maps application program interface (API) and Google Maps Geocoding API function, which allows users to

find addresses and convert them into latitude and longitude coordinates.

5.3.1 Generating a Spatial Sampling Frame

In pursuit of obtaining an EPI sample and using it for estimations in a study, all relevant household coordinates first need to be located to form a sampling frame. Working with the Google Maps API, it is possible to extract relatively new information about a geographical area and tabulate household information using some programming language. This provides the ability to quickly obtain household information and assess the EPI method for some survey regarding a target characteristic.

Due to the fact that this technique is experimental and being introduced as a novel method of studying the EPI method, the applications are rather limited. Specifically, the information that can be extracted from the Google Maps API is limited by Google who implement API keys that place a restriction on the amount of information that can be recorded and processed. For some Google account, a user can access 12 API keys each allowing 2,500 queries. After this limit, a service fee is charged of \$0.50/1000 queries up to a maximum of 100,000 additional queries per day (Google, 2017). With that being said, a study on actual cities is not feasible and instead this method will be used to study inclusion probabilities of households for a particular geographical region.

The Google Maps API allows various techniques to extract information that has been recorded on its server. One particular method that will be explored in this paper is to generate household addresses using Google Maps API and then inputting a batch job into the Google Maps Geocoding API to convert the addresses into latitude and longitude coordinates. Initiating this process begins by defining a central address,

which can be an existing household. Although selecting an existing household as a geographic center produces an unorthodox layout for the EPI method, at times of an immediate crisis it may prove to be more quick and affordable to target households in a specific region using some known focal point. Afterwards, the data can be rescaled to shift any household at the center of the collected cluster.

Using the Google Maps Geocoding API, the focal address is converted to its (x, y) coordinates and a circle is produced from the geographic center with radius r. Within this geographical circle centered around the focal address, all household addresses located within its radius could be mapped, recorded, and inputted as batch jobs for the Google Maps Geocoding API to extract coordinates. Under the assumption that N_i households are required for the sampling frame, the radius will grow incrementally by some small value such that new households will be constantly introduced until a geographic region of size N_i is obtained. The algorithm to execute this method is:

1. Select a starting address and let the radius be $\kappa = 0.0000001$ units. Input the starting address through the Google Maps Geocoding API to retrieve a latitude and longitude. If the starting address is a household, record the address along with its latitude and longitude. Otherwise, proceed to the next step.

2. Locate all addresses within the circle of radius κ, centered at the starting address.

3. If new addresses are located, record and input the addresses through the Google Maps Geocoding API. Tabulate the addresses, latitudes and longitudes of each new observation recorded. If no new addresses are located, proceed to the next step. Note, Google sometimes clusters a group of households as one address

and they should be completely ignored and discarded if located.

4. Increase the radius of the circle incrementally by 0.0000001 after each iteration of step 3.

5. Repeat steps 3 and 4 until a sample size of N_i is obtained.

Figure 5.10: Demonstration of extracting household coordinates using Google Maps API from a residential area in Hamilton, Ontario, Canada. The household identified with a red icon was used to initialize the Google Maps API procedure. All locations marked with a yellow icon are the subsequent households collected.

Note, the size of the radius could be increased or decreased depending on prior knowledge of the households in a particular region. Typically, neighbouring households

have latitude and longitude coordinates which are nearly indistinguishable. The radius provided in the algorithm should be suitable to capture each sequentially.

With this progression forward in mapping real household coordinates to study the EPI method, many challenges have yet to be met. One of the more common problems using this procedure is that the sampling frame must be cleaned to account for some unexpected data; like apartments, stores, factories and other non-residential structures and locations which get recorded in the process. Some addresses are incorrectly marked on Google Maps or simply do not exist, producing inaccurate household information. Due to many countries' laws and regulations, Google is unable to accurately address mapping information problems in their systems and applications. For example, China requires special authorization for individuals seeking geographic information in China (NASG, 2002).

5.3.2 Study on Real Household Coordinate Inclusion Probabilities Using the Strip and Sector Method

Even with the ability to obtain household address information using the Google Maps API, the limitations mentioned prevent two stage studies of the EPI method as done in sections 5.1 and 5.2. To reiterate, the Google Maps API procedure restricts the population size that can be acquired within any city because of the query limit and the computational time. In fact, it would take a minimum of 6 hours to extract information from 10,000 or more households. If disease patterns were simulated on real household data, it would produce results that are partially realistic because many assumptions still need to be held. The strongest assumption being that every

(a) Two household addresses incorrectly mapped to the corner of their respective streets.

(b) Non-residential addresses mixed in the collection process.

(c) Collection of a non-existent address.

Figure 5.11: Common issues using the Google Maps API to obtain household addresses.

household has an eligible candidate who can be used to study the characteristic of interest.

Another aspect of the EPI method that could be analyzed in depth are the household inclusion probabilities. In her thesis, Reyes (2016) explored a method of sampling households called arc or sector sampling, which targets a sector of a geographic region relative to the town center with span $\gamma \in (0, 2\pi]$. Strips and sectors are distinct methods of sampling at the household level, but fundamentally

their procedures are similar. They both begin at the center of the town and initiate through the means of a random direction. Using sector sampling, households along the same path or direction relative to the town center have the same probability of being selected first for the household picked; whereas in strip sampling they generally do not.

Since the strip method uses a base of size α and the sector method uses an arc of span γ, a possible setting to compare the two sampling methods is by finding the base and span which produces the same effective area of a strip and sector respectively on a geographic location. The process used to find the effective area of a strip and sector begins by first selecting a geographic region and generating N_i address coordinates using the Google Maps API. After the data has been cleaned to account for non-residential addresses and a household list has been generated, the house farthest from the geographic region's center is marked and denoted as $r^* = \max_{i \in N_i}\{r_i\}$. All strips and sectors generated in this geographic region will thereafter start at the geographic center and move outward a maximum distance of r^*.

