Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Julia Gallini Date

Optimizing Cluster Survey Designs for Trachomatous Inflammation-Follicular in Amhara Region, Ethiopia

By

Julia Gallini

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

Paul Weiss, M.S.

Committee Chair

Scott Nash, Ph.D.

Committee Member

Optimizing Cluster Survey Designs for Trachomatous Inflammation-Follicular in Amhara Region, Ethiopia

By

Julia Gallini

B.S.P.H; B.A.

The University of North Carolina at Chapel Hill

2017

Thesis Committee Chair: Paul Weiss, M.S.

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2019

Abstract

Optimizing Cluster Survey Designs for Trachomatous Inflammation-Follicular in Amhara Region, Ethiopia

By Julia Gallini

Background: Trachoma is the leading infectious cause of blindness globally, with some of the highest rates observed in Amhara National Regional State, Ethiopia. An international effort led by the World Health Organization (WHO) aims to reduce the prevalence of the trachomatous inflammation-follicular (TF) among children ages 1 to 9 years to below 5% globally by 2020. A key component of the strategy to eliminate trachoma as a health problem is the mass drug administration (MDA) of the antibiotic azithromycin. MDA decisions are made based on prevalence estimates from two stage cluster surveys. Work remains to formally mathematically evaluate the WHO recommended trachoma survey design.

Objective: Characterize the effects of the number of clusters and the number of households sampled on the precision and accuracy of TF estimates as well as costs of surveying in Amhara.

Methods: We simulated a population from which samples of varying numbers of clusters and households were selected. Bootstrapping techniques were used for variance estimation. Sampling schemes were evaluated on the following metrics: precision (through 95% uncertainty intervals), proportion of incorrect and low MDA decisions made (less MDA prescribed than warranted), design effect, effective sample size, and percent of sample efficiently used. Costs for each design were estimated based on previous work.

Results: The number of clusters sampled has a greater impact on the precision of the estimate than the number of households. Increasing households sampled yields a diminishing return in precision beyond 30 households. In low prevalence regions, cluster number has a lesser impact on precision than in higher prevalence regions. Samples are most mathematically and cost efficient used when sampling less than 30 clusters and households. The number of clusters drives survey cost more than the number of households sampled.

Conclusions: We recommend that past data on a region inform the survey design decision. For lower prevalence areas (less than 10%) we recommend 20 clusters of 20-30 households; for moderate to high prevalence regions (greater than 10%) we recommend 15 clusters of 20-30 households. Efficient use of surveying funds now will allow sustained surveying as Amhara approaches TF elimination as a public health problem.

Optimizing Cluster Survey Designs for Trachomatous Inflammation-Follicular in Amhara Region, Ethiopia

By

Julia Gallini

B.S.P.H; B.A.

The University of North Carolina at Chapel Hill

2017

Thesis Committee Chair: Paul Weiss, M.S.

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2019

Acknowledgements

First and foremost, I would like to thank the Biostatistics Department at Emory for their constant support in all areas of my education. I would especially like to thank my committee chair and advisor Paul Weiss, M.S. who has been an invaluable mentor throughout this process, providing me ample opportunity to explore this project to my own liking while guiding me throughout the entire experience. His enthusiasm for this project has been contagious. I would also like to thank my faculty mentor Renee' Moore, PhD. for her continual guidance throughout my experience at Emory.

This study would not have been possible without the collaboration of the trachoma control team at The Carter Center. I would like to thank my committee member Scott Nash, PhD. for his constant support and advice. With his guidance, I learned more than I ever anticipated about the application of biostatistics in the real world of public health. I would like to thank Kelly Callahan, M.P.H. for her support of this project as the director of the program. I would also like to thank Aisha Stewart, M.P.H. and the rest of the trachoma control team for first sparking my interest in the disease. Working with the trachoma team provided countless opportunities to not only learn, but to apply my skills to real world problems.

I would like to thank my biostatistics cohort members, who have simultaneously been the best study group and the best way to decompress during stressful moments. I would not be the biostatistician I am today without their support, nor would I have had as enjoyable a time at Emory without their friendship.

Lastly, I would like to thank my family for their unwavering support, answering all of my phone calls, and talking me through every situation whether panicked, frustrated, or celebratory. I attribute every step of my educational and professional experience to your constant positivity and encouragement.

Table of Contents

I. Introduction 1

II. Background 2

III. Methods 6

Generation of Population Dataset 6

Methods for Cluster Level Analysis 13

Methods for Household Level Analysis 14

Statistical Methods for Cost Analysis 15

Bootstrapping Methodology 17

Proportion Incorrect and Low Methodology 17

Design Effect and Effective Sample Size Methodology 18

IV. Results: Gott Level 20

Proportion of Incorrect and Low Decisions Results 25

Design Effect Results 26

V. Results: Household Level 28

Design Effect and Effective Sample Size Results 32

VI. Cost Analysis Results 35

VII. Future Directions 38

VIII. Discussion 39

Cluster Level Analysis 39

Household Level Analysis 41

Reappearance of Trachoma.....43

Cost Analysis.....44

Recommendations 45

Limitations 45

Conclusion.....47

IX. Appendix 52

Supplemental Figures to Gott Level Results.....52

Supplemental Figures for Household Level Results 57

SAS Macros.....82

Population Simulation Macro.....82

Drawing Samples of 30 Households (Using Segment Structure) Macro.....84

Drawing Samples of Varying Numbers of Households (Ignoring Segment Structure) Macro 88

I. Introduction

Trachoma (a neglected tropical disease or NTD) is the world's leading infectious cause of blindness[1]. It is a bacterial infection spread from person to person and the active infection is both preventable and treatable. After repeated infection over many years scarring on the eyelids causes the eyelashes to turn inwards (entropian), scratching the cornea. Without treatment this scratching will eventually cause corneal opacity, or irreversible blindness[1]. Trachoma is endemic in areas of the world where crowded living conditions and inadequate sanitation are common, with 85% of cases (roughly 18 million) occurring in Africa [1]. In 1996, the World Health Organization (WHO) adopted the four-part intervention strategy to eliminate trachoma as a public health problem which includes Surgery for trachomatous trichiasis, Antibiotics for the infection, Facial cleanliness (to reduce transmission), and Environmental improvement (to better sanitation infrastructure) (SAFE). In recent years, several countries have successfully eliminated trachoma as a public health problem such as Cambodia, Ghana, Nepal, and Mexico[2].

In countries where the disease remains endemic like Ethiopia and South Sudan[1], trachoma prevalence is monitored through regular sample-based surveying performed by the governments of the countries often with the support of nonprofit organizations. Based on the WHO simplified grading scheme there are five grades of trachoma: trachomatous inflammation- follicular (TF), trachomatous inflammation- intense (TI), trachomatous scarring (TS), trachomatous trichiasis (TT), and corneal opacity (irreversible blindness)[1]. The WHO hopes to eliminate trachoma as a public health problem by

2020, meaning that in every country in the world, the prevalence of TF will be below 5% among children aged 1-9 years, and the prevalence of TT will be below 0.1% among individuals of all ages[3]. Accurate and precise prevalence estimates are necessary not only for the justified validation of the eventual elimination of the disease, but also for program evaluation to determine the efficacy of interventions. While formal analysis of optimal survey design has been performed for other NTD's [4-6] there has been less formal analysis conducted on trachoma survey design.

II. Background

Ethiopia is one of the relatively few remaining countries with consequentially high trachoma rates. In particular, the Ethiopian region of Amhara is considered hyperendemic for trachoma, [7, 8] and has been since the area was first surveyed in 2001 [9]. Early surveys estimated that 45% of TT in all of Ethiopia was estimated to occur in Amhara, with roughly 1 in 20 adults suffering [9]. The SAFE strategy was implemented throughout the region starting in 2007. Despite the effort, surveys conducted between 2011 and 2015 demonstrated that 94% of woredas (similar to districts) remained above the TF elimination threshold, suggesting that trachoma will remain a problem in Amhara for many years to come [8].

The antibiotic portion of the SAFE strategy is designed combat ocular infection with *Chlamydia trachomatis,* and thus to reduce the clinical sign TF. [10]. The WHO has established guidelines recommending mass drug administration (MDA) in woredas where the TF prevalence is above the threshold (5% in children ages 1-9 years). MDA entails administering azithromycin to all residents 1 year and above of a woreda regardless of

their disease status with a goal of 80% coverage; the idea being that the drug will reach even those that are asymptomatic at the moment [11]. The guidelines recommend that in woredas where the TF prevalence in children 1-9 years is 5%-10% one year of MDA be implemented, with the area being resurveyed the following year. In woredas where the prevalence is 10%-30%, the recommendation is that MDA be implemented for 3 years before resurveying, and in woredas with a TF prevalence above 30%, the recommendation is that MDA be implemented for 5 years before resurveying. In woredas where TF prevalence is below the 5% threshold, no MDA is recommended, as this is considered to be a low enough prevalence that the disease is not a public health or economic burden [12]. Typically, woredas that come in below the 5% TF threshold are resurveyed after 2 years or more to ensure that trachoma has not reappeared [12]. TF reappearance is an issue faced in hyperendemic areas such as Ethiopia.

Sampling schemes for determining TF prevalence are relatively similar across countries, though there are some differences due to varying administrative structures. The Global Trachoma Mapping Project (GTMP) set a precedent with a 2015 paper that discusses sample size calculations for TF surveys. Aiming for ±3% precision and assuming 10% prevalence and a design effect of 2.65 the authors calculated a necessary sample size of roughly 1200 children aged 1-9 years in each evaluation unit (in Amhara, a woreda)[13]. Building on the GTMP paper, the WHO published survey design recommendations for TF in 2018. Survey designs assume 10% prevalence with ±3% precision in woredas where TF is believed to be above the 5% threshold, and assume 4% prevalence with ±2% precision in woredas where TF is believed to be near the 5%

threshold [14]. Because of this assumption the example of a woreda (district) with TF prevalence is 4% is used throughout this study (among other examples).

In Ethiopia, the currently implemented sampling design is a two-stage cluster design. Trachoma prevalence is measured at the woreda level, meaning that a prevalence estimate is obtained for every woreda in the country. In order to reach the required number of children (usually about 1200) and taking into account what percentage of the population is aged 1-9 years, the following sampling plan is implemented. In the first stage, 30 gotts (villages) are randomly selected from a comprehensive list of gotts within the woreda. In this paper, "cluster" (the general sampling term) and "gott" (the specific Ethiopian term for these analyses) are used interchangeably. In the second stage, one development team (similar to a neighborhood) is randomly selected from each sampled gott (development teams are also referred to as "segments"). Field teams go to all households in the segment that they can reach in one day: usually about 30 households [15]. Teams evaluate every individual over 1 year old for TF, though WHO TF thresholds are based on the prevalence in children aged 1-9 years[16]. By this process, a woreda level TF prevalence is determined and MDA guidelines are followed as appropriate.

The cost of trachoma surveys is based on a large number of factors including field team training costs, supplies, transportation, data technology support, etc. A 2011 study examined the average cost of trachoma surveys in 8 African countries, including Ethiopia. The authors found that of the four cost domains studied (training, field work, supervision, data entry), fieldwork, specifically personnel and transportation during fieldwork, were the main cost drivers. The median cost per cluster was found to be $311, while the median cost per survey was found to be $4,784 [17]. A 2017 study by the

Global Trachoma Mapping Project (GTMP) found an average per cluster cost of $692 across all of the countries monitored [18] by GTMP, including Ethiopia, Uganda, Pakistan, Laos, and many others [18]. A 2019 cost study of Amhara alone also found that personnel and transportation were the main drivers of cost. On the other hand, the study found an average per cluster cost of $752, a considerable increase from the previous studies [5, 18, 19]. The increase is likely due to the study being 8 years later, as TF survey costs have greatly increased over time [19].

In general NTD surveillance is an expensive endeavor, given their rarity as well as their tendency to exist in remote areas that are often difficult and/or time-consuming to access. To design cost-efficient surveys without sacrificing quality of results, simulation studies have been performed for a number of NTD's such as schistosomiasis, soil-transmitted helminths, and trachomatous trichiasis [4-6]. Using a computer simulated population dataset based on empirical data, a 2017 schistosomiasis study examined the number of schools sampled as well as the number of students sampled within each school, and optimized relative to the cost of the survey. The researchers found an optimal number of 15-20 schools sampled per district and 20-30 children per school such that the cost of the survey remained relatively low, while the precision of the prevalence estimate remained adequate [5]. While the specifics of what constitutes "low" cost and "adequate" precision are different for each NTD, the underlying idea of mathematically optimizing cost and survey design has potential to greatly inform the decision of what sampling scheme to use for trachoma surveillance. While some simulation-based work comparing sampling methods in trachoma has been done, one such work compared cluster random sampling to integrated threshold mapping [20], while the other looked only at TT [6].

Additionally, trachomatous trichiasis (TT) has much lower prevalence than trachomatous inflammation-follicular (TF), so sampling analyses for TF alone would be beneficial to the trachoma community. The body of work comparing different cluster random sample designs in TF is lacking. The currently recommended design (nicknamed “30 by 30”) needs to be further mathematically analyzed to determine the potential for substantial improvement in the realm of cost-efficiency.

Therefore, there are three main aims of this study. The first aim is to examine the relationship between the number of gotts (villages) sampled and the precision of the TF prevalence estimates, relative to the MDA thresholds as per the WHO recommendations. The second aim is to build upon the first aim by adding the element of varying the number of development teams (neighborhoods) sampled within each gott (and thus number of households), and again examine the precision of the TF prevalence estimates. Finally, the third aim is to optimize the cost of the survey design with the precision of the TF prevalence estimate, and subsequently inform the decision of what sampling scheme to use in TF surveillance from a mathematical standpoint.

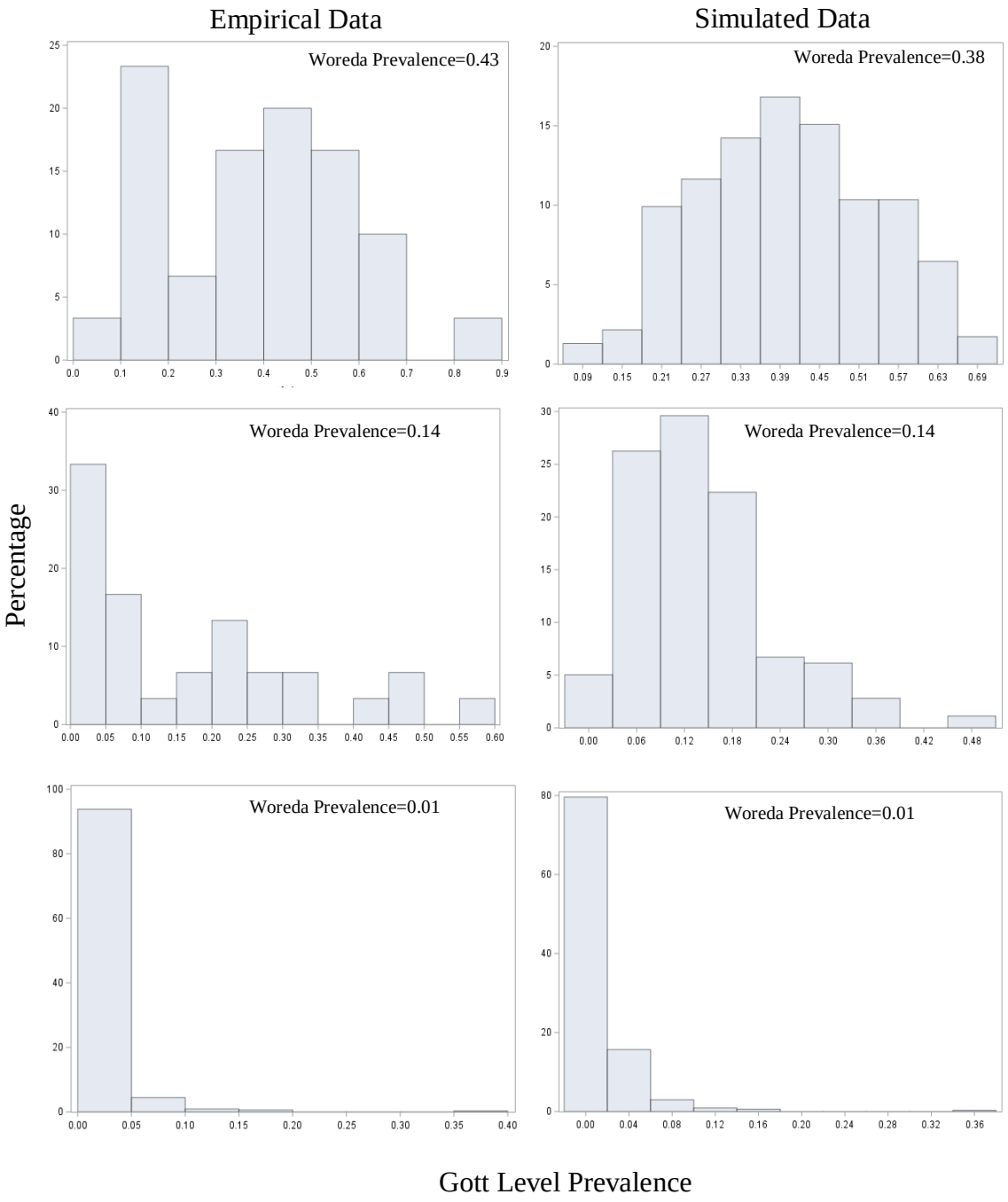
III. Methods

Generation of Population Dataset

To compare the precision and cost of different sampling schemes, we created a population database using SAS 9.4 to represent the population of Amhara, Ethiopia. The distribution of TF in the population was characterized using empirical surveys supported by The Carter Center in late 2017 and conducted using the Tropical Data system [21]. Empirical distributions of gott level prevalence within various woreda can be seen in Figure 1 side by side with simulated gott level prevalence rates.