The area covered using the sector method will effectively capture any household that lies within its radius r^*, but the strip method produces two corner areas that lie outside r^*. Since this area is redundant and can not capture any households, it should be removed from the rectangular strips to produce areas that effectively capture households. Looking at Figure 5.12, the area needed to generate a strip that effectively captures households is shaded in blue and green. In other words, the effective area of the strip is the area of the sector of span ζ and radius r^*, and the two triangular regions of base $\frac{\alpha}{2}$ and height $d = \sqrt{(r^*)^2 - \left(\frac{\alpha}{2}\right)^2}$. The explicit formula for the effective area of the strip of base α, $A_{\text{strip},\alpha}$, is

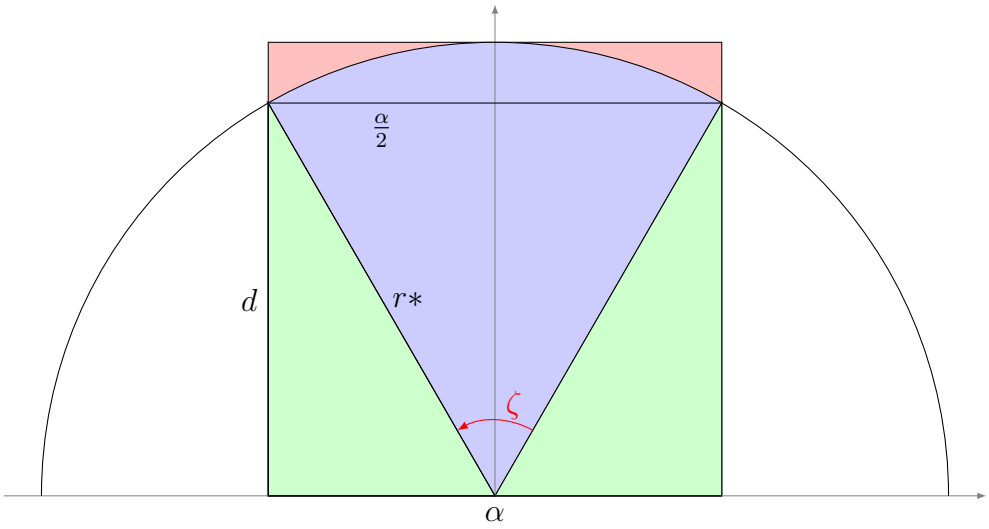


Figure 5.12: The effective area of a strip of base α. The area in red will not capture any households since all households will be less than or equal to r^* units from the origin. Consequently, it is not needed for the strip and should be removed when calculating the effective area of the strip.

$$\begin{aligned} A_{\text{strip},\alpha} &= 2 \times \frac{d \times \frac{\alpha}{2}}{2} + \pi (r^*)^2 \times \frac{\zeta}{2\pi} \\ &= \frac{d \times \alpha}{2} + \frac{(r^*)^2}{2} \times \zeta \\ &= \frac{d \times \alpha}{2} + \frac{(r^*)^2}{2} \times 2 \times \sin^{-1}\left(\frac{\alpha}{2r^*}\right) \\ &= \frac{d \times \alpha}{2} + (r^*)^2 \times \sin^{-1}\left(\frac{\alpha}{2r^*}\right). \end{aligned} \tag{5.2}$$

The effective area of a sector with span γ, $A_{\text{sect},\gamma}$, must be equal to $A_{\text{strip},\alpha}$ so that the methods are comparable. The corresponding formula for γ would then be:

$$\gamma = \frac{2 \times A_{\text{strip},\alpha}}{(r^*)^2}. \tag{5.3}$$

An implication of using area as a basis of comparison is that γ is restricted to the set $(0, \pi]$ because it is impossible to produce an effective area greater than $\pi \cdot (r^*)^2$

using strips of base $\alpha \in (0, 2 \max_{i \in \{1,\ldots,N_i\}} r_i]$.

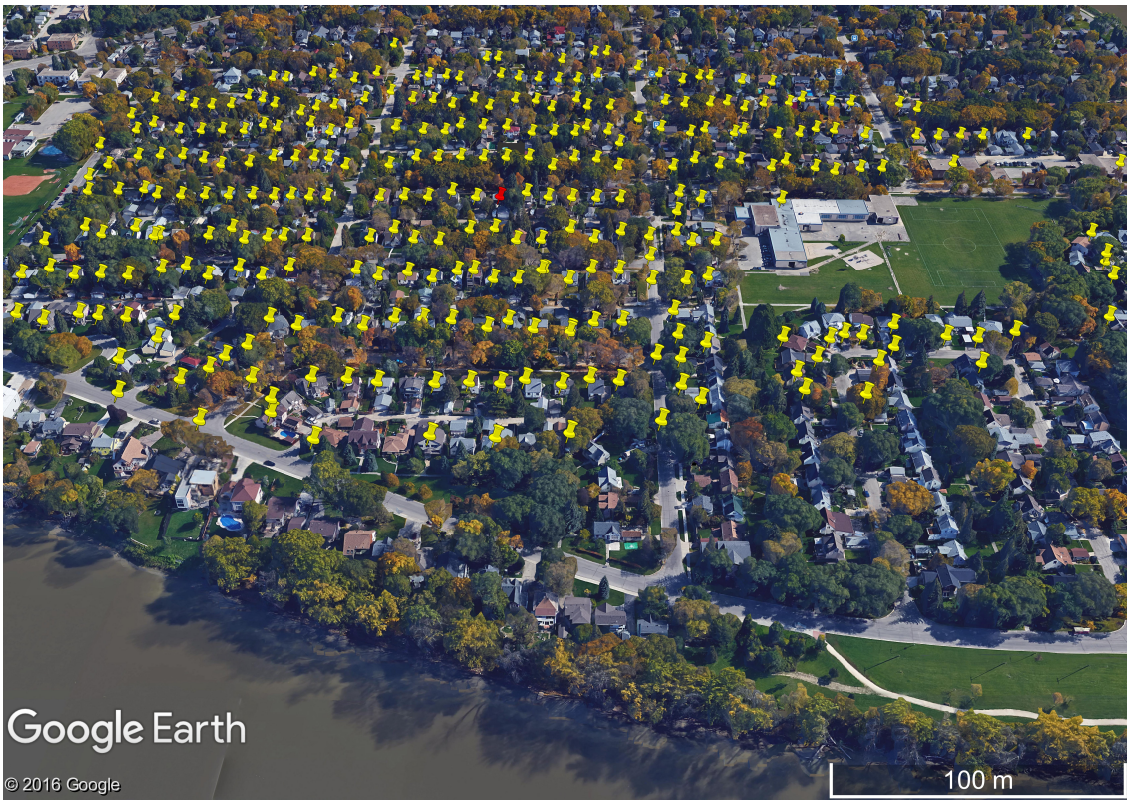


Figure 5.13: Google Earth aerial view of 303 residential households from a neighbourhood in Winnipeg, Manitoba, Canada. The household used to initialize the collection of this information from this geographical region is identified with a red pin, and all subsequent households collected using the Google Maps API procedure have been identified by yellow pins.