Each observation in the dataset represented an individual aged 1-9 years in the population, since this age group is used to determine the district prevalence of TF. In all, there were 1,012,474 children aged 1-9 years in the dataset representing 30 woredas. There are actually 165 woredas in Amhara [22], but it was determined that 30 woreda were sufficient to be representative of possible TF distributions for the simulation, and that representing all 165 woreda would become unnecessarily repetitive. The SAS macro written to create the dataset can be seen in the Appendix. User inputs to the macro include the potential for 14 different prevalence rates at the woreda level.

Figure 1. Gott Level Prevalence in Empirical Woredas versus Simulated Woredas



For 24 woredas we set the default prevalence between 8% and 40% and for 6 woredas we set the default prevalence between 0 and 5%. These values were chosen to represent probable woreda prevalence rates in Amhara as of 2017[23]. As the TF rates are expected to drop over time on account of the MDA pressure, the same macro can be used to simulate a more realistic Amhara population, and recreate the analyses seen in this study.

We used the following procedure to randomly generate the population dataset. First, 30 woredas were generated numbered 1-30. Within each woreda a random number of gotts were generated using the generalized negative binomial distribution. The traditional negative binomial distribution is defined:

$$P_{r,p}(x) = \binom{x+r-1}{r-1} p^r (1-p)^x$$

and models the probability of $r-1$ successes and x failures in $x+r-1$ trials, with the $x+r$ trial being a success [24]. Its advantage in modeling discrete events over other distributions, like Poisson, lies in the fact that the variance is not required to be the same as the mean. The mean of the negative binomial distribution is $r\frac{1-p}{p}$, and the variance is $r\frac{1-p}{p^2}$.

The generalized negative binomial distribution (defined by Jain and Consul in 1971)[25] allows for even further flexibility in the shape of the distribution by permitting a non-integer form of x, which takes the form $n+\beta u-u$ in the generalized distribution function:

$$P_{\beta(u,n,o)} \frac{n\Gamma[n+\beta u]}{u!\,\Gamma[n+\beta u-u+1]} p^u (1-p)^{n+\beta u-u}$$

where $n > 0, u = 0,1,2,$ ... and $|\beta p| < 1$. Thus, the term $n + \beta u - u$ has the ability to take on non-integer values. Since the definitions of p and $1 - p$ are arbitrary, we see that u itself can take on non-integer values. The generalized negative binomial distribution reduces to the traditional negative binomial distribution when $\beta = 1$. The parameters $p, r,$ and x are more recognizable with the negative binomial distribution so this paper will maintain this symbology for the use of the generalized negative binomial.

Using the parameters $p = 0.01$ and $r = 1.51$ in the generalized negative binomial distribution (x being the outcome in this case), an expected value of $x = 150$ gotts per woreda was achieved, as in the surveys performed in late 2017[23], with a variance of 14,949.

To generate a random value for TF prevalence for each gott, the beta distribution was used. The beta probability function takes the form:

$$P(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}(1 - x)^{\beta - 1}x^{\alpha - 1}$$

The beta distribution function takes on values between 0 and 1 as does prevalence, and like the negative binomial distribution, the mean and variance are not dependent on one another[24]. The mean of the beta distribution is $\frac{\alpha}{\alpha+\beta}$, while the variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. As a result, the distribution of TF prevalence can be adjusted to the user's liking, though currently, it mimics the distribution seen in late 2017 surveys. As for parameter values, the TF prevalence was generated follow a beta distribution around a mean prevalence assigned to the respective woreda (also a user input value). The first parameter, α, was defined as equal to the prevalence, and β was defined as equal to (1 - prevalence). Woreda level prevalences, as discussed above, ranged from 0.4 to 0.002 for this study.

The variance of each beta distribution was designed to vary by prevalence: woredas with higher prevalence tend to have higher variance, and woredas with lower prevalence tend to have smaller variance, as seen in the 2017 data available[23]. To implement differing variances, in high prevalence woredas (prevalence ≥ 0.3), a multiplier of 20 was used for both the α and β parameters. In moderate prevalence woredas ($0.1 \leq$ prevalence ≤ 0.3) and low prevalence woredas (prevalence < 0.1), a multiplier of 25 was used, resulting in substantially less variation in individual gott level prevalence rates. Figure 1 depicts the resulting distributions of gott level TF prevalence rates within several woredas from the simulated data compared to empirical data from 2017.

Next, within each gott a random number of segments were generated using the generalized negative binomial distribution with an expected value of 3.24, the calculated average number of segments per gott as of late 2017 ($p = 0.1, r = 0.25$). There is expected to be some degree of correlation of TF rates within segments since they are geographically determined units, and TF is the clinical manifestation of an infectious disease. To account for this correlation, a random segment level prevalence was generated for each segment using the beta distribution with an expected value of the corresponding gott level prevalence. A multiplier of 200 was used for beta distribution parameters such that the variance was relatively small, since while there is some variation of prevalence rates within gotts, the variance is limited in comparison to the variation between gotts (as mentioned above, multipliers of 20 and 25 were used for this component of variance). Though some segments in Amhara do not contain exactly 30 households, this portion is relatively small and thus exactly 30 was used for simplicity's sake.

A random number of children aged 1-9 were for each household using the Poisson distribution with expected value 1.107 (and consequently variance 1.107 since both the mean and the variance of the Poisson and λ). The mean value was calculated as the number of children per household from the late 2017 surveys. The Poisson distribution takes the form:

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

In this case, based on 2017 data[23], the variance of children per household is similar to the mean, so the Poisson distribution was utilized to most closely approximate the empirical distribution[24].

Finally, a random indicator variable for presence of TF was generated using the Bernoulli distribution, which takes the form:

$$P(x) = p^{x}(1-p)^{1-x}$$

where x is equal to 0 or 1[24]. The Bernoulli distribution allows the simulation to assign an indicator (0 or 1; no or yes) to each observation (child) with probability p. In this case, p was set equal to the segment level prevalence rate. The mean of the distribution is p and the variance is $p(1-p)$.

The preceding procedure resulted in the 1,012,474 children aged 1-9 years and their TF status comprising the simulated Amhara population. Samples were drawn from this population for purposes of comparing precision and cost effectiveness across sampling designs.

Methods for Cluster Level Analysis

As mentioned previously, the currently implemented sampling scheme for TF surveillance in Amhara is a two-stage cluster design stratified by woreda. In the first stage, 30 gotts are sampled from each woreda; in the second stage, one segment (roughly 30 households) is sampled per gott. All children aged 1-9 in sampled households are evaluated for signs of active TF.

The first aim sought to evaluate the extent to which adjusting the number of gotts selected affected the precision of the TF prevalence estimate obtained. Instead of only selecting samples with 30 gotts, samples were drawn with 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, and 34 gotts. The high number of values was selected to be able to see a trend in prevalence precision as number of gotts increased relative to the default value of 30. There may be woredas in which even 34 clusters does not provide the desired precision. In the second stage, for all sampling iterations, one segment (30 households exactly for this simulation) was selected and all children ages 1-9 were used in the sample. Weighted TF prevalence estimates were calculated for each of the 30 woreda for all samples.

Survey weights were calculated by first determining the probability of selection at each stage of the design. In first stage, the probability of selection reduces to:

$$p_{Stage1} = \frac{Number\ of\ Gotts\ Selected}{Number\ of\ Gotts\ in\ Woreda}$$

Similarly, the probability of selection in the second stage reduces to:

$$p_{Stage2} = \frac{Number\ of\ Households\ Selected}{Number\ of\ Households\ in\ Gott}$$

The weight for each selected unit is then calculated as the inverse of the overall probability of selection[26]:

$$Weight = \frac{1}{p_{Stage1} * p_{Stage2}}$$

Solely examining the relative precision of sampling designs with varying numbers of clusters was not expected to yield mathematically interesting results on its own: the precision of estimates should improve as more clusters are included in the sample. However, this analysis becomes interesting when evaluated relative to the current MDA cutoff guidelines, which will be addressed in chapter IV.

The following metrics were used to assess the cluster level analysis: the width of the 95% uncertainty interval, the proportion of incorrect MDA decisions made, the proportion of low MDA decisions made, the design effect, the effective sample size, and the percent of sample size used efficiently. Methods for each of these metrics are discussed later in this section.

Methods for Household Level Analysis

To evaluate the second aim and analyze the effect of the number of second stage units selected (households), the segment structure was temporarily ignored. To simplify the process and observe the overall trend, samples were drawn using only 15, 20, 25, and 30 gotts in the first stage, as opposed to all values used in the cluster level analysis. Iterations for the second stage were 10, 20, 30, 40, 50, and 60 households. All possible combinations of these two stages resulted in 24 different sampling schemes.

From past surveys done in Amhara, the average survey team can evaluate about 30 households in one day with respect to time[15]. Many gotts in Amhara have less than 60 households[23], therefore, this analysis did not explore beyond two days' worth of surveying as it was deemed impractical and unrealistic. When sampling 40, 50, or 60 households in gotts with only 30 households, all units were selected. Weighted TF

prevalence estimates were calculated for each of the 30 woreda across all 24 sampling combinations using the same sampling weights described in earlier methods. When all units were selected from a gott, p_{stage2} reduces to 1.

The following metrics were used to assess the household level analysis: the width of the 95% uncertainty interval, the proportion of incorrect MDA decisions made while holding the number of clusters constant, the proportion of low MDA decisions made while holding the number of clusters constant, the design effect, the effective sample size, and the percent of sample size used efficiently. Methods for each of these metrics are discussed later in this section.

Statistical Methods for Cost Analysis

In general, the cost of surveys is comprised of a fixed cost (that of getting to the cluster) and a variable cost (varying based on the number of secondary units selected). To calculate an estimated cost for each sample design, two components were needed: the cost of measuring one household, and the cost of measuring a gott (aside from the cost of measuring households within the gott). As mentioned previously, the work that has been done on trachoma survey costs has not been extensive, nor has it been conclusive [17-19]. Therefore, to estimate costs for this analysis, several assumptions were made. The burden of these assumptions is somewhat lessened by the fact that relative cost of sampling designs is of interest for this study, and not exact cost.

The basis for cost estimates for this study came from work done by Slaven *et al.* in 2019. The authors calculated an average per cluster (gott) cost of $752 in Amhara. Of this cost, an estimated 86.4% was spent on field work, of which as estimated 4.6% was spent of TF evaluation supplies [19], and thus after simple calculations it was estimated

for this study that $29.87 worth of supplies was needed to sample one household. Slaven found that personnel incurred an average of 37.2% of fieldwork costs, so for this study a value of $241.70 per day spent sampling was used. Although the travel time within a village most likely differs widely by setting, in Amhara, it takes most field teams one day to measure 30 households [15]. Therefore, for this study we assumed that when sampling 30 or fewer households, the sampling would theoretically take one day per gott, and when sampling more than 30 households, the sampling would take 2 days per gott. To estimate the cost per gott apart from the cost of measuring individual households, it was assumed that a negligible percentage of transportation costs were accrued from household to household, and that the grand majority came from traveling to and from the gott. Thus, using Slaven's estimate of 58% of fieldwork costs being due to travel, and 13.6% of the overall cost per cluster coming from training, for this study it was estimated that $479.11 is spent per gott, apart from the costs incurred by sampling households. In summary, the following function was used to calculate overall cost of a given sampling design:

$$Cost = 479.11 * gotts + 29.87 * households + 241.70(1 + I_{>30\ households})$$

The notation $I_{>30\ households}$ represents an indicator variable for sampling beyond the first 30 households in the gott. Using this formula, the approximate cost was calculated for each of the aforementioned sampling schemes[19].

The following metrics were used to assess the cost level analysis: cost range, the relationship between uncertainty interval width and cost, and cost wasted. The cost range was calculated by adding and subtracting $20 to the per household cost, and adding and subtracting $100 to the per cluster cost, such that a range of possible costs was calculated for each sampling scheme given that all reference costs used are approximations, and not

exact. Cost waste was calculated using the percentage of the sample efficiently used (discussed later in this chapter) and the cost estimates derived above for each sampling scheme using the following formula: $Waste = Cost * (1 - p_{used})$, where p_{used} is the percentage of the sample used efficiently.

Bootstrapping Methodology

To determine relative precision of prevalence estimates, bootstrapping techniques were used. Bootstrapping allows for estimation of parameters like the mean and variance, and consequently the calculation of empirical confidence intervals, referred to here as uncertainty intervals (UI). The advantage of bootstrapping lies in the limited assumptions related to normality compared to the standard confidence interval. Additionally, bootstraps are asymptotically more accurate than standard intervals using sample standard error[27].

1,000 samples were drawn for each sampling scheme, i.e. 1,000 samples of 30 households in each of 14 gotts were drawn; 1,000 samples of 40 households in each of 20 gotts were draw, etc. Weighted prevalence estimates were calculated for each of the 30 woredas across each of the 1,000 samples. Within each woreda, the 1,000 estimates were sorted from highest to lowest. The 2.5^{th} percentile and the 97.5^{th} percentile were used as the lower and upper bounds of 95% uncertainty intervals. The width of the intervals was compared across sampling schemes as the metric for comparing precision.

Proportion Incorrect and Low Methodology

Knowles et al. used the following methods in analyzing schistosomiasis survey designs [5]. They have been replicated here for TF.

To compare sampling schemes to one another relative to MDA guidelines, the proportion of the 1000 samples that resulted in an incorrect MDA decision relative to the true woreda level prevalence was calculated. For instance, if the true prevalence of the woreda fell within the range for 3 rounds of MDA (population prevalence of 0.1-0.3) but the sample yielded an estimate that warranted 5 round of MDA (sample prevalence of >0.3), this was considered an incorrect MDA decision.

Additionally, the proportion of incorrect and low MDA decisions was calculated. The example given above is consequential as far as supplying excessive antibiotics before resurveying, but unproblematic relative to the actual elimination of the disease. Thus, estimating the prevalence at lower than it truly is more problematic than overestimating, so the proportion of the 1000 samples in which a low incorrect MDA decision was made (i.e., 1 round of MDA instead of 3 rounds) was calculated for each sampling scheme as an additional metric.

Design Effect and Effective Sample Size Methodology

The design effect (often "$deff$") is a quantification of the amount of survey error induced by using an alternative sampling scheme in place of the gold standard simple random sample. A design effect of 1 indicates the sampling scheme is mathematically identical to a simple random sample. There are many formulas for determining the design effect of a sampling scheme, including:

$$D_{eff} = 1 + (m - 1)\rho,$$

where m is the number of observations in each cluster and ρ is the intra-cluster correlation coefficient, a measure of the correlation between observations within a cluster. Another (mathematically identical) formula for the design effect is:

$$D_{eff} = \frac{Var_{Design}}{Var_{SRS}}$$

offering an additional interpretation of the design effect as the ratio of variance of the estimates between the implemented design and a simple random sample[26].

An empirical version of the design effect, denoted $\widetilde{D_{eff}}$, was calculated for each sampling scheme such that the precision of each scheme relative to a simple random sample could be compared. The empirical design effect was calculated as follows:

$$\widetilde{D_{eff}} = \frac{\widetilde{Var(p)}}{Var_{SRS}}$$

In the above equation, $\widetilde{Var(p)}$ denotes the empirical variance of the prevalence estimates across the 1,000 replications for each sampling design within each woreda, serving as an estimator for the variance of the sampling design. Thus, the variance of each sampling scheme was calculated empirically rather than theoretically. Var_{SRS} denotes the theoretical variance of the prevalence estimate were the same sample size to have been selected using a simple random sample instead of a two-stage cluster design. The theoretical variance of the prevalence estimate from simple random sample is calculated:

$$Var_{SRS} = \left(\frac{N-n}{N-1}\right)\left(\frac{p(1-p)}{n}\right)$$

In the above equation, N represents the total population size, n represents the sample size, and p represents the true population level prevalence for the given woreda[26, 28].

The total size of the simulated Amhara population dataset was N = 1,012,474 individuals. The true woreda level prevalences p were derived from the simulated data. Across the 1,000 samples drawn for each scheme, there was some degree of variation in sample size due to the variation in number of children per household. However, this

difference was assumed to be negligible, and thus the mean sample size across the 1,000 samples for each scheme was used as n.

Empirical design effects were compared for each woreda for all sampling schemes, such that the precision could be compared from scheme to scheme. Additionally, the effective sample size was calculated for each sampling scheme as: $n_{eff} = \frac{n}{\widetilde{D_{eff}}}$, where n is the actual sample size. The effective sample size measures what sample size under a simple random sample would be equivalent to the actual sample size[26]. Thus, the higher the effective sample size, the more precise the point estimate. Effective sample sizes were calculated for all sampling schemes as well.

Additionally, the "percent of the sample size effectively used" was calculated to compare sampling schemes in terms of their efficiency design-wise. This metric was calculated as the inverse of the design effect multiplied by 100.

$$p_{used} = \frac{100}{\widetilde{D_{eff}}}$$

It essentially represents the percent of each sample that is actually used in estimating the prevalence. It will always be less than 100% in a cluster sampling design since cluster designs do not achieve perfect efficiency relative to a simple random sample[26, 28].

IV. Results: Gott Level

The length of the 95% uncertainty interval was examined relative to the number of clusters drawn in the first stage of the sample. Relative to the MDA guidelines, woredas with TF prevalence over 0.3 were considered high prevalence; woredas with prevalence between 0.1 and 0.3 were considered moderate prevalence, and woredas with prevalence

less than 0.1 were considered low prevalence. Figures 2a-2c illustrate the relationship between clusters and precision in these three progressive tiers.

In general, the width of the uncertainty interval decreases as clusters increase across all three tiers of prevalence, implying an improvement in precision as clusters increase. Mathematically, this is expected, as precision of point estimates should improve as more of the population is selected for sampling. The high prevalence woredas have the widest uncertainty intervals, while the low prevalence woredas have the narrowest. Additionally, the relationship between width of uncertainty interval and number of clusters appears to be much stronger in high prevalence woredas compared to moderate prevalence woredas, and higher in moderate prevalence woredas compared to low prevalence woredas.

Figures 3a-3e display the upper and lower bounds of the 95% uncertainty intervals for the prevalence estimate, as well as the true woreda level prevalence and MDA guideline cut points. While only five woredas are represented by these five figures, the other 25 figures can be seen in the appendix. Figure 3a illustrates the progression of the upper and lower bounds of the 95% uncertainty interval as more clusters are drawn in a woreda with TF prevalence 0.38. Between drawing 16 clusters and 18 clusters, the lower bound of the confidence limit crosses the MDA cut point of 0.3 (below 0.3 warrants 3 rounds of MDA; above 0.3 warrants 5 rounds of MDA). Essentially, in this theoretical woreda, if at least 18 clusters are sampled, 95% of samples will result in the correct decision for MDA per the WHO guidelines.

Figure 3b illustrated the same progression as Figure 3a, but with a prevalence rate of 0.31. Regardless of how many clusters are sampled, even when sampling beyond 30 clusters, the incorrect decision (3 rounds of MDA instead of 5 rounds of MDA) would be

made with a proportion of samples given how close the true prevalence is to the MDA cut point of 0.3. The proportion will be more precisely characterized later in the chapter. Figure 3c displays a woreda with prevalence 0.075. Depending on the sample, one of three MDA decisions could be made: 3 rounds of MDA, 1 round of MDA (the correct decision according to the WHO) or terminate MDA programs. No matter the number of clusters sampled, any of these three decisions is possible.

Figure 3d illustrates a woreda with prevalence rate 0.04. Here, no matter the number of clusters sampled there is a risk of making an incorrect MDA decision and performing a year of MDA when in reality the prevalence is below the elimination threshold.

Figure 3e displays an example of the final woreda classification for this analysis. Due to the low prevalence, regardless of how many clusters are sampled, the correct decision to terminate MDA programs will be made 95% of the time.

Figure 2a. High Prevalence Woredas

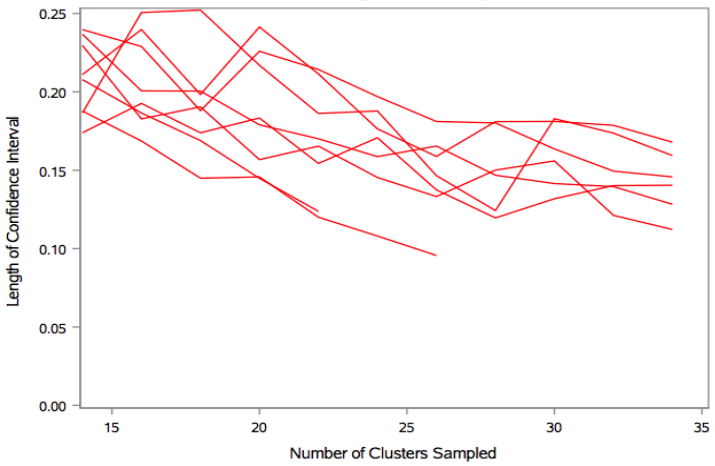


Figure 2b. Moderate Prevalence

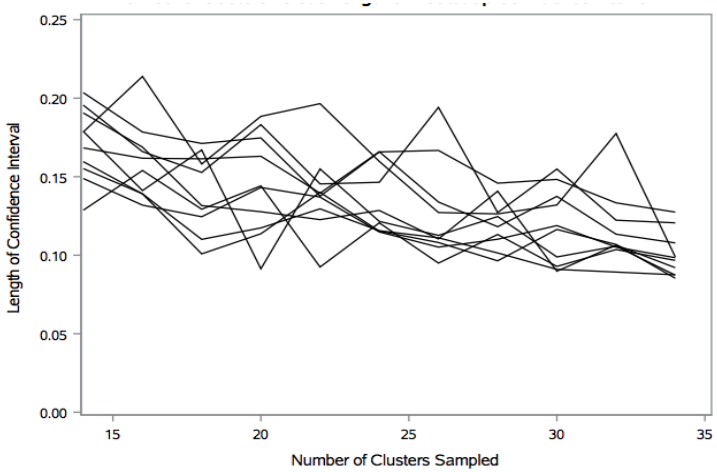
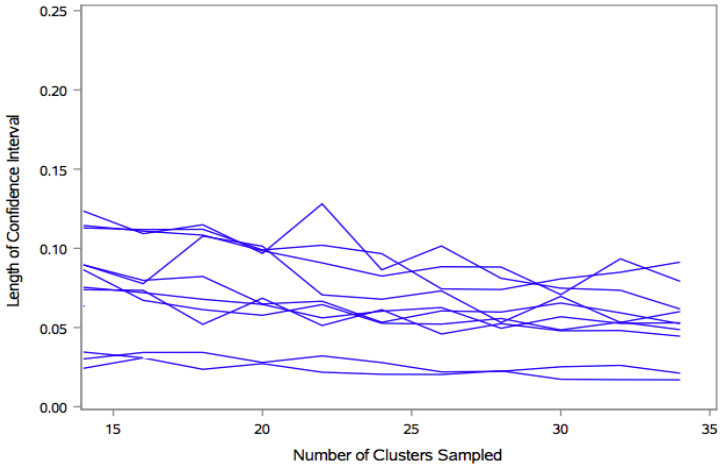


Figure 2c. Low Prevalence Woredas



Note: Lines that end prematurely represent woredas that contain less than 34 gotts, and thus achieve perfect precision prior to reaching 34 clusters sampled

Figure 3a. Woreda with True Prevalence 0.38

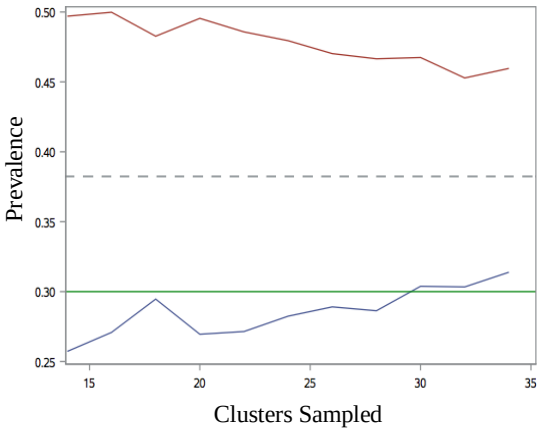


Figure 3b. Woreda with True Prevalence 0.31

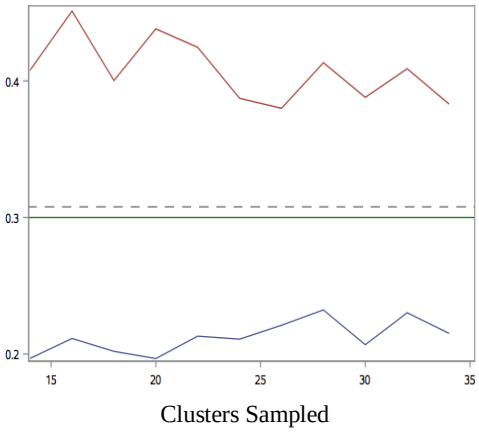


Figure 3c. Woreda with True Prevalence 0.08

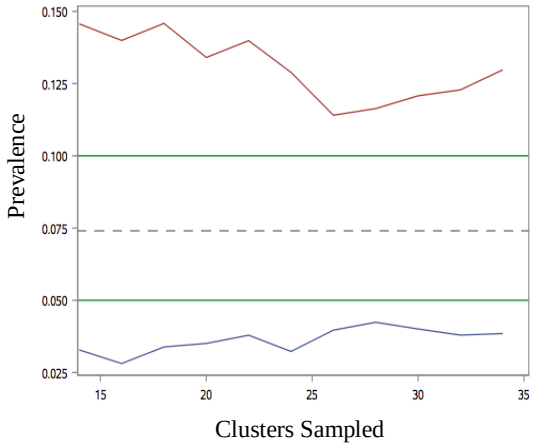


Figure 3d. Woreda with True Prevalence 0.04

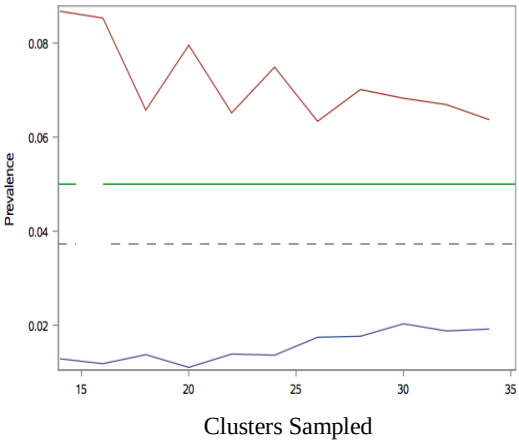
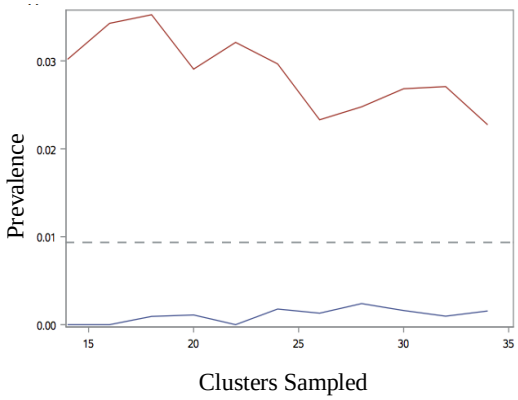
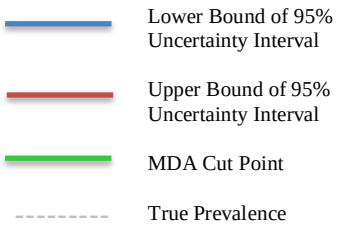


Figure 3e. Woreda with True Prevalence 0.01





Proportion of Incorrect and Low Decisions Results

Figure 4 displays the proportion of times in the 1000 samples the incorrect MDA decision was made relative to the true woreda level prevalence. Designs using 14, 20, 24, 30, and 34 clusters are displayed. The proportion of wrong decisions peaks around treatment decision cut points (0.05, 0.10, 0.30), which is to be expected given that no amount of precision (unless the entire population is selected) will yield a consistently correct decision when the true prevalence itself is on the borderline. The proportion of wrong decisions drops to its lowest point among all cluster levels midway in between treatment decision cut points. As far as differing results across number of clusters selected, there is some visible separation between 0.10 and 0.30 true prevalence. Selecting 34 clusters, while impractical, yields a lower proportion of wrong decisions, while selecting only 14 clusters yields the highest proportion of wrong decisions in this range, as would be expected. However, the separation amongst the cluster levels is minimal, and likely inconsequential given the potential cost savings (to be discussed in Chapter X).

Figure 4. Proportion of Wrong MDA Decisions by Cluster Number and True Prevalence

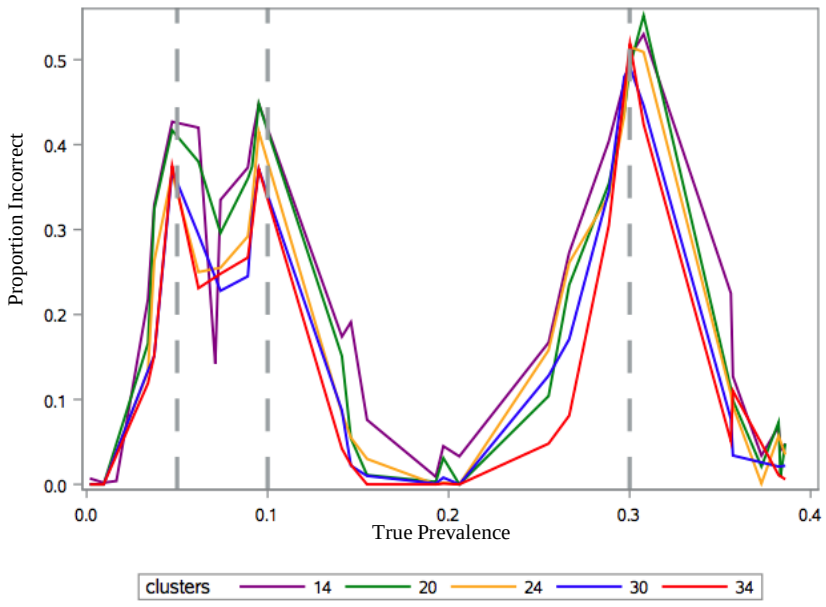
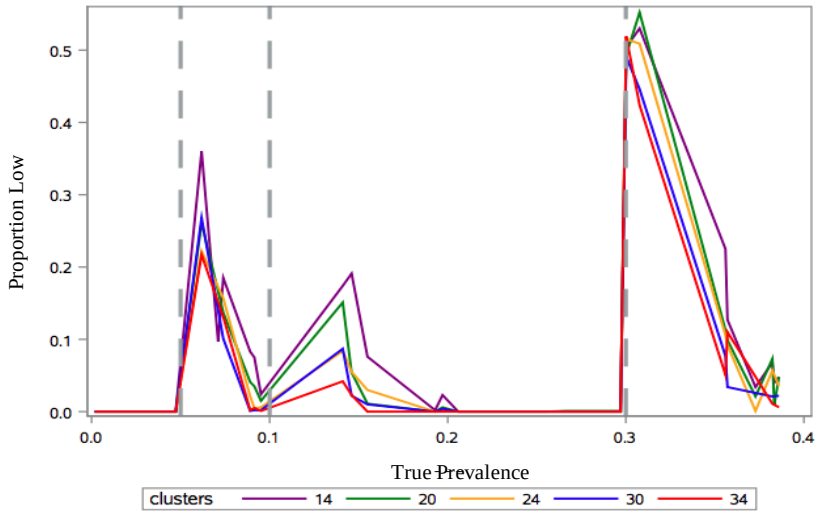


Figure 5 displays the proportion of the 1000 samples that made not only an incorrect MDA decision given the true prevalence, but also a low MDA decision, i.e., the true prevalence warranted 3 rounds of MDA and the sample suggested 1 round of MDA was sufficient. Again, peaks are seen near treatment decision cut points as expected, with the proportion of wrong decisions dropping off between cut points. There is some separation seen at certain prevalence levels across number of clusters, but even less than seen when looking at the incorrect decision plot.