To carry out a study comparing the two methods, 303 household coordinates were extracted from a suburban residential area in Winnipeg, Manitoba, Canada using the Google Maps API procedure. Since the coordinates look almost indistinguishable on a two dimensional surface using their latitudes and longitudes, they were rescaled and shifted allowing simpler selections of base α. For the study, $\alpha = \{129.9844, 259.9688,$

389.9533, 519.9377, 649.9221, 779.9065} were selected to compute the inclusion probabilities on the transformed coordinates of the Winnipeg data, as seen in Figure 5.13. Correspondingly, the arc span for the sector which produces the identical effective areas relative to a strip of base α were γ={0.6635673, 1.3082132, 1.9132230, 2.4532633, 2.8915062, 3.1415927}, respectively. Under the strip and sector sampling methods, the inclusion probabilities were based off sample sizes of $n = \{7, 15, 30, 60, 90\}$.

Analyzing the density plot in Figure 5.14 of the neighbourhood from Winnipeg, it is observed that for $n = \{7, 15, 30, 60\}$ and γ={0.6635673, 1.3082132, 1.9132230, 2.4532633} the sector method produces smaller variations in inclusion probabilities relative to the strip method. When $n = 120$, the strip and sector methods have identical spreads in inclusion probabilities. This occurrence arises because households in dense areas appear more often in smaller sized samples, whereas larger sample sizes will include households that are not located in dense clusters leading to greater inclusion probabilities. As the sample size increases or the base size and arc span increase proportionately, the inclusion probabilities for both methods are approximately uniform in distribution. It is to be expected as the sample size approaches the town size, the household inclusion probabilities should become more uniform because more households are being included in the sample. In the most extreme case where all households are included in every sample, every household must have an inclusion probability of 1 regardless of the secondary sampling method used.

Another interesting deduction from the study of the neighbourhood in Winnipeg is that when the base of the strip approaches $2r^*$, implying that $\gamma = \pi$, the inclusion probabilities under both methods of sampling are identical. This has to do with the fact that when strips of size $2r^*$ are used, every household is captured in a span of π.

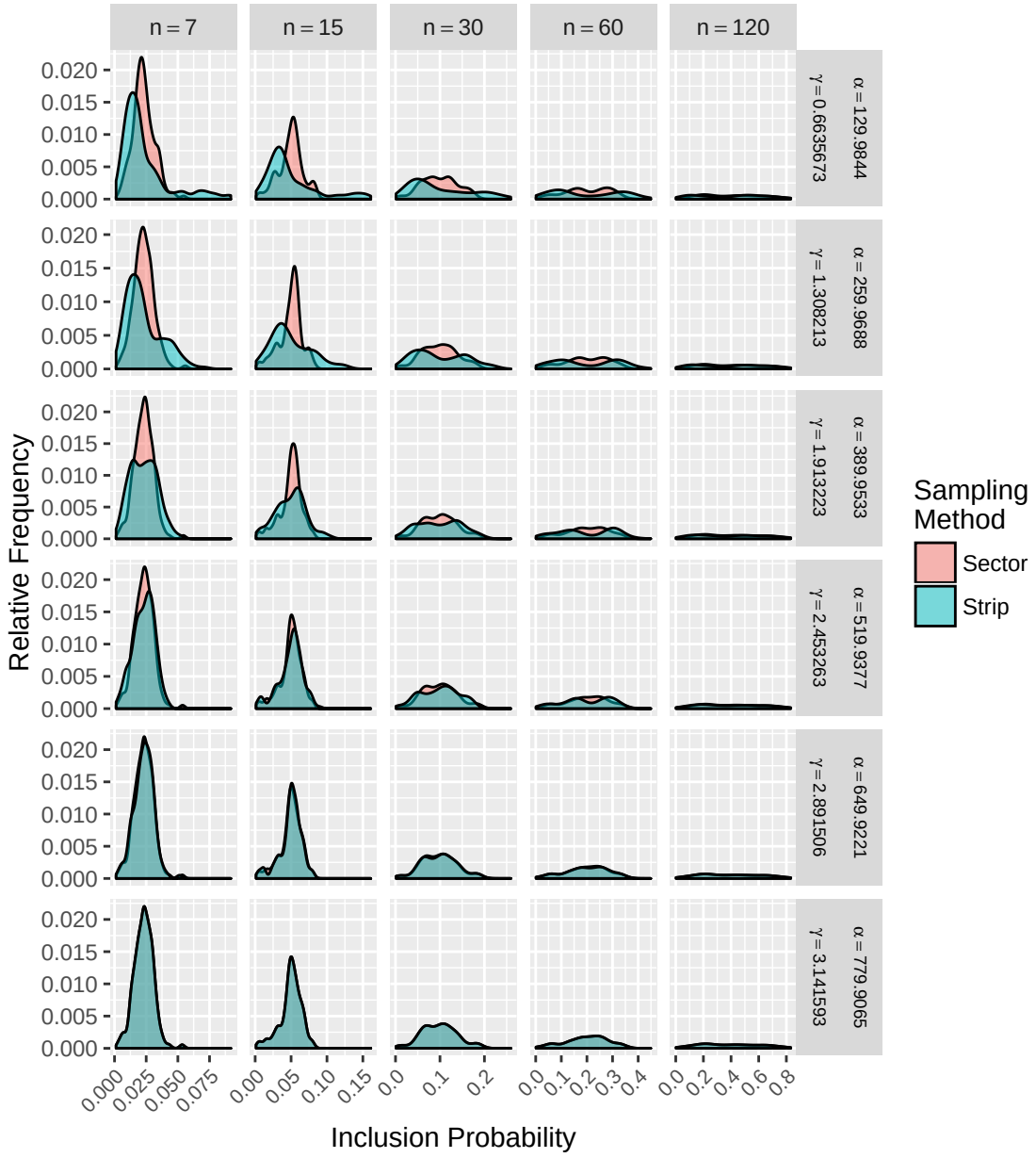


Figure 5.14: Density plot of the inclusion probabilities using the strip and sector sampling methods on 303 households from a neighbourhood from Winnipeg, Manitoba, Canada.

A span of π for every household also corresponds to the sector method of sampling where $\gamma = \pi$. In general, the household inclusion probabilities of a strip with base α approaching $2r^*$ become more identical to those respective household inclusion probabilities obtained using a sector with span π.

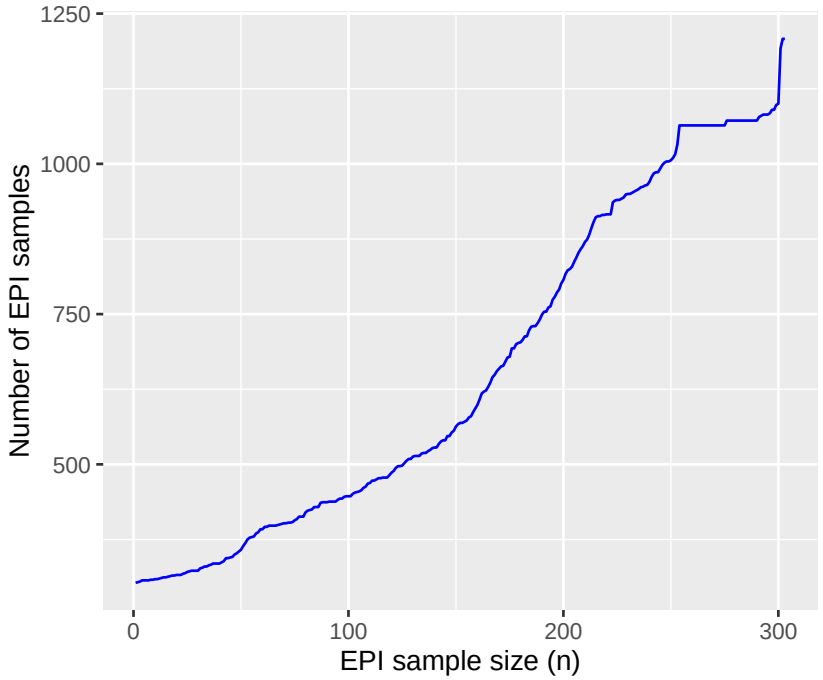


Figure 5.15: Connected scatter plot showing the number of EPI samples relative to the EPI sample size n from the data set acquired using the Google Maps API procedure in neighbourhood from Winnipeg Manitoba, Canada.

It can be seen that real households are generally spread in a pattern which is a combination of the *grid* and *circ* spatial patterns, as seen in Figure 5.10 and 5.13. Since extracting households using the Google Maps API produces some inaccurate information regarding the mapping of households, it is currently difficult to obtain a precise understanding of the spread of households using the procedure. It does however give a very rough understanding of the relationship between the size of the

EPI samples at the household level and the number of samples produced at each respective level, for a given geographical area. Looking at Figure 5.15, it is seen that there are multiple nearest neighbours leading more than 303 possible EPI path samples. As n gets larger, the number of possible EPI samples increases roughly linearly. This cannot be generalized because every neighbourhood sampled will vary from region to region depending on the geographic area of study around the globe.

Chapter 6

Summary, Discussion and Future Directions

The methodological aspects and application of the EPI method produce insightful results and advantages. The traditional approach can be applied quickly with very little cost and no sampling frame; the only information required being the sizes of each town. With the modifications proposed in this paper, a sampling frame is required to extract the household inclusion probabilities for the HT estimators, but not required for the EW estimators.

Utilizing the Google Maps API allows for the construction of sampling frames reflective of the real world, which could be used to study and get a better understanding of the EPI method's application. However, challenges in processing larger data still exist. Depending on the spatial distribution of households, the number of EPI samples could be extremely large with respect to the sample size required for research.

6.1 Summary and Discussion

A major emphasis of this thesis went to assessing the EPI method with respect to both of its stages. At the primary level, a variety of common techniques were applied and the associated statistics were explored to obtain exact figures. The conventional method of performing systematic sampling, as explained in the EPI manual, was generalized to account for non-integer sampling intervals. Formulating the inclusion probabilities lead to a closed form equation, as seen in Equation (2.4).

Sampling WOR and WR using PPS were two methods researched, however studying the EPI method while sampling at the primary stage WOR using PPS was much more difficult to implement. The was mainly attributed to the fact that when there are many towns included in the sampling frame, it is far too time consuming to compute the inclusion and joint inclusion probabilities. A possible solution to this problem is to use SHARCNet, a network of high performance computers, to process the inclusion and joint inclusion probabilities as separate jobs and later using the output in the assessment phase of the EPI method. On the other hand, sampling WR using PPS was far simpler to integrate as part of the EPI method's assessment because manipulating the inclusion probability formulas lead to closed form equations of the inclusion and joint inclusion probabilities that were quick to process.

Employing strips at the secondary stage of sampling was another equally important concept used to study the EPI method. Much like sectors, strips reduce possible bias produced by surveyors collecting data going towards the edge of the town in some random direction relative to the center of a town. The most crucial difference between strips and sectors is that strips tend to sample those households

that are closer to the center of the town. It was established that for strips with base sizes $\alpha = 2 \cdot r^*$, a sector with span $\gamma = \pi$ will compute identical inclusion probabilities for each household individually. Finally, since strips are more geometrically complex compared to sectors, extracting inclusion probabilities using strips is on average two to three times longer than using sectors.

Performing an exact assessment of the EPI method lead to generalized statistics for any two stage sampling technique, conditional that inclusion and joint inclusion probabilities are accessible for both stages. One of the original studies on the EPI method lead to the use of the EW estimator, as shown by Lemeshow and Robinson (1985). Aiming to construct an unbiased estimator guided the research towards the discovery and implementation of the well known HT estimator. Systematic sampling and sampling WR using PPS were the only two primary sampling techniques that were feasible with respect to computational time and had methodology to produce exact inclusion probabilities. Consequently, they were used to construct a family of two EW and four HT estimators.

The six estimators were applied to a variety of spatial and disease patterns which were studied in Reyes (2016). Population and sampling plans were constructed by considering a list of factors like the number of towns generated, the sizes of each town, the dimensions of each town, the number of towns sampled at the primary stage, the number of households sampled from the sampled towns at the secondary stage and the prevalence of disease for some geographic region. This produced 108 population and sampling plans which were used to study each of the six estimators.