Figure 5. Proportion of Low MDA Decisions by Cluster Number and True Prevalence



Design Effect Results

Figure 6 illustrates the relationship between the design effect and the number of clusters sampled. Three examples of woreda are displayed, though the trend remains consistent throughout all woredas. The empirical design effect stays relatively consistent as the number of clusters increases, which is what is mathematically expected given the formula for the theoretical design effect. Only the number of units in each cluster is a term, the number of clusters selected is not accounted for at all (see design effect methodology). The same does not hold for the effective sample size, displayed in the same way in Figure 7. As the number of clusters increase, the effective sample size

increases as well, which predictable given that the overall sample size is increasing. However, the effective sample size increasing does not ensure that the sample is being used efficiently in terms of providing a precise estimate. For instance, measuring many more individuals in a cluster with high prevalence of TF will increase the actual sample size and subsequently the effective sample size, but may not necessarily inform the overall woreda prevalence estimate any more than selecting half that number of individuals in the cluster. Therefore, the percent of the sample used efficiently was calculated for each sampling scheme and is displayed in Figure 8. While there is some fluctuation, in general, there is no clear trend in percentage of the sample used efficiently as the number of clusters increase. Additionally, the

Figure 6. Empirical Design Effect versus Number of

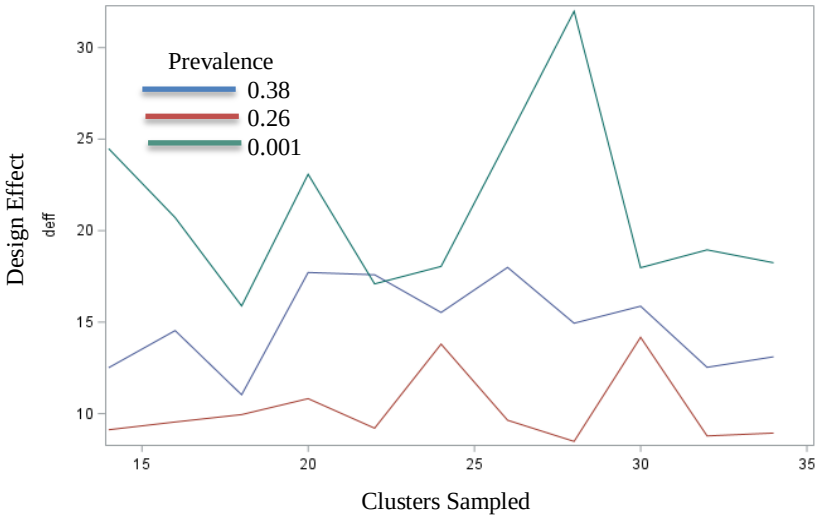
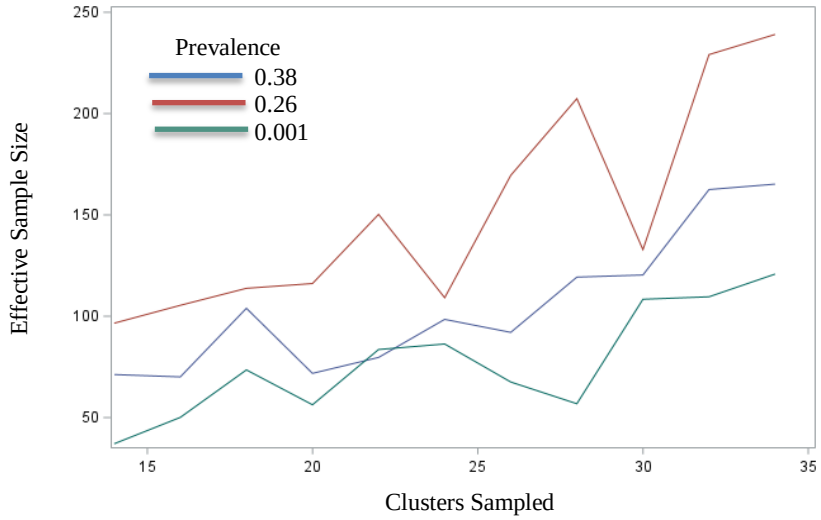
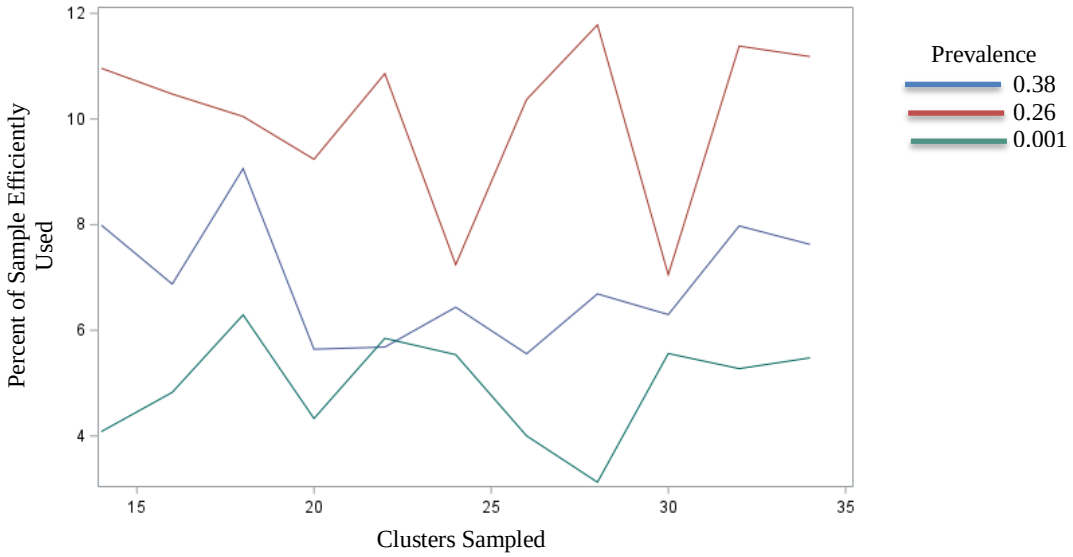


Figure 7. Effective Sample Size versus Number of Clusters



percentages observed in the three example woredas are quite low due to the highly correlated nature of TF clusters.

Figure 8. Percent of Sample Size Used Efficiently versus Clusters

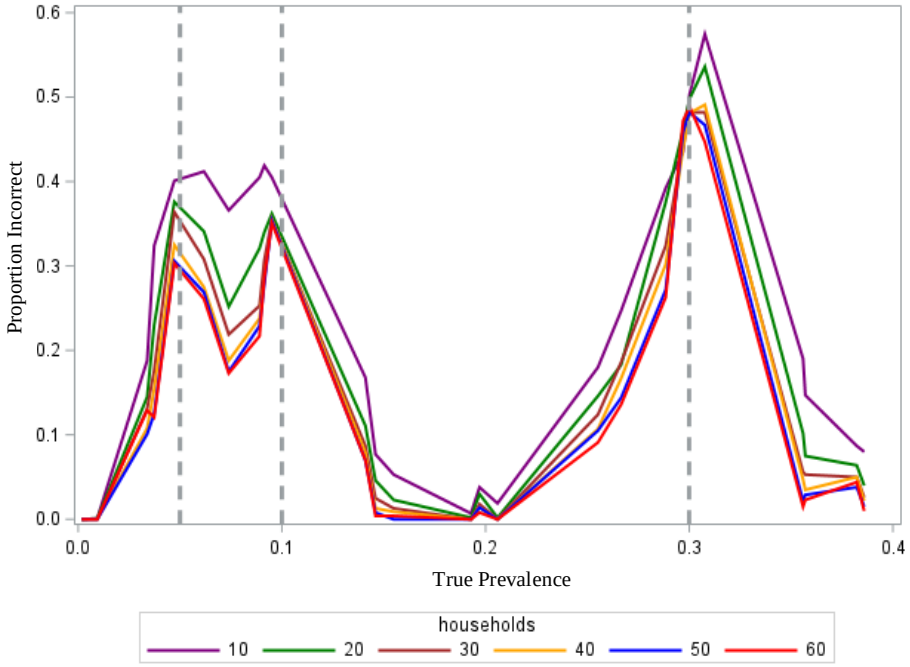


V. Results: Household Level

Figure 9a displays the proportion of incorrect MDA decisions made across the 1000 samples held constant at 30 clusters sampled. Similarly to Figure 4, peaks in incorrect decisions occur close to treatment decision cut points and decrease between treatment cut points. There is very little separation across number of households selected (clusters selected held constant at 30), though the highest amount of separation occurs between prevalence levels of 0.10 and 0.30. Figure 9b displays the proportion of low MDA decisions made, where similar trends are seen compared to Figure 5. Very little separation occurs across number of households selected.

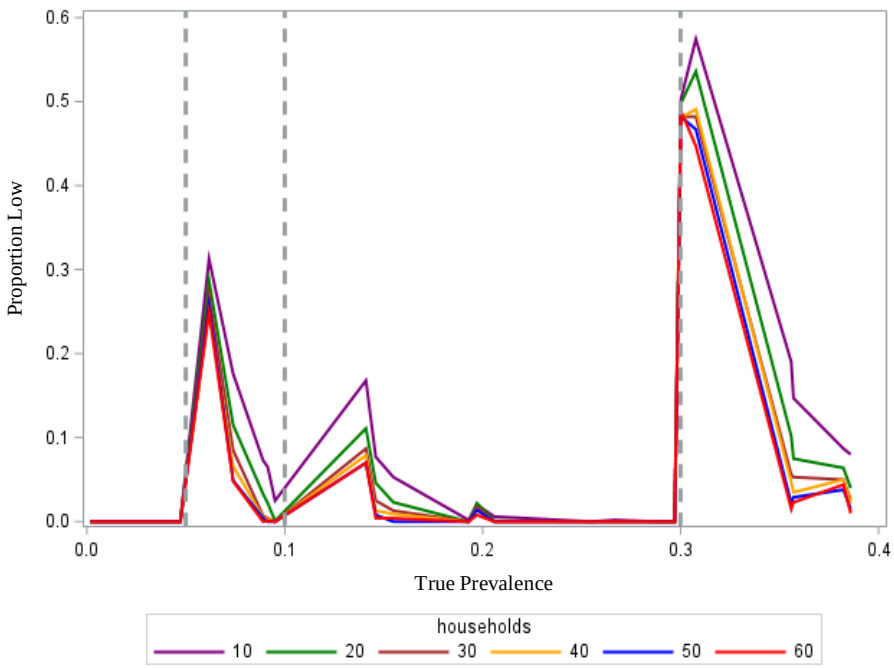
24 different sampling schemes were used to address the second main aim. These 24 schemes were comprised of all combinations of drawing 15, 20, 25, and 30 clusters in the first stage and 10, 20, 30, 40, 50, 60 households in the second stage across all 30 woredas.

Figure 9a. Proportion of Incorrect Decisions by Household Number and True Prevalence



Figures 10a-10d display several representative examples of trends in precision across the 24 sampling schemes within 4 woredas. The rest of the figures for the additional 26 woredas can be seen in the Appendix. Figure 10a illustrates the trends for a woreda with TF prevalence of 0.38. The improvement in precision as a result of increasing number of households selected is minimal in comparison to the improvement in precision as a result of selecting higher numbers of gotts, especially after 30 households. Figure 10b illustrates the trends for a woreda with TF prevalence of 0.10. In this case, increasing the number of clusters beyond 20 does not substantially improve precision. We observe the same diminishing return beyond 30 households. Figure 10c displays a woreda with true prevalence 0.04. Here, we again see that increasing the

Figure 9b. Proportion of Low Decisions by Household Number and True Prevalence



number of clusters beyond 25 does not make a substantial difference in terms of precision. Finally, Figure 10d illustrates the trends for a woreda with TF prevalence 0.01.

The number of clusters selected does not appear to have a great influence on precision, nor does the number of households given that the length of the confidence intervals only drop by less than 0.02 from 10 households to 60 households. Especially beyond 30 households, the improvement in precision is negligible.

Figure 10a. Woreda with True Prevalence 0.38

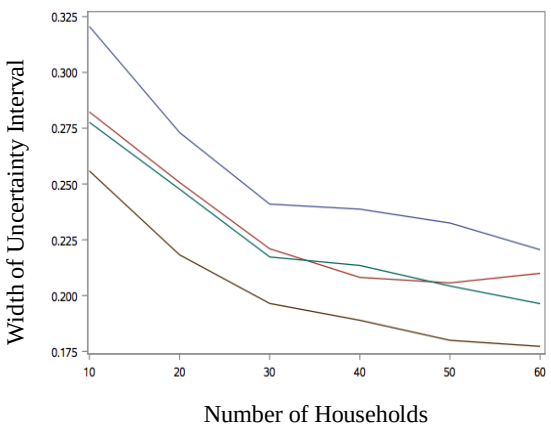


Figure 10b. Woreda with True Prevalence 0.10

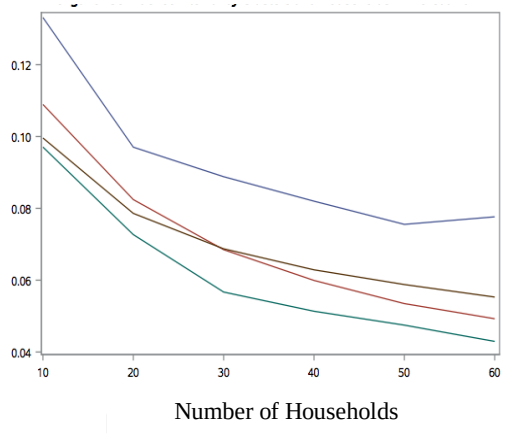


Figure 10c. Woreda with True Prevalence 0.04

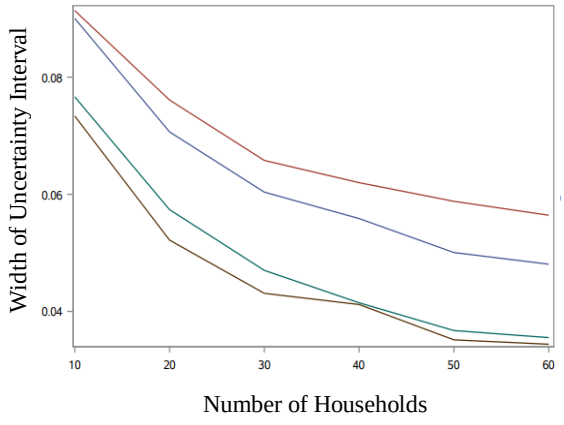
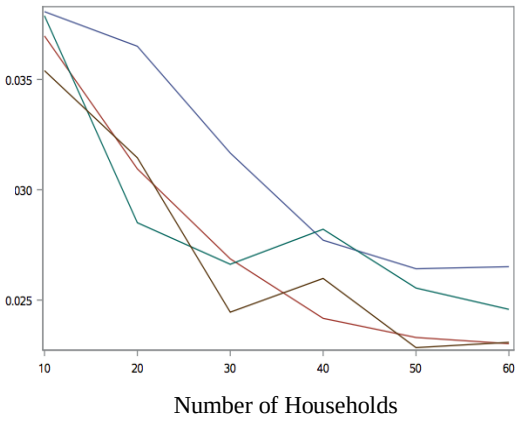
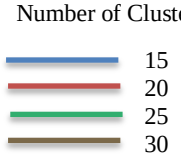


Figure 10d. Woreda with True Prevalence 0.01





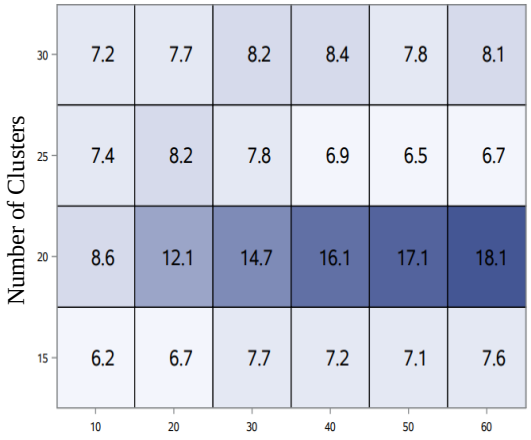
Design Effect and Effective Sample Size Results

The design effect and effective sample size were calculated for each sampling scheme in each woreda. Figures 11a-11c display heat maps of design effect by number of clusters selected and number of households selected for three example woredas; the rest of the

Figure 11a. Design Effects for Woreda With True Prevalence 0.38

Figure 11b. Design Effects for Woreda with True Prevalence 0.15

Figure 11c. Design Effects for Woreda with True Prevalence 0.04



figures for the remaining 27 woredas can be seen in the Appendix. In the first two example woredas, the design effects increase substantially as the number of households increase, and less substantially as the number of clusters increase. In Figure 11c, there is not a real trend seen in the design effects, with the high design effects at 20 clusters likely being due to a unique but random TF distribution within this particular woreda. We repeated this analysis with different samples and saw a similar trend. This brings up an important point: in practice, particular woredas may deviate from the trends

reported in this thesis. We plan to investigate the unique distribution of design effects in this woreda in future research. In general, the design effect is much lower in low prevalence woredas compared to moderate and high prevalence woredas as seen in woredas displayed in the Appendix.. The magnitude of the design effects is noteworthy as they are all much larger than the assumed 2.65 in the GTMP paper[13], which may be due to higher intra-cluster correlation in the simulated population than were assumed for the GTMP calculation.