From the results that were obtained, the HTSYS and HTWR estimators were unbiased as expected while all other estimators were biased. The largest spread in

bias and MSE measurements across the 108 sampling and population plans was produced by the EW family of estimators. The least amount of spread in MSE with respect to prevalence was seen in the HTSYS and HTSRS1 estimators. As mentioned before, the HTSRS2 estimator always underestimated the population prevalence and it distinctly classified the bias and MSE measurements into three groups which corresponded to each of the three prevalence values used in the study.

Evaluating each estimator under the three disease patterns concludes that the EW1 estimator should always be selected when the disease follows a random dispersion pattern because it produces very accurate and precise estimates. Although, fundamentally the HTSYS and HTSRS1 estimators were unique estimators with respect to one another, the MSE evaluations under both showed a nearly identical spread. Conclusively, it can be stated that the HTSYS estimator should be used when diseases are spread through a central gradient or pocket pattern because it is an unbiased estimator that produced the most precise estimates relative to the other estimators.

It would be unlikely that groups or organizations carrying out the EPI method in field work would actually know the prevalence before hand, as it would typically defeat the purpose of doing a survey. As a result, no sensible recommendations could be made regarding which estimator to use for specific levels of prevalence. However, the HTSYS estimator is the ideal estimator to use when looking at spatial patterns and prevalence values simultaneously across the study. As the prevalence approaches a value of $p = 0.5$, the spread in the bias and MSE measurements increases across all estimators. This is a logical occurrence because when the prevalence is close to $p = 0$ or $p = 1$, each estimator typically increased in both accuracy and precision.

When the sample size at the secondary stage increased for a fixed prevalence, the spread in bias and MSE slightly decreased across all estimators.

A separate study on the effect of the strip's base size at the secondary stage of sampling across the various spatial and disease patterns was conducted after fixing the sample size to $n = 7$ and prevalence to $p = 0.5$. If households are dispersed in an aggregated gradient spatial pattern or if the disease is spread randomly, the EW1 estimator is recommended to be implemented in the sampling design of the EPI method because it produces very little bias and has the smallest MSE. When the households were spread in a circular gradient pattern and the disease assumes a pocketed or circular gradient pattern, then the EW2 estimator should be implemented. Depending on a particular researcher's goal, they might decide to choose an estimator that is more accurate versus precise. When the spatial pattern follows the *runi*, *grid* or *aggr* pattern and the disease is spread through pocketing, the researcher can either choose the EW estimator to get more precise estimates or the HTSRS1 estimator to get more accurate estimates.

6.2 Future Directions

From an applications stand point, there are a handful of alternative sampling methods that can be implemented at the primary level of sampling for the EPI method. Although sampling WOR using PPS could not be studied due to calculation constraints, it is possible with the aid of computational databases like SHARCNet to do so. Many other methods were considered to integrate as part of the primary

stage of sampling, like Lahiri's method which is an accept-reject algorithm that typically samples WR using PPS. Due to the complex sampling structure, these unique methods present difficulties in obtaining exact inclusion probabilities. One possible solution to help study these types of methods is to use a Monte Carlo simulation to obtain close estimates of the inclusion probabilities. Although they will not be exact, they will give a close understanding of how the methods might affect the assessments of the EPI method actual field work.

Studying the EPI method using the sector method was not emphasized within this paper. But, using the comparative setting of effective areas, it is possible to study the effect of using strips versus sectors at the secondary stage for a variety of two stage estimators. In Reyes (2016), alternate EPI paths were explored sampling every k^{th} eligible household from a sampled town. Using the sector sampling method, the inclusion probabilities for $k = \{3, 5\}$ produced significantly different results compared to when $k = 1$ for small secondary sample sizes of $n = \{7, 15\}$. It is unclear how sampling every k^{th} household would effect the inclusion probabilities when sampling using the strip, however this is an aspect of the EPI method that can be studied in future projects.

Introducing non-eligible candidates to the study would greatly improve the representation of the EPI method's results. It is clear that visiting a household that does not contain an eligible candidate that contributes as a member of the EPI sample affects the EPI path. This in turn can have a significant impact on the inclusion probabilities, however it is unclear how it would affect them. Future studies should integrate the study of households that do not contain eligible candidates to mimic a real world setting as close as possible.

Developing the strip method gave rise to another secondary sampling technique, which is initiated by generating a circle with a random radius from the town center. After a circle has been generated, a household is randomly selected from within the defined circle to initiate the nearest neighbour algorithm as traditionally used in the EPI method. The most notable advantage of this method is that for reasonable radius values, it can be very cost effective in locating all households within the specified area as opposed to using a strip or a sector which requires surveyors to go to the boundary of the town for any base size of arc span.

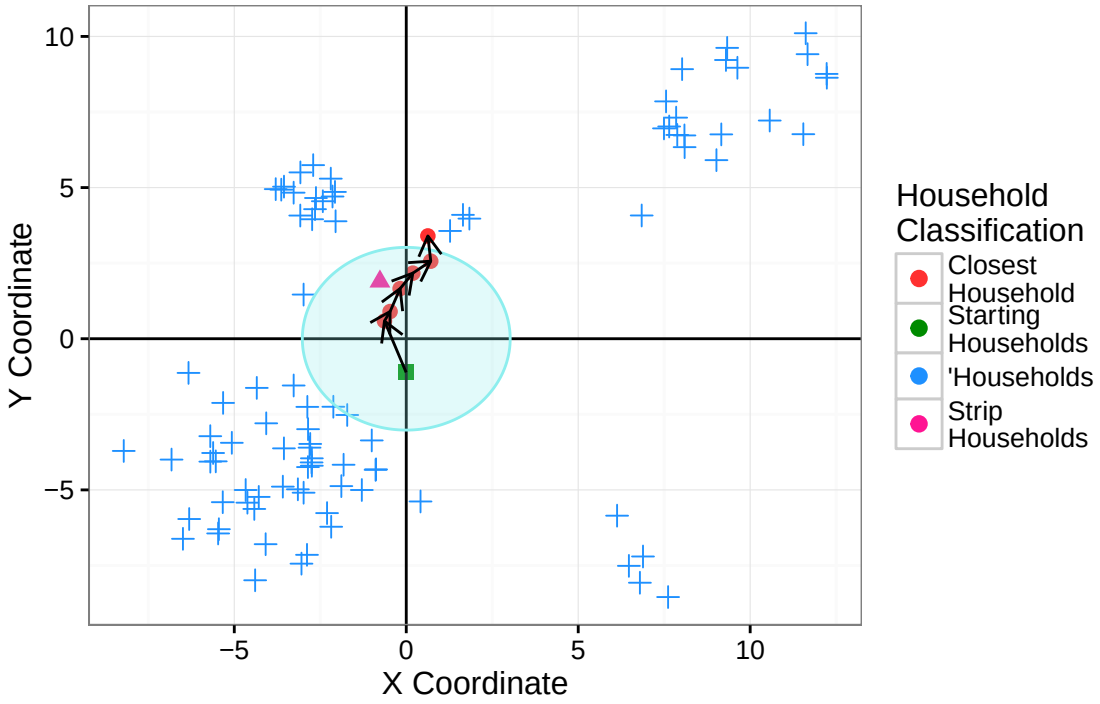


Figure 6.1: Illustration of the random radius method selecting a sample of $n = 7$ households at the secondary stage of the EPI method on a sampled town with 100 households. Arrows were used to show the sequence of households selected in the process of selecting nearest neighbours.

A major drawback of the random radius method is that it tends to select households close to the town center more likely than those further away, much like the strip method. This implies that the farthest households from the town center are generally never selected as the initial households for an EPI path. Careful methodology must be created to ensure that households bordering the edge of the town are given at least some non-zero probability of initial selection. With regards to computational time, the method itself has no variable to study making it far faster to process compared to the sector or strip method. To compare strips, sectors and the random radius methods, the effective area is still a suitable basis of comparison because a radius can always be found such that the effective areas across each sampling method are identical.

With the exponential rate of progress in technology and methodology, it would be wise to look at techniques to study the EPI method using real world spatial data. The four spatial patterns studied in this paper are common patterns in specific areas, but they do not capture the intricate structure of real world household patterns. Using the Google Maps API is definitely a useful technique in studying the EPI, however it is challenged with the issue of the amount of data it can capture and how quickly it can capture it.

In Yuan (2016), buildings and residential households are extracted from aerial scenes using convolutional networks. This will allow one to accurately isolate geographic regions aerially, and process the information through the developed convolution network to obtain precise household positioning. A recurring problem, which also appeared in the Google Maps procedure, is that all non-residential structures

are recorded using the convolution network. A solution to this problem is to selectively choose aerial imaging so that malls, plazas and other locations containing non-residential buildings are avoided. Using this alternate method to construct reasonable household information for geographic regions should enable more accurate and affordable studies of the EPI method.

Figure 6.2: Example of the convolution network developed by Yuan (2016) in a residential area from Washington, D.C.. Transparent red areas represent the infrastructures extracted using the convolution network and the blue pixels indicate the boundaries (Yuan, 2016).

Appendix A

Proofs

A.1 Validating the Inclusion Probabilities for Sampling WOR with PPS

Recall, in Section 2.1.5, it was mentioned by Lohr (2009) that the sum of inclusion probabilities when sampling WOR must equal m. Thus, the development of Equations (2.10) and (2.16) could be verified through this result. Both equations could be shown by induction. Beginning with Equation (2.10) and letting $m = 1$, the following result is obtained

$$\begin{aligned}\sum_{i=1}^{M} \alpha_i &= \sum_{i=1}^{M} P(\text{town } i \text{ is included in } s_{(1)}) \\ &= \sum_{i=1}^{M} \frac{N_i}{N} = \frac{N_1 + \ldots + N_M}{N} = 1\end{aligned}$$

Before showing the next result, it is important to note that the total size of the population, N excluding the size of town i N_i is simply the sum of all town sizes, excluding town i. Thus,

$$N - N_i = \sum_{j=1}^{M} N_j - N_i = \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} N_{k_1}$$

Continuing the induction proof, it is useful to show that Equation (2.6) holds for sizes of $m = 2$. Thus,

$$\begin{aligned}
\sum_{i=1}^{M} \alpha_i =& \sum_{i=1}^{M} P(\text{town } i \text{ is included in } s_{(2)}) \\
=& \sum_{i=1}^{M} \left[\frac{N_i}{N} + \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_{k_1}}{N} \frac{N_i}{N - N_{k_1}} \right] \\
=& \sum_{i=1}^{M} \frac{N_i}{N} + \sum_{i=1}^{M} \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_{k_1}}{N} \frac{N_i}{N - N_{k_1}} \\
=& 1 + \frac{N_1}{N} \frac{N_2}{N - N_1} + \frac{N_1}{N} \frac{N_3}{N - N_1} + \ldots + \frac{N_1}{N} \frac{N_M}{N - N_1} + \frac{N_2}{N} \frac{N_1}{N - N_2} + \frac{N_2}{N} \frac{N_3}{N - N_2} \\
& + \ldots + \frac{N_2}{N} \frac{N_M}{N - N_2} + \ldots + \frac{N_M}{N} \frac{N_1}{N - N_M} + \frac{N_M}{N} \frac{N_2}{N - N_M} + \ldots + \frac{N_M}{N} \frac{N_{M-1}}{N - N_M} \\
=& 1 + \sum_{i=1}^{M} \frac{N_i}{N} \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_{k_1}}{N - N_i} = 1 + \sum_{i=1}^{M} \frac{N_i}{N} \sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} \frac{N_{k_1}}{\sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} N_{k_1}} = 1 + \sum_{i=1}^{M} \frac{N_i}{N} \frac{\sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} N_{k_1}}{\sum_{\substack{k_1=1 \\ k_1 \neq i}}^{M} N_{k_1}} \\
=& 1 + \sum_{i=1}^{M} \frac{N_i}{N} = 1 + 1 = 2
\end{aligned}$$

To show the induction step, suppose that for a sample of size $m - 1$, the following

holds

$$\begin{aligned}\sum_{i=1}^{M}\alpha_i &= \frac{N_i}{N} + \sum_{\substack{k_1=1\\k_1\neq i}}^{M}\frac{N_{k_1}}{N}\frac{N_i}{N-N_{k_1}} + \sum_{\substack{k_1=1\\k_1\neq i}}^{M}\sum_{k_2\neq\{i,k_1\}}^{M}\frac{N_{k_1}}{N}\frac{N_{k_2}}{N-N_{k_1}}\frac{N_i}{N-N_{k_1}-N_{k_2}} + \dots \\ &\quad + \sum_{\substack{k_1=1\\k_1\neq i}}^{M}\sum_{k_2\neq\{i,k\}}^{M}\cdots\sum_{k_{m-2}\neq\{i,k_1,\dots,k_{m-2}\}}^{M}\frac{N_{k_1}}{N}\frac{N_{k_2}}{N-N_{k_1}}\cdot\dots\cdot\frac{N_i}{N-N_{k_1}-\dots-N_{k_{m-2}}} \\ &= m-1\end{aligned}$$