Figure 12a. Effective Sample Size for Woreda with True Prevalence 0.38

Figure 12b. Effective Sample Size for Woreda with True Prevalence 0.15

Figure 12c. Effective Sample Size for Woreda with True Prevalence 0.04

Figures 12a-12c display heat maps for the effective sample size by cluster and households for the same three woredas (the additional 27 heat maps can be seen in the Appendix part b). In general, the effective sample size increases with both higher numbers of clusters and higher numbers of households selected. The relative increase in effective sample size is greater in low prevalence woredas (Figure 12c) due to

lower variance between clusters. As a result, the effective sample size increases at a much quicker rate when selecting more units. In contrast, in high prevalence woredas (Figure 12a), while the effective sample size increases when selecting more units, the increase is not as substantial. However, for the same reasons as discussed in the previous chapter, the effective sample size is not the best indicator of the most efficient sampling scheme.

Figure 13a. Percent of Sample Efficiently Used for Woreda with True Prevalence 0.38

Figure 13b. Percent of Sample Efficiently Used for Woreda with True Prevalence 0.29

Figure 13c. Percent of Sample Efficiently Used for Woreda with True Prevalence 0.04

Therefore, the percent of sample efficiently used ($\frac{100}{\widetilde{Deff}}$) is displayed in Figures 13a-13c for the same three woredas. In the woredas with prevalence 0.38 and 0.15, it is clear the sample size is used most efficiently when sampling only 10 households per gott. In the low prevalence woreda, while the trend remains the same, all percentages are somewhat similar to one another in magnitude. Sampling schemes do not differ much in terms of efficiency in this woreda.

VI. Cost Analysis Results

Figure 14 displays the approximate costs of sampling schemes using 15, 20, 25, 30 clusters and 10, 20, 30, 40, 50 households, while allowing for a range of per cluster and per household costs. As expected, overall cost increases with higher numbers of clusters as well as with higher numbers of households. In general, the per cluster cost has a much greater influence on the overall cost of the design in comparison to the per household

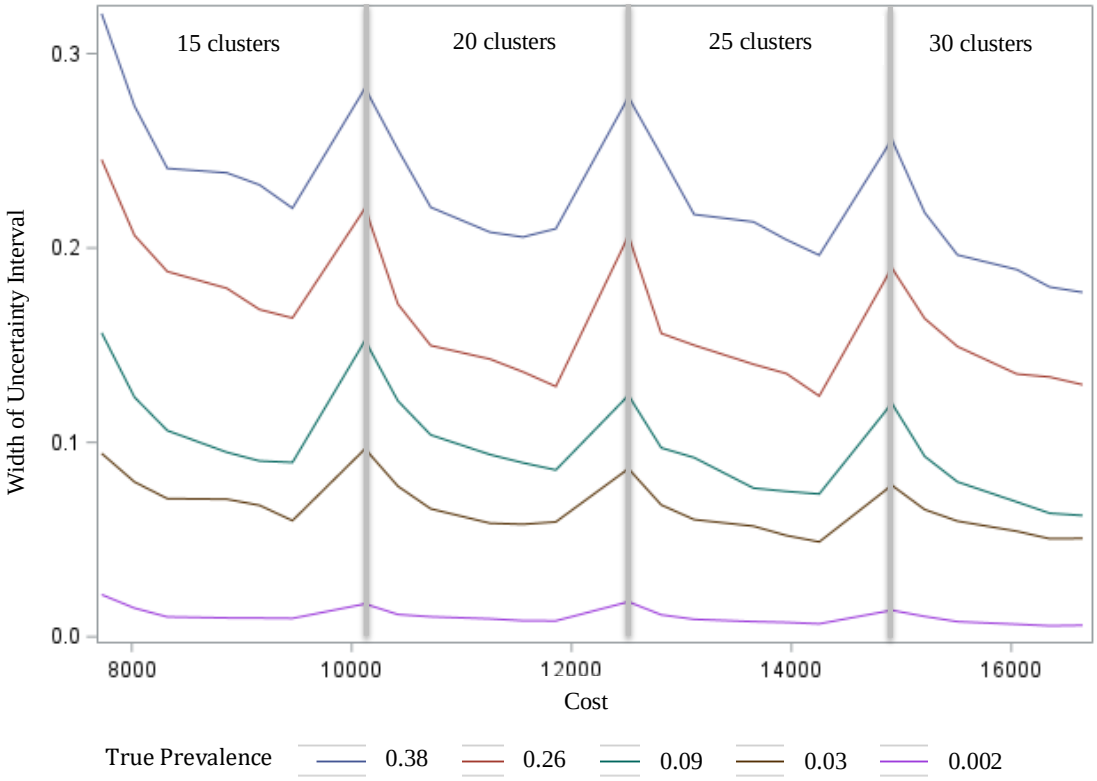
Figure 14. Range of Cost of Sampling Schemes by Number of Clusters and Households

Households		-100	Cost	+100	-100	Cost	+100	-100	Cost	+100	-100	Cost	+100
		15 Clusters			20 Clusters			25 Clusters			30 Clusters		
60	+20	9162.3	10662	12162	11058	13058	15058	12953	15453	17953	14849	17849	20849
	Cost	7962.3	9462.3	10962	9857.8	11858	13858	11753	14253	16753	13649	16649	19649
	-20	6762.3	8262.3	9762.3	8657.8	10658	12658	10553	13053	15553	12449	15449	18449
50	+20	8663.6	10164	11664	10559	12559	14559	12455	14955	17455	14350	17350	20350
	Cost	7663.6	9163.6	10664	9559.1	11559	13559	11455	13955	16455	13350	16350	19350
	-20	6663.6	8163.6	9663.6	8559.1	10559	12559	10455	12955	15455	12350	15350	18350
40	+20	8164.9	9664.9	11165	10060	12060	14060	11956	14456	16956	13852	16852	19852
	Cost	7364.9	8864.9	10365	9260.4	11260	13260	11156	13656	16156	13052	16052	19052
	-20	6564.9	8064.9	9564.9	8460.4	10460	12460	10356	12856	15356	12252	15252	18252
30	+20	7424.5	8924.5	10424	9320	11320	13320	11216	13716	16216	13111	16111	19111
	Cost	6824.5	8324.5	9824.5	8720	10720	12720	10616	13116	15616	12511	15511	18511
	-20	6224.5	7724.5	9224.5	8120	10120	12120	10016	12516	15016	11911	14911	17911
20	+20	6925.8	8425.8	9925.8	8821.3	10821	12821	10717	13217	15717	12612	15612	18612
	Cost	6525.8	8025.8	9525.8	8421.3	10421	12421	10317	12817	15317	12212	15212	18212
	-20	6125.8	7625.8	9125.8	8021.3	10021	12021	9916.9	12417	14917	11812	14812	17812
10	+20	6427.1	7927.1	9427.1	8322.6	10323	12323	10218	12718	15218	12114	15114	18114
	Cost	6227.1	7727.1	9227.1	8122.6	10123	12123	10018	12518	15018	11914	14914	17914
	-20	6027.1	7527.1	9027.1	7922.6	9922.6	11923	9818.2	12318	14818	11714	14714	17714

cost. Additionally, there is significant overlap between costs of sampling schemes when taking into account the potential range of costs.

Figure 15 illustrates the relationship between cost and the width of the uncertainty interval for prevalence estimate for 5 different example woredas (ranging from high prevalence to extremely low prevalence). As cost increases, the overall improvement in precision is limited in all woredas, but especially so in low prevalence woredas. The width of the uncertainty interval spikes in each woreda when the design transitions to a higher number of clusters selected (the main driver of the cost). This trend is suggestive of the high potential to wastefully spend money with the intent of improving precision of estimates. In reality, arbitrarily sampling more individuals does not automatically improve precision due to the intra-cluster correlation of TF, but inarguably increases costs.

Figure 15. Width of Confidence Interval versus Cost: 5 Example Woredas



Next, cost waste was calculated for each sampling design in each woreda. Cost waste was calculated based on the cost estimates and the percentage of the sample size used efficiently using the following formula $Waste = Cost * (1 - p_{used})$, where p_{used} is the percentage of the sample used efficiently. Figures 16a-16c display cost waste for woredas with true prevalence of 0.38, 0.15, and 0.04 respectively. Across the board, the most cost is wasted in sampling when more clusters are selected. Additionally, more cost is wasted when more households are selected as well, though this effect is much smaller in magnitude since households are cheaper to sample than clusters. Cost waste rates increase more quickly in the higher prevalence woredas compared to the lower prevalence woreda.

Figure 16a. Cost Waste by Sampling Design in Woreda with True Prevalence 0.38

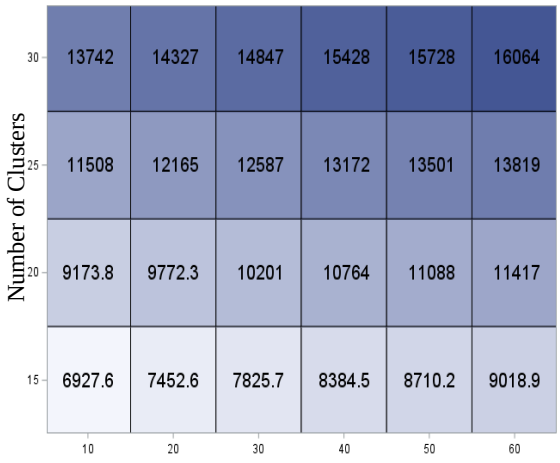


Figure 16b. Cost Waste by Sampling Design in Woreda with True Prevalence 0.15

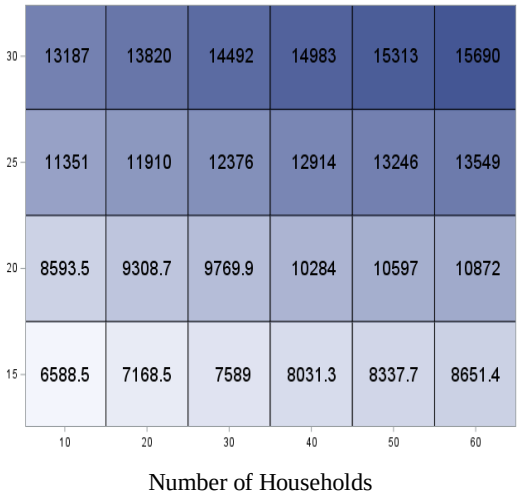
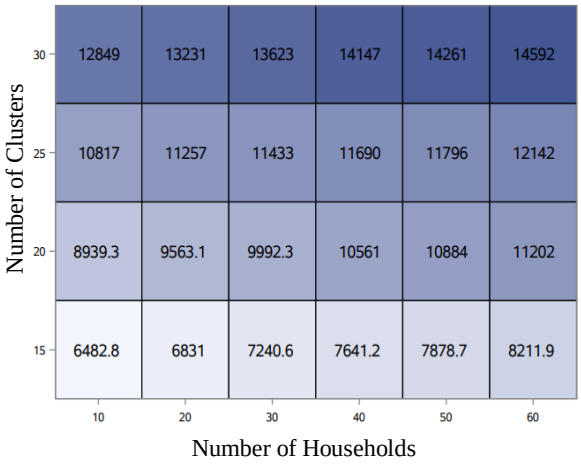


Figure 16c. Cost Waste by Sampling Design in Woreda with True Prevalence 0.04



VII. Future Directions

There are many possible expansions of this work. First, we only addressed woredas with adequate sample size for sampling up to 34 gotts and up to 60 households per gott. In practice, there are several woredas with less than 30 gotts. From a mathematical perspective, it is clearly unnecessary to sample the entire population of the woreda to get an adequately precise prevalence estimate to determine the best MDA course. However, there has been limited statistical analysis with regard to exactly what proportion of the gotts need to be sampled to achieve this precision.

Additionally, the metric of percent of sample used efficiently could be used to evaluate sampling schemes in other neglected tropical diseases such as schistosomiasis and soil-transmitted helminths. While precision and cost-efficiency have been evaluated for these diseases using other metrics [4, 5], percent of sample used efficiently is a novel method for comparing the efficiency of sampling schemes that would greatly inform those funding the monitoring of these diseases.

The unique distribution of design effects as seen in Figure 11c is worthy of further investigation. Since other woredas display a more expected distribution of design effects and the same procedure was sued to calculate design effects for each woreda, it is possible that there is particularly unique TF distribution within this woreda.

Only TF was evaluated in this paper. Also monitored is the prevalence of TT, the stage of trachoma just before irreversible blindness. Prevalence rates of TT are considerably lower compared to TF[1], resulting in difficulty finding cases with random samples. However, the implications of TT are great enough that just a few cases going unnoticed is considered a public health burden[29]. Therefore, mathematical evaluation

of sampling methods for TT should be performed in future work and should consider the use of non-random samples.

Non-random samples should also be considered for TF surveillance. For instance, as opposed to randomly selecting households to measure within a randomly selected gott, announcing that antibiotics will be available at a certain location and allowing parents and children to come on their own accord would save considerable time and resources. Furthermore, gotts with high prevalence would quickly be determined after several children are positive for TF, at which point evaluation could stop and MDA could be implemented. Spending time and resources to determine whether the gott level prevalence is 0.32 versus 0.38 is a waste: many years of MDA will be implemented regardless.

Lastly, as the prevalence of trachoma in Amhara drops over the coming years, the methods presented in this paper in addition to the SAS macros written to perform the simulations (see the Appendix) will be invaluable tools to reevaluate trachoma sampling methodology.

VIII. Discussion

Cluster Level Analysis

As is mathematically expected, improved precision was observed with the more clusters (gotts) selected. However, these trends yield important results when examined relative to MDA decision cut points. In high prevalence woredas that warrant 5 rounds of MDA before re-surveying (greater than 30% prevalence), 95% uncertainty intervals remain as wide as 20% even when sampling an unrealistically high 34 clusters. While

this observation is initially frustrating to trachoma teams attempting to achieve at least 3% precision in all woredas [13], in practice 3% precision is not needed when the TF prevalence is as high as 35%. The difference between a woreda with 30% TF prevalence and 40% TF prevalence is negligible: both woredas will need MDA for many years to come if past trends in Amhara are any indication [30]. Thus, the question arises why surveying teams are attempting to achieve 3% precision in all woredas by sampling large numbers of gotts when 1) this precision is unrealistic to achieve in the first place, and 2) if it could be achieved, does not substantially improve the quality of treatment decisions in high prevalence woredas. On that note, assuming the prevalence of TF is 10% in suspected high and moderate prevalence woredas is unrealistic given the high burden of TF in much of Amhara. Instead of basing sample size calculations off of 10% ± 3% for moderate and high prevalence woredas, sample size calculations should be based on more realistic numbers for each woreda using past data.

In moderate prevalence woredas (10%-30% prevalence), uncertainty intervals (created using bootstrapping) get down to widths of 10%-15% when sampling 34 clusters, which is still quite wide relative to treatment decisions (Figures 2a and 2b). However, the point remains that if MDA is to continue in a woreda, spending money to resurvey regularly and determine the exact prevalence is likely a high-cost low-reward scenario with respect to precision. Repercussions for making a low MDA decision are limited compared to the cost savings given that the area will be resurveyed in future years.