For samples of size m, the sum of the inclusion probabilities show that

$$\begin{aligned}\sum_{i=1}^{M}\alpha_i =& \sum_{i=1}^{M}\Bigg(\Bigg[\frac{N_i}{N} + \sum_{\substack{k_1=1\\k_1\neq i}}^{M}\frac{N_{k_1}}{N}\frac{N_i}{N-N_{k_1}} + \sum_{\substack{k_1=1\\k_1\neq i}}^{M}\sum_{k_2\neq\{i,k_1\}}^{M}\frac{N_{k_1}}{N}\frac{N_{k_2}}{N-N_{k_1}}\frac{N_i}{N-N_{k_1}-N_{k_2}} + \dots \\ &+ \sum_{\substack{k_1=1\\k_1\neq i}}^{M}\sum_{k_2\neq\{i,k\}}^{M}\cdots\sum_{k_{m-2}\neq\{i,k_1,\dots,k_{m-3}\}}^{M}\frac{N_{k_1}}{N}\frac{N_{k_2}}{N-N_{k_1}}\cdot\dots\cdot\frac{N_i}{N-N_{k_1}-\dots-N_{k_{m-2}}}\Bigg] \\ &+ \sum_{\substack{k_1=1\\k_1\neq i}}^{M}\sum_{k_2\neq\{i,k\}}^{M}\cdots\sum_{k_{m-1}\neq\{i,k_1,\dots,k_{m-2}\}}^{M}\frac{N_{k_1}}{N}\frac{N_{k_2}}{N-N_{k_1}}\cdot\dots\cdot\frac{N_i}{N-N_{k_1}-\dots-N_{k_{m-1}}}\Bigg) \\ =& m-1+ \\ &\sum_{i=1}^{M}\frac{N_i}{N}\sum_{\substack{k_1=1\\k_1\neq i}}^{M}\frac{N_{k_1}}{N-N_i}\sum_{k_2\neq\{i,k\}}^{M}\frac{N_{k_2}}{N-N_i-N_{k_1}}\cdots\sum_{k_{m-1}\neq\{i,k_1,\dots,k_{m-2}\}}^{M}\frac{N_{m-1}}{N-N_i-\dots-N_{k_{m-2}}} \\ =& m-1+ \\ &\sum_{i=1}^{M}\frac{N_i}{N}\sum_{\substack{k_1=1\\k_1\neq i}}^{M}\frac{N_{k_1}}{\sum_{\substack{k_1=1\\k_1\neq i}}^{M}N_{k_1}}\sum_{k_2\neq\{i,k\}}^{M}\frac{N_{k_2}}{\sum_{k_2\neq\{i,k\}}^{M}N_{k_2}}\cdots\sum_{k_{m-1}\neq\{i,k_1,\dots,k_{m-2}\}}^{M}\frac{N_{m-1}}{\sum_{k_{m-1}\neq\{i,k_1,\dots,k_{m-2}\}}^{M}N_{k_{m-1}}}\end{aligned}$$

$$=m-1+\sum_{i=1}^{M}\frac{N_i}{N}\times\frac{\sum_{\substack{k_1=1\\k_1\neq i}}^{M}N_{k_1}}{\sum_{\substack{k_1=1\\k_1\neq i}}^{M}N_{k_1}}\times\frac{\sum_{\substack{k_2=1\\k_2\neq\{i,k\}}}^{M}N_{k_2}}{\sum_{\substack{k_2=1\\k_2\neq\{i,k\}}}^{M}N_{k_2}}\times\ldots\times\frac{\sum_{\substack{k_{m-1}=1\\k_{m-1}\neq\{i,k_1,\ldots,k_{m-2}\}}}^{M}N_{m-1}}{\sum_{\substack{k_{m-1}=1\\k_{m-1}\neq\{i,k_1,\ldots,k_{m-2}\}}}^{M}N_{k_{m-1}}}$$

$$=m-1+1=m$$

The assertion is proven and it can be stated that Equation (2.10) is a valid formula to compute inclusion probabilities sampling WOR with PPS because it satisfies Equation (2.5). This implies that all the special cases while sampling WOR using PPS must adhere to Equation (2.5) as well.

A.2 Inclusion probabilities for towns of equal size when sampling WOR using PPS

Suppose $m=1$, then

$$\begin{aligned}\alpha_i =&P(\text{town } i \text{ is included in } s_{(1)})\\ =&\frac{N_i}{N}=\frac{Q}{MQ}=\frac{1}{M}\end{aligned}$$

Now, suppose $m=2$. Using Equation (2.11),

$$\begin{aligned}\alpha_i =&P(\text{town } i \text{ is included in } s_{(2)})\\ =&P(\text{town } i \text{ is selected in the } 1^{st} \text{ draw})+P(\text{town } i \text{ is selected in the } 2^{nd} \text{ draw})\\ =&\frac{N_i}{N}\left(1+\sum_{\substack{k_1=1\\k_1\neq i}}^{M}\frac{N_{k_1}}{N-N_{k_1}}\right)=\frac{Q}{MQ}\left(1+\sum_{\substack{k_1=1\\k_1\neq i}}^{M}\frac{Q}{MQ-Q}\right)=\frac{N_i}{N}\left(1+\sum_{\substack{k_1=1\\k_1\neq i}}^{M}\frac{N_{k_1}}{N-N_{k_1}}\right)\end{aligned}$$

$$=\frac{1}{M}\left(1+\frac{(M-1)Q}{(M-1)Q}\right)$$

, since k_1 can take on all values, except i. Therefore k_1 has $M-1$ choices.

$$=\frac{1}{M}(1+1)=\frac{2}{M}$$

To show the induction step, suppose that

$$P(\text{town } i \text{ is selected up to the } (m-1)^{th} \text{ draw}) = \frac{m-1}{M}.$$