Getting precise estimates becomes most relevant when determining if a woreda is under the 5% TF threshold, when TF is no longer considered to be a public health

problem. Encouragingly, in the lowest prevalence woredas where TF is almost non-existent, 2% precision is achieved. In these same woredas, precision of estimates barely improves between sampling 14 clusters and sampling 34 clusters due to how little variation there is in prevalence among gotts. Therefore, while it is understandably tempting to lean towards sampling as many gotts as possible in low prevalence woredas to ensure that TF is in fact beneath the 5% threshold, this tendency does not actually result in substantially more precise estimates, as seen in Figure 2c. There will always be a degree of error in sampling. It is a question of how to optimize the resources available to achieve the most precise results, not a question of achieving highly precise results no matter the cost.

Figure 4 gives more insight as to the effect of number of clusters on prevalence estimates. When the true prevalence is near an MDA decision cut point, an incorrect decision will be made in around 50% of samples regardless of how many clusters are selected. When the true prevalence is between treatment cut points, there is some separation seen between number of clusters sampled, but not substantial. While sampling 30 clusters makes more correct decisions than sampling 14 clusters, the increased cost is not always worth the minor improvement in precision, especially when the prevalence is above 10%. When the prevalence is below 10% the separation between numbers of clusters is more distinct, so sampling higher numbers of clusters is more justified. Though, sampling 30 clusters remains inefficient in terms of cost.

Household Level Analysis

As seen in Figures 8 and 9 there is very little separation between household values in the proportion of samples making an incorrect MDA decision, as well as in the

proportion of samples making a low MDA decision. Number of households selected has very little impact on the accuracy of the sample, especially when the true prevalence is low. Sampling more than 30 households has little impact on precision in woredas of all prevalence rates, as seen in Figures 10a-10d. The number of clusters has more of an impact on precision in moderate and high prevalence woredas, while in low prevalence woredas neither number of households nor number of clusters has a great impact on the precision given the general uniformity of low prevalence woredas.

The above results suggest that having an estimate of prevalence in a given woreda may greatly inform the most efficient sampling scheme since the most efficient scheme varies based on the true prevalence in the woreda. These results suggest that not using available past information on TF in a woreda results in unnecessarily inefficient samples and funds wasted.

As we would expect, the number of clusters selected has little impact on the design effect, while the number of households selected has a major impact. In moderate and high prevalence woredas, the smallest design effect is seen when sampling only 10 households per gott. In low prevalence woredas, there is no trend with respect to design effect and units sampled due to the limited intra cluster correlation in low prevalence woredas. To have an overall TF prevalence below 5%, an overwhelming percentage of gotts must be TF free; thus, continuing to sample households that are TF free causes the sample to become less and less efficient.

With regard to effective sample size, the highest values are seen when sampling the most number of households: 30 clusters of 60 households each. However, this statistic is misleading and does not imply that 30 clusters of 60 households each is the most

efficient sampling scheme. It merely implies that this is the sampling scheme in which the most individuals are sampled. The more useful metric for optimization purposes is the percent of sample effective used, seen in Figures 13a-13c. Across all tiers of prevalence, the sample is used most efficiently when only ten households are sampled per gott. Due to the high intra-cluster correlation, sampling many households per gott is not yielding any more information about the prevalence in that gott, and thus incites inefficiency. It could certainly be argued that sampling another ten or 20 households in a day is not much extra effort or cost for the field team, who likely cannot survey another gott that day anyway. This practicality should be taken into account when making survey design decisions, but so should the mathematical inefficiencies of these schemes demonstrated in this thesis. It should be noted that in order to sample individual households instead of segments, the structure of the sampling process would need to change, though this change would likely be worth the slightly more representative sample achieved by randomly selecting households.

Reappearance of Trachoma

It is a relatively common dilemma that trachoma seems to reappear in woredas where TF was below the 5% elimination threshold in the previous survey. The results from this study provide strong evidence that concluding TF has resurged in these woredas is misguided. Far more likely is one of the two following scenarios: TF was never below the threshold to begin with and the previous survey happened to be a bit low, or TF is still below the threshold and the current survey happens to be a bit high. Even using the 30 by 30 design in woredas with TF around 5% allows 2% or 3% precision, meaning that

surveys returning 4% prevalence and 7% prevalence in consecutive years (for example) would hardly be surprising.

To truly understand the TF prevalence in an area many years of surveying are necessary. If a woreda remains below the 5% threshold for several consecutive years, our confidence that TF is in fact below the threshold increases in accordance with the number of surveys yielding this result.

Cost Analysis

Costs of survey designs are significantly more affected by the number of clusters sampled compared to the number of households. Additionally, overlap between sampling designs when accounting for a range of costs suggests that more conclusive cost analyses may be performed in the future when more is understood on the cost of TF surveys in Amhara. Research on the cost of surveys is crucial to using funding efficiently in Amhara, which will be essential to continuing the fight against trachoma in Amhara for years to come.

Spending more on surveys does not necessarily result in better TF estimates. In high prevalence woredas, there is improvement in precision, but this improvement is negligible given that surveys will be needed in these woredas for many years regardless of whether the prevalence is (for example) 25% or 35%. In low prevalence woredas, the precision hardly improves at all as cost increases, suggesting that cheaper surveys are adequate in woredas where TF is below threshold, as most survey sizes will reflect the absence of TF if TF is truly gone. Surveys must be designed efficiently in order to ensure funds are used effectively. Additionally, the most funding was wasted in designs with high numbers of clusters and households selected. Understanding how to efficiently

spend money on trachoma surveys is crucial not only from a statistical perspective, but from the perspective of potential donors.

Recommendations

Context matters in TF surveillance. Taking all of the above analyses into consideration, we recommend that TF survey designs in Amhara be based on previous information about each woreda. For woredas that are suspected to have TF prevalence below 10%, we recommend 20 clusters in the first stage and 20-30 households in the second stage. 20 clusters yield adequately precise estimates in woredas below the 5% threshold in comparison to 30 clusters. Sampling beyond 30 households does not greatly improve precision, but does require field teams to spend an extra day in the gott.

In moderate and high prevalence woredas (suspected above 10%), we recommend sampling 15 clusters and 20-30 households per cluster. 15 clusters constitute a significant cost savings compared to 30 while still yielding comparable results with respect to precision and probability of making correct MDA decisions.

In woredas where little is known about the TF prevalence, we recommend sampling 20 gotts and 20-30 households. If the woreda is low prevalence, then we achieve adequate precision. If the woreda is high or moderate prevalence, we achieve unnecessarily high precision, but can adjust accordingly for future surveys.

Limitations

There were several limitations to this study. One of the primary limitations was the design of the population dataset and the assumptions associated with it. Only data Amhara were used to approximate prevalence distributions, though other non-profit

groups and governmental agencies have collected many more data on TF in other regions. As a result, the distributions may not be as accurate a representation of the population as desired if these data from other areas differ in their distributions. However, given the empirical data available to us for this project, the simulated population is representative.

An additional limitation was ignoring the segment structure in the household level analysis. Due to the nature of infectious disease, there is a slight clustering effect within neighborhoods (segments) within the gotts themselves. Ignoring the segment structure and taking a simple random sample of households ignores a degree of intra-cluster correlation, resulting in slightly more accurate prevalence estimates than might be seen in practice. However, since the estimates seen would only be less precise in practice, this limitation does not change any of the conclusions reached in this paper.

Furthermore, in this study we ignored sources of non-sampling error such as missingness in measurement error. These errors play a very real role in TF sampling, and can lead to higher chances of incorrect MDA decisions. Due to the direction of this trend, this limitation also does not affect any of the conclusions reached in this thesis.

Another limitation was the lack of substantial information on cost of TF surveys. Exact cost estimates reported in this paper should be regarded with caution, as we made many assumptions in calculating costs that may or may not hold for future sampling surveys. For instance, the per cluster cost may be over or under estimated, as well as the per household cost. However, we expect the relative cost conclusions to more or less hold for future surveys. The number of gotts should have a much greater impact on the cost than number of households in the sample.

Conclusion

Overall, this study provides substantial and pragmatic information on optimal TF survey sampling in Amhara. The methods used in this thesis including the SAS macros can be applied to survey design in TF in areas beyond Amhara as well as to other NTD's.

TF surveys designs in Amhara should take into account previous TF data from the woreda to achieve the most efficient use of funds. Achieving 3% precision in high and moderate prevalence woredas is unrealistic in practice. However, in many instances lack of precision will not affect the MDA decision, and when it does, the difference between 3 and 5 rounds of MDA is not particularly consequential in terms of public health of the region. Additionally, the assumption of a prevalence of 10% in unrealistic in many woredas in Amhara: it is better to use past survey information to inform the prevalence assumption. TF is not going to disappear from high prevalence woredas in a year, or even in three years, so resurveying will be necessary in high prevalence woredas for many years to come. Surveying costs will be incurred in future years regardless of how many gotts are sampled in these high prevalence woredas. Consequently, it makes more sense to save funds for future surveys when TF is close to being eliminated in these woredas than to waste the funds now on unnecessarily precise TF estimates.

In woredas with low prevalence near the elimination threshold of 5%, high precision is possible, but does not substantially improve by sampling more clusters or more households. If TF is essentially eliminated, it will become clear when sampling only 20 gotts and only 20-30 households per gott because the huge majority of children aged 1-9 years will not have TF in a woreda with overall prevalence this low. Given the ability of most survey teams to complete 30 households in one day, the limited costs associated with surveying more households on the same day, and the noticeable

diminishing return in precision beyond 30 households, we recommend continuing to sample 20-30 households per gott.

Additionally, even when using the 30 by 30 design, incorrect MDA decisions are made a portion of the time. Therefore, when woredas that have previously come in below the 5% threshold rise above it in later years, caution should be used in concluding a resurgence of TF. Multiple years of surveys should be used to make program decisions with respect to elimination, not a single survey.

Furthermore, throwing funding at a survey will not guarantee more precise samples if they are not designed to be efficient. Samples should be designed with past evidence of the TF prevalence in mind. To achieve the most mathematically efficient surveys, sampling fewer households is better, allowing more funds to be spent on visiting additional clusters. Further, sampling 30 clusters is not necessary in the grand majority of woredas. This study found that designs with 20 clusters in low prevalence woredas and 15 clusters in high prevalence woredas are likely sufficient for making MDA decisions. This will substantially lower survey costs. Ideally, funds should be used incrementally over many years, since in all practical terms TF is going to be in Amhara for a long time. Overspending funds now may result in decreased donor interest in trachoma when results don’t come quickly. Until TF prevalence estimates in most of Amhara drop considerably, it does not make sense to implement samples designed to achieve large sample sizes since doing so is a waste of funds.

Using the methods and metrics described in this study to continue to evaluate TF survey design will greatly increase the efficiency of trachoma sampling in Amhara and

around the world. Ultimately, by evaluating and improving the systems already in place to fight trachoma, we can achieve our trachoma elimination goals.

References

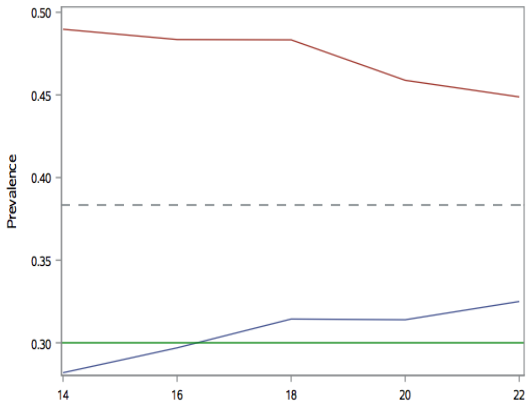
1. WHO. *Trachoma*. 2018 [cited 2018 December 1]; Available from: https://www.who.int/trachoma/disease/en/.
2. WHO, *Trachoma Fact Sheet.* 2018.
3. Solomon, A., *Validation of the Elimination of Trachoma as a Public Health Problem.* WHO, 2016.
4. Sturrock, H.J., et al., *Optimal survey designs for targeting chemotherapy against soil-transmitted helminths: effect of spatial heterogeneity and cost-efficiency of sampling.* Am J Trop Med Hyg, 2010. **82**(6): p. 1079-87.
5. Knowles, S.C.L., et al., *Optimising cluster survey design for planning schistosomiasis preventive chemotherapy.* PLoS Negl Trop Dis, 2017. **11**(5): p. e0005599.
6. Rebecca Mann Flueckiger, P.C., David C. W. Mabey, Rachel L. Pullan, Anthony W. Solomon, *Design and Validation of a Trachomatous Trichiasis-Only Survey*, in *Strategic and Technical Advisory Group for Neglected Tropical Diseases*, W.H. Organization, Editor. 2017.
7. Ngondi, J., et al., *Estimation of effects of community intervention with antibiotics, facial cleanliness, and environmental improvement (A,F,E) in five districts of Ethiopia hyperendemic for trachoma.* Br J Ophthalmol, 2010. **94**(3): p. 278-81.
8. Nash, S.D., et al., *Trachoma prevalence remains below threshold in five districts after stopping mass drug administration: results of five surveillance surveys within a hyperendemic setting in Amhara, Ethiopia.* Trans R Soc Trop Med Hyg, 2018. **112**(12): p. 538-545.
9. Ngondi, J., et al., *Evaluation of three years of the SAFE strategy (Surgery, Antibiotics, Facial cleanliness and Environmental improvement) for trachoma control in five districts of Ethiopia hyperendemic for trachoma.* Trans R Soc Trop Med Hyg, 2009. **103**(10): p. 1001-10.
10. West, S.K., *Azithromycin for control of trachoma.* Community Eye Health, 1999. **12**(32): p. 55-6.
11. Rono, H.K., *Mass treatment for trachoma: how does it all work?* Community Eye Health, 2013. **26**(82): p. 38-9.
12. Solomon, A., *Trachoma Control: A Guide for Program Managers*. 2006, World Health Organization: Geneva, Switzerland.
13. Solomon, A.W., et al., *The Global Trachoma Mapping Project: Methodology of a 34-Country Population-Based Study*. Ophthalmic Epidemiol, 2015. **22**(3): p. 214-25.
14. Anthony W. Solomon, C.K.M., Rebecca M. Flueckiger, Tawfik Al-Khatib, *Design Parameters for Population-Based Trachoma Prevalence Surveys*, in *Strategic and Technical Advisory Group for Neglected Tropical Diseases*, W.H. Organization, Editor. 2018.
15. Missamou, F., et al., *A Population-Based Trachoma Prevalence Survey Covering Seven Districts of Sangha and Likouala Departments, Republic of the Congo.* Ophthalmic Epidemiol, 2018. **25**(sup1): p. 155-161.
16. WHO, *Report of the Third Global Scientific Meeting on Trachoma*. 2010, WHO.

17. Chen, C., et al., *Incremental cost of conducting population-based prevalence surveys for a neglected tropical disease: the example of trachoma in 8 national programs.* PLoS Negl Trop Dis, 2011. **5**(3): p. e979.
18. Trotignon, G., et al., *The cost of mapping trachoma: Data from the Global Trachoma Mapping Project.* PLoS Negl Trop Dis, 2017. **11**(10): p. e0006023.
19. Randall P. Slaven, A.E.P.S., Mulat Zerihun, Eshetu Sata, Tigist Astale, Berhanu Melak, Scott D. Nash, Melsew Chanyalew, Paul M. Emerson, Zerihun Tadesse, E. Kelly Callahan, Deborah A. McFarland, *A cost-analysis of conducting population-based prevalence surveys for the validation of the elimination of trachoma as a public health problem in Amhara, Ethiopia*, T.C. Center, Editor. 2019.
20. Smith, J.L., et al., *Comparing the performance of cluster random sampling and integrated threshold mapping for targeting trachoma control, using computer simulation.* PLoS Negl Trop Dis, 2013. **7**(8): p. e2389.
21. SightSavers. *Tropical Data*. 2019 [cited 2019; Available from: https://www.sightsavers.org/programmes/mhealth/tropical-data/.
22. *Ethiopian Government Portal: Amhara Regional State*. 2018 [cited 2019; Available from: http://www.ethiopia.gov.et/amhara-regional-state.
23. Center, T.C., *TF in Amhara, Fall 2017*. 2017.
24. DeGroot, M.H., *Probability and statistics*. 2nd ed. 1986, Reading, Mass.: Addison-Wesley Pub. Co. xi, 723 p.
25. G.C. Jain, P.C.C., *A Generalized Negative Binomial Distribution.* SIAM Journal on Applied Mathematics, 1971. **21**(4): p. 501-513.
26. Kish, L., *Survey Sampling*. 1965: John Wiley & Sons, Inc.
27. Thomas J. CiCiccio, B.E., *Bootstrap Confidence Intervals.* Statistical Science, 1996. **11**(3): p. 189-228.
28. Cochran, W.G., *Sampling techniques*. 3d ed. Wiley series in probability and mathematical statistics. 1977, New York: Wiley. xvi, 428 p.
29. Rajak, S.N., J.R. Collin, and M.J. Burton, *Trachomatous trichiasis and its management in endemic countries.* Surv Ophthalmol, 2012. **57**(2): p. 105-35.
30. Ngondi, J., et al., *Risk factors for active trachoma in children and trichiasis in adults: a household survey in Amhara Regional State, Ethiopia.* Trans R Soc Trop Med Hyg, 2008. **102**(5): p. 432-8.

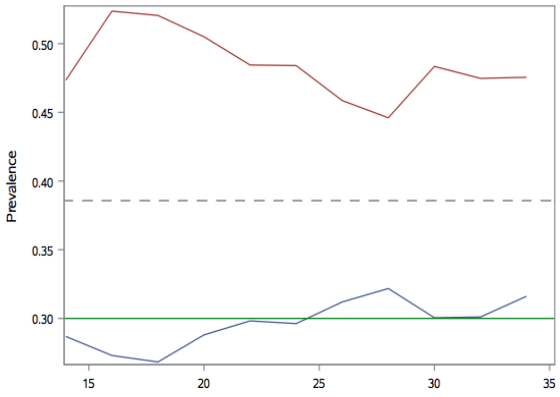
IX. Appendix

Supplemental Figures to Gott Level Results

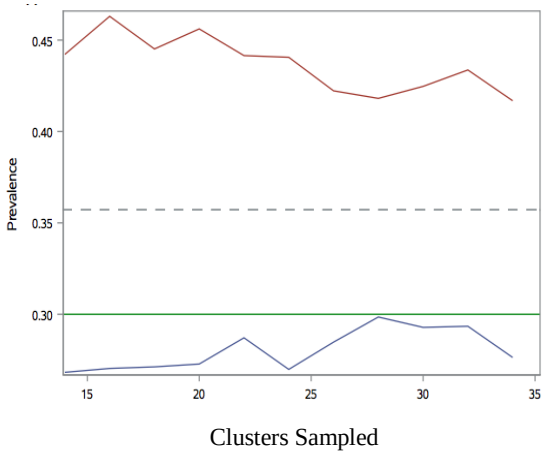
Woreda with True Prevalence 0.38



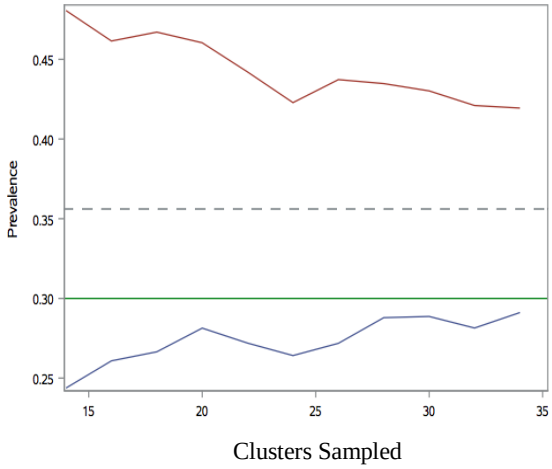
Woreda with True Prevalence 0.39



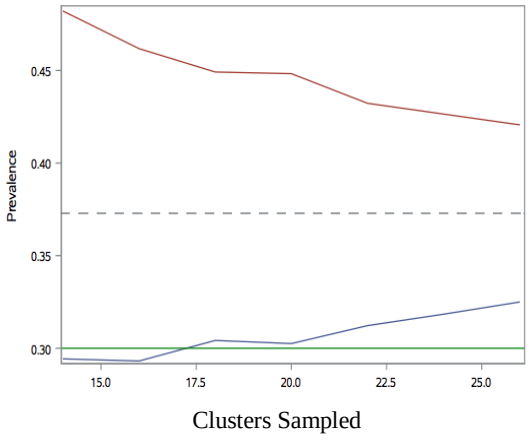
Woreda with True Prevalence 0.36

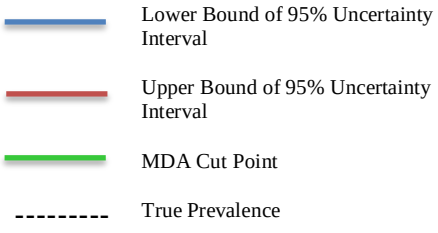


Woreda with True Prevalence 0.36

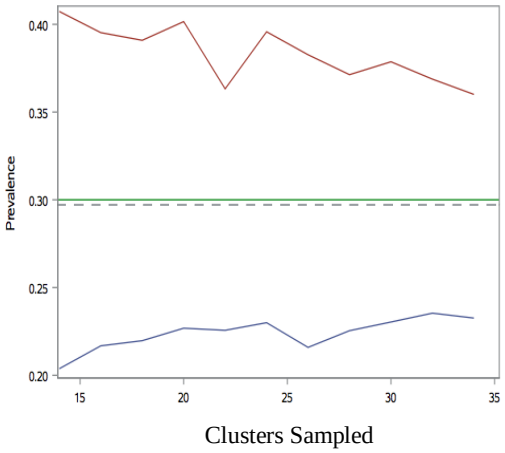


Woreda with True Prevalence 0.37

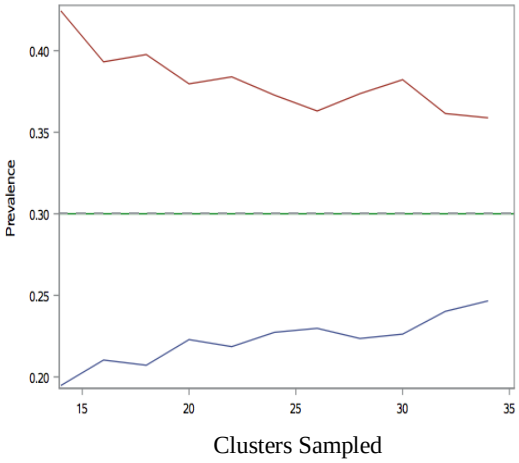




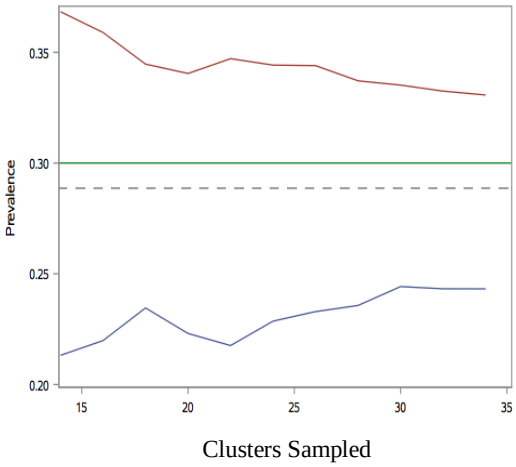
Woreda with True Prevalence 0.30



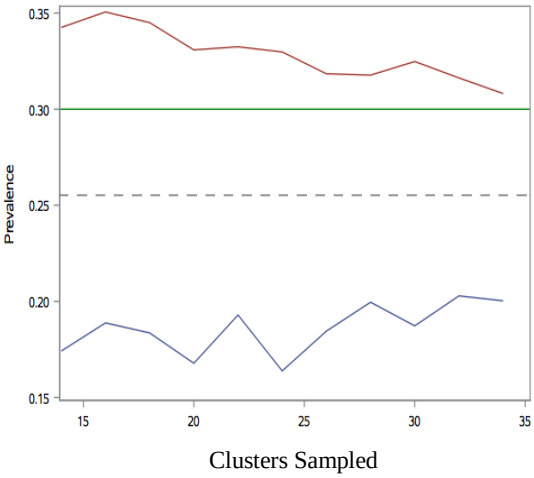
Woreda with True Prevalence 0.30



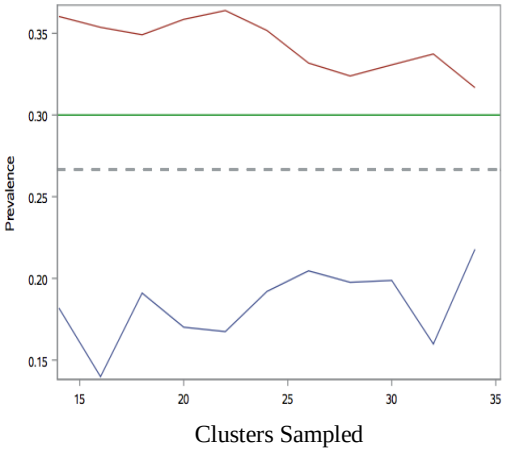
Woreda with True Prevalence 0.29



Woreda with True Prevalence 0.26



Woreda with True Prevalence 0.27



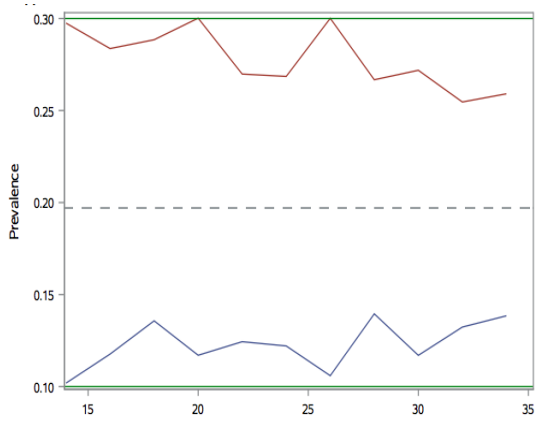
Lower Bound of 95% Uncertainty Interval

Upper Bound of 95% Uncertainty Interval

MDA Cut Point

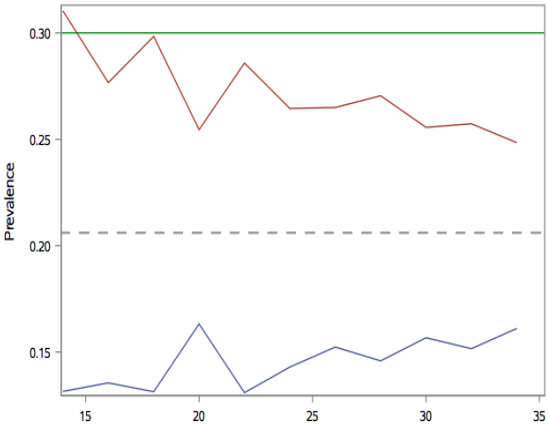
True Prevalence

Woreda with True Prevalence 0.20



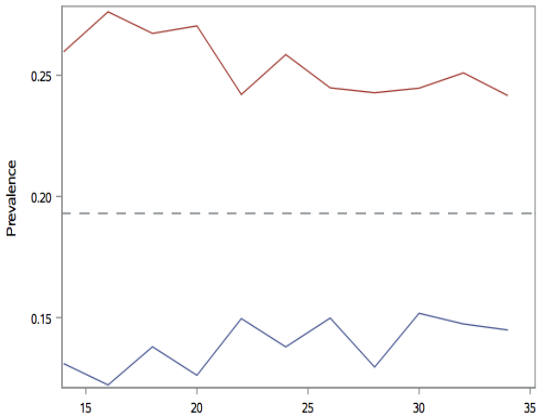
Clusters Sampled

Woreda with True Prevalence 0.21



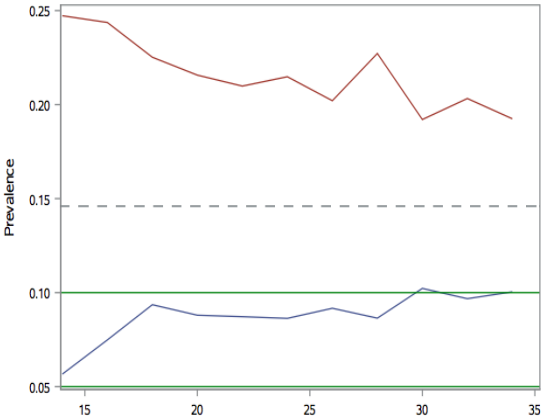
Clusters Sampled

Woreda with True Prevalence 0.19



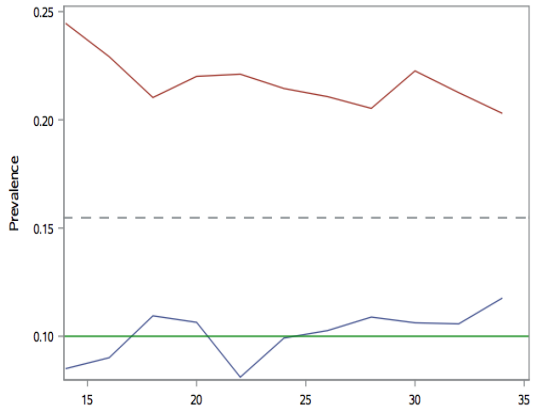