Then using Equation (2.10), if a sample of size m is to be collected,

$$\begin{aligned}
\alpha_i =& P(\text{town } i \text{ is included in } s_{(m)})\\
=&\frac{N_i}{N}+\sum_{\substack{k_1=1\\k_1\neq i}}^{M}\frac{N_{k_1}}{N}\frac{N_i}{N-N_{k_1}}+\sum_{\substack{k_1=1\\k_1\neq i}}^{M}\sum_{\substack{k_2=1\\k_2\neq\{i,k_1\}}}^{M}\frac{N_{k_1}}{N}\frac{N_{k_2}}{N-N_{k_1}}\frac{N_i}{N-N_{k_1}-N_{k_2}}+\dots\\
&+\sum_{\substack{k_1=1\\k_1\neq i}}^{M}\sum_{\substack{k_2=1\\k_2\neq\{i,k\}}}^{M}\cdots\sum_{\substack{k_{m-1}=1\\k_{m-1}\neq\{i,k_1,\dots,k_{m-2}\}}}^{M}\frac{N_{k_1}}{N}\frac{N_{k_2}}{N-N_{k_1}}\cdot\dots\cdot\frac{N_i}{N-N_{k_1}-\dots-N_{k_{m-1}}}\\
=&\frac{N_i}{N}\left(1+\sum_{\substack{k_1=1\\k_1\neq i}}^{M}\frac{N_{k_1}}{N-N_{k_1}}+\sum_{\substack{k_1=1\\k_1\neq i}}^{M}\frac{N_{k_1}}{N-N_{k_1}}\sum_{\substack{k_2=1\\k_2\neq\{i,k_1\}}}^{M}\frac{N_{k_2}}{N-N_{k_1}-N_{k_2}}+\dots\right.\\
&\left.+\sum_{\substack{k_1=1\\k_1\neq i}}^{M}\frac{N_{k_1}}{N-N_{k_1}}\sum_{\substack{k_2=1\\k_2\neq\{i,k\}}}^{M}\frac{N_{k_2}}{N-N_{k_1}-N_{k_2}}\cdots\sum_{\substack{k_{m-1}=1\\k_{m-1}\neq\{i,k_1,\dots,k_{m-2}\}}}^{M}\frac{N_{k_{m-1}}}{N-N_{k_1}-\dots-N_{k_{m-1}}}\right)\\
=&\frac{1}{M}\left(1+\frac{(M-1)Q}{(M-1)Q}+\frac{(M-1)Q}{(M-1)Q}\frac{(M-2)Q}{(MQ-Q-Q)}+\dots\right.\\
&\left.+\frac{(M-1)Q}{(M-1)Q}\frac{(M-2)Q}{(M-Q-Q)}\cdot\dots\cdot\frac{(M-(m-1))Q}{MQ-Q-\dots-Q}\right)\\
=&\frac{1}{M}\left(1+1+1+\dots+1\right)=\frac{1+(m-1)}{M}=\frac{m}{M}.
\end{aligned}$$

Bibliography

1. Bennett, S., Radalowicz, A., Vella, V., and Tomkins, A. (1994). A computer simulation of household sampling schemes for health surveys in developing countries. *International Journal of Epidemiology*, 23(6):1282–1291.

2. Bennett, S., Woods, T., Liyanage, W. M., and Smith, D. L. (1991). A simplified general method for cluster-sample surveys of health in developing countries. *World Health Statistics Quarterly*, 44(3):98–106.

3. Bolker, B. (2008). *Ecological Models and Data in R*. Princeton University Press.

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

5. Brogan, D., Flagg, E. W., Deming, M., and Waldman, R. (1994). Increasing the accuracy of the Expanded Programme on Immunization's cluster survey design. *Annals of Epidemiology*, 4(4):302–311.

6. Ferrante, M. and Frigo, N. (2014). On the expected number of different records in a random sample.

7. Galway, L., Bell, N., SAE, A. S., Hagopian, A., Burnham, G., Flaxman, A., Weiss,

W. M., Rajaratnam, J., and Takaro, T. K. (2012). A two-stage cluster sampling method using gridded population data, a GIS, and Google Earth^{TM} imagery in a population-based mortality survey in Iraq. *International Journal of Health Geographics*, 11:12.

8. Google (2017). Google maps APIs: Documentation.

9. Henderson, R. H., Davis, H., Eddins, D. L., and Foege, W. H. (1973). Assessment of vaccination coverage, vaccination scar rates, and smallpox scarring in five areas of West Africa. *Bulletin of the World Health Organization*, 48(2):183–194.

10. Henderson, R. H. and Sundaresan, T. (1982). Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bulletin of the World Health Organization*, 60(2):253–260.

11. Katz, J., Yoon, S. S., Brendel, K., and West, K. P. (1997). Sampling designs for xerophthalmia prevalence surveys. *International Journal of Epidemiology*, 26(5):1041–1048.

12. Lemeshow, S. and Robinson, D. (1985). Surveys to measure programme coverage and impact: a review of the methodology used by the expanded programme on immunization. *World Health Statistics Quarterly*, 38(1):65–75.

13. Lindsey, J. (2017). rmutil: Utilities for nonlinear regression and repeated measurements models. R package version 1.0.

14. Lohr, S. L. (2009). *Sampling: Design and Analysis*. Duxbury Press, Boston, 2nd edition.

15. National Administration of Survey, Mapping and Geoinformation of China (2002). Surveying and mapping law of the people's republic of china.

16. Pearson, A. L., Rzotkiewicz, A., and Zwickle, A. (2015). Using remote, spatial techniques to select a random household sample in a dispersed, semi-nomadic pastoral community: utility for a longitudinal health and demographic surveillance system. *International Journal of Health Geographics*, 14(1):33.

17. Reyes, M. (2016). An analysis of equally weighted and inverse probability weighted observations in the expanded program on immunization sampling method.

18. Serfling, R. E. and Sherman, I. L. (1965). *Attribute Sampling Methods for Local Health Departments*. U. S. Dept. of Health, Education, and Welfare, Public Health Service, Communicable Disease Center, Epidemiology Branch, Atlanta.

19. Wampler, P. J., Rediske, R. R., and Molla, A. R. (2013). Using arcmap, google earth, and global positioning systems to select and locate random households in rural haiti. *International Journal of Health Geographics*, 12(1):3.

20. World Health Organization (2008). Training for Mid-Level Managers (MLM) Module 7: The EPI Coverage Survey.

21. Yuan, J. (2016). Automatic building extraction in aerial scenes using convolutional networks. *CoRR*, abs/1602.06564.