Clusters Sampled

Woreda with True Prevalence 0.15



Clusters Sampled

Woreda with True Prevalence 0.15

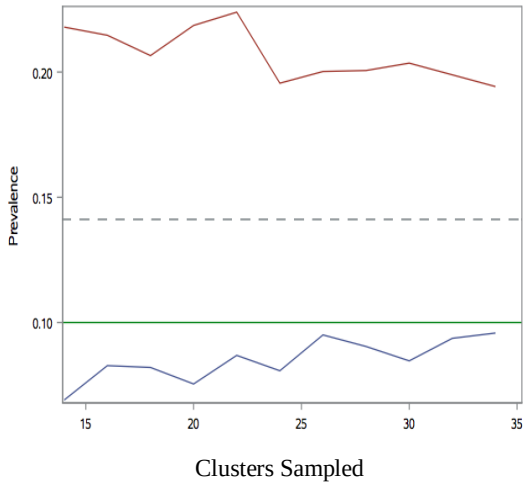


Clusters Sampled

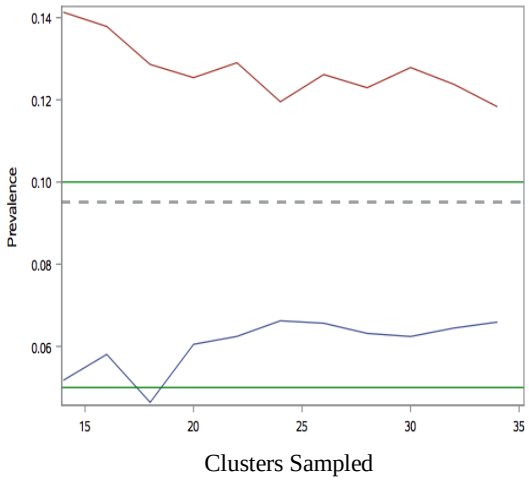
Lower Bound of 95% Uncertainty Interval

Upper Bound of 95% Uncertainty Interval

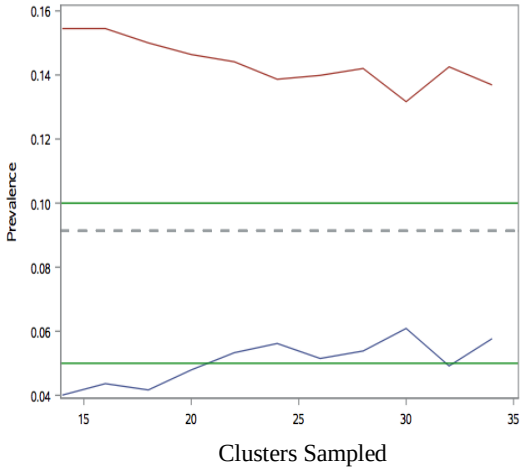
Woreda with True Prevalence 0.14



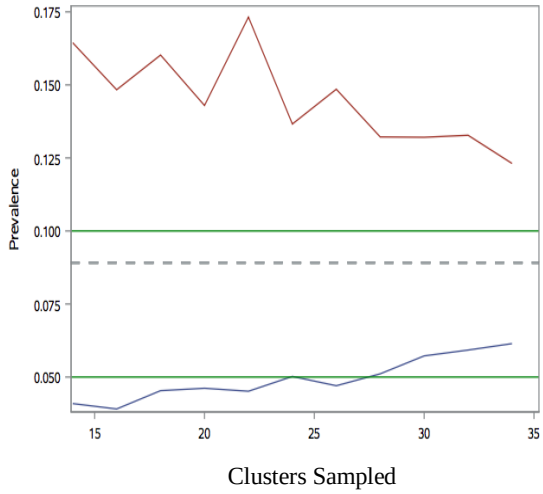
Woreda with True Prevalence 0.09



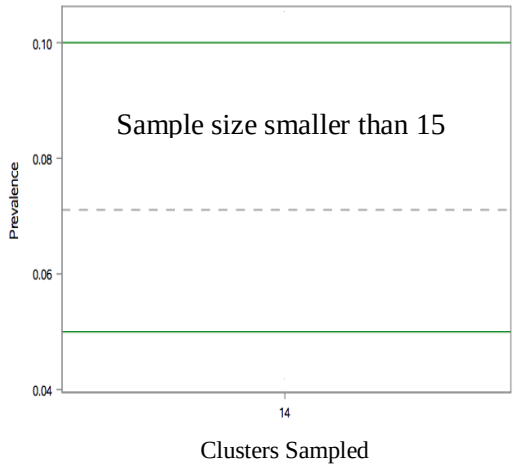
Woreda with True Prevalence 0.09

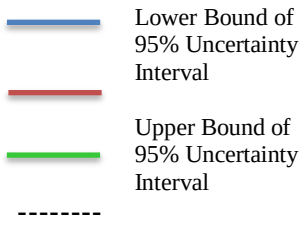


Woreda with True Prevalence 0.09

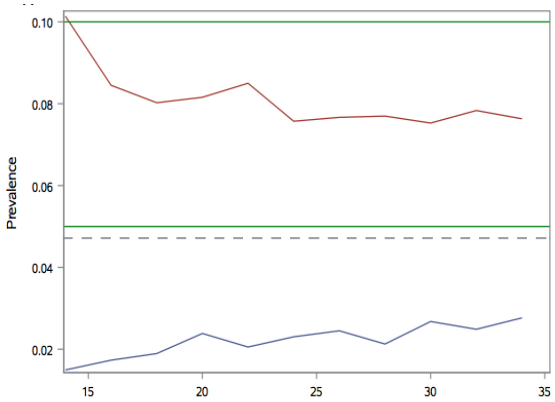


Woreda with True Prevalence 0.07



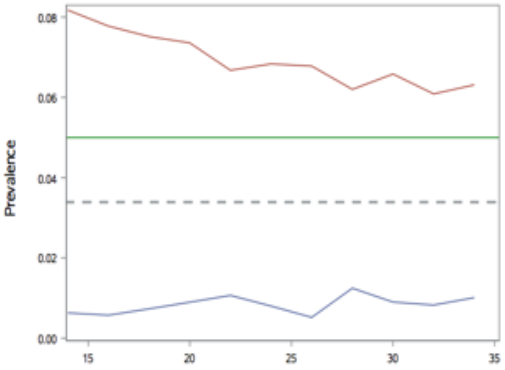


Woreda with True Prevalence 0.05



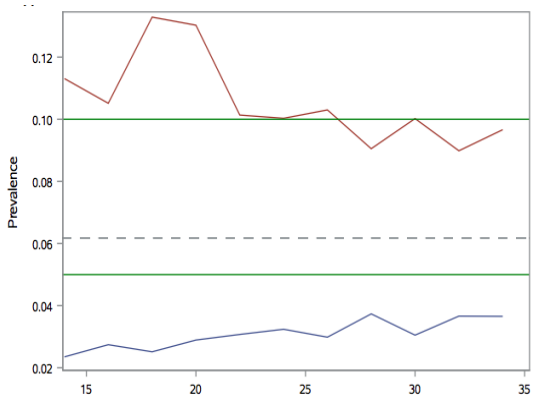
Clusters Sampled

Woreda with True Prevalence 0.03



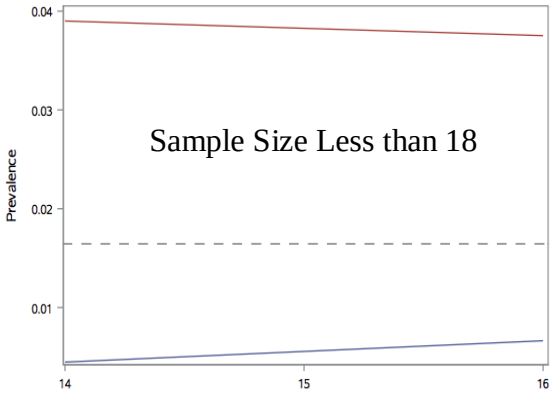
Clusters Sampled

Woreda with True Prevalence 0.06



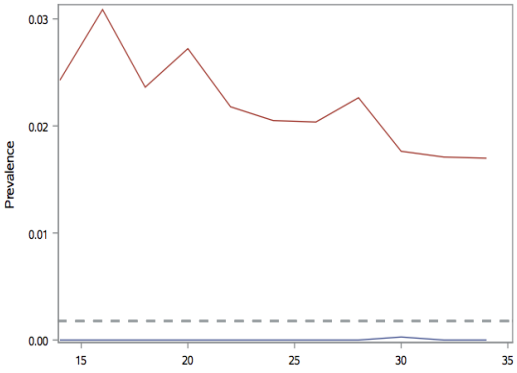
Clusters Sampled

Woreda with True Prevalence 0.02



Clusters Sampled

Woreda with True Prevalence 0.002

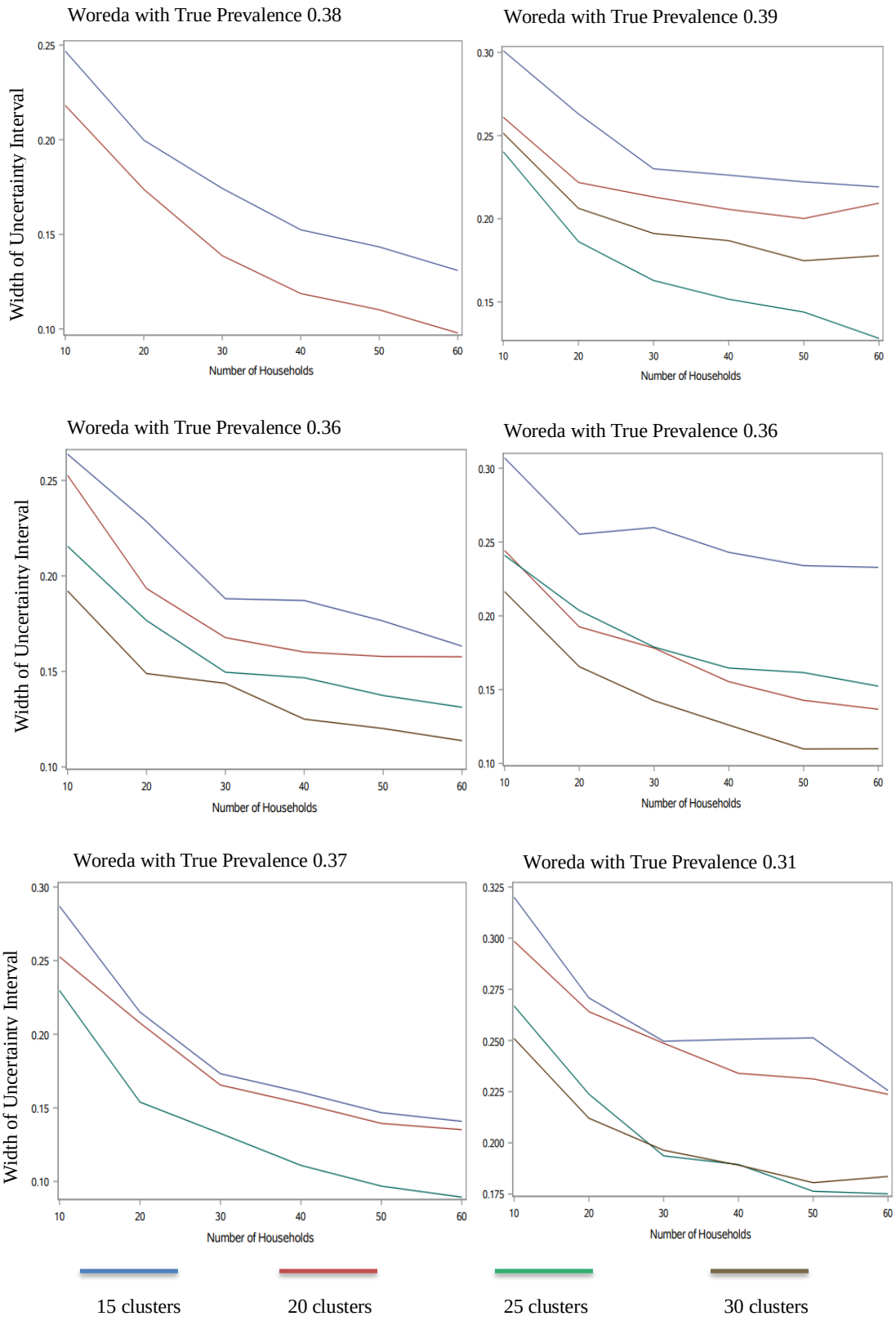


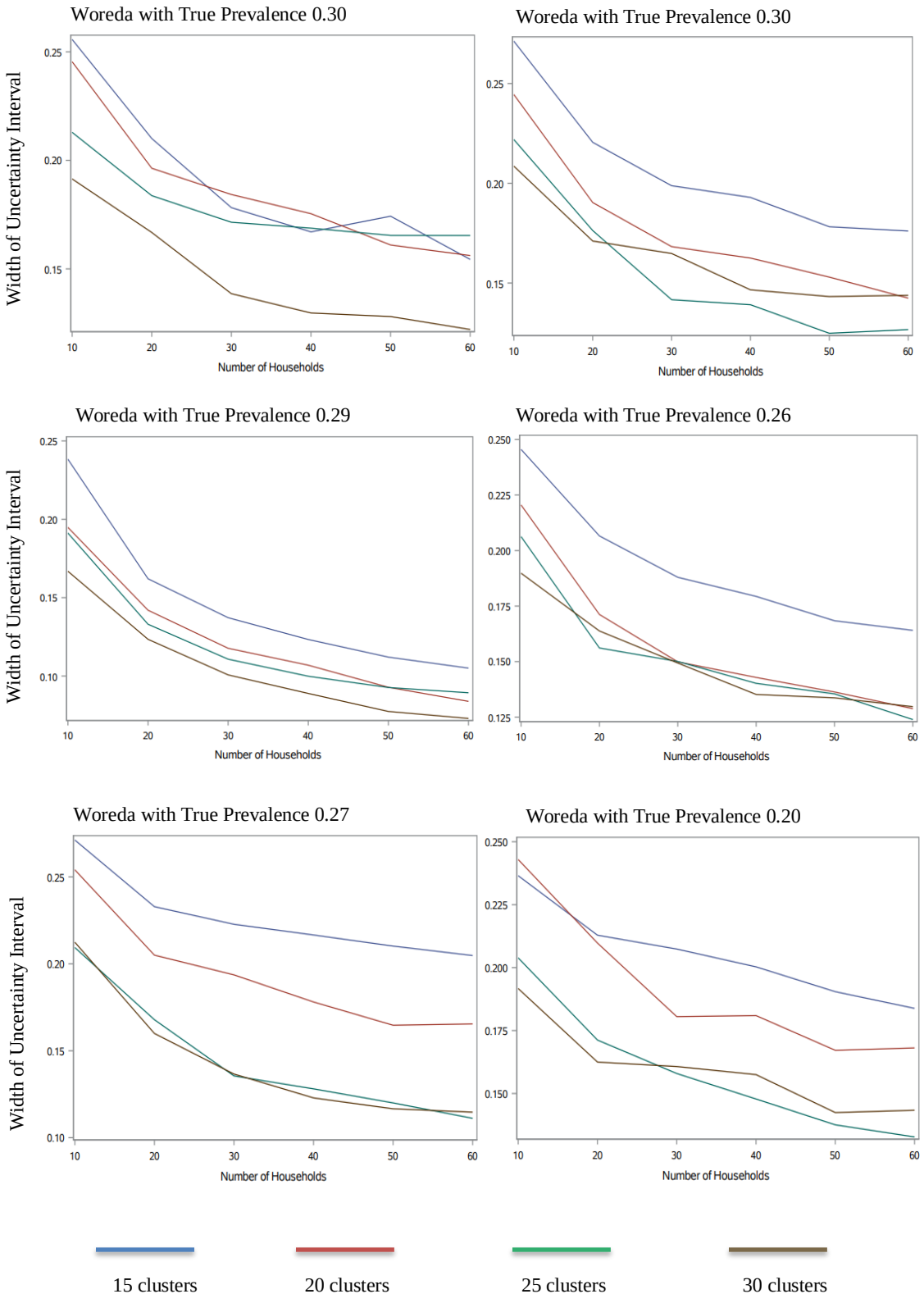
Clusters Sampled

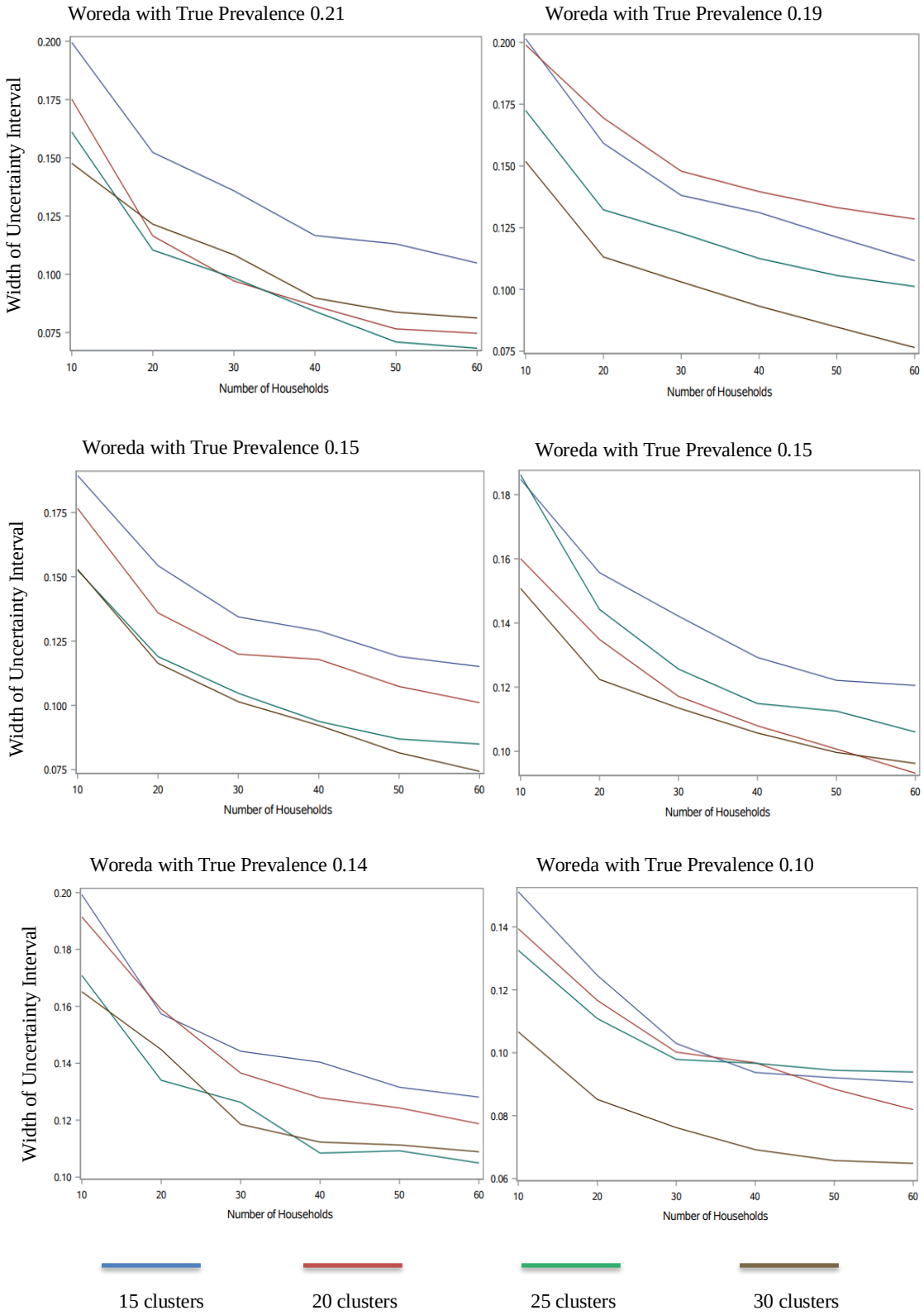
Lower Bound of 95% Uncertainty Interval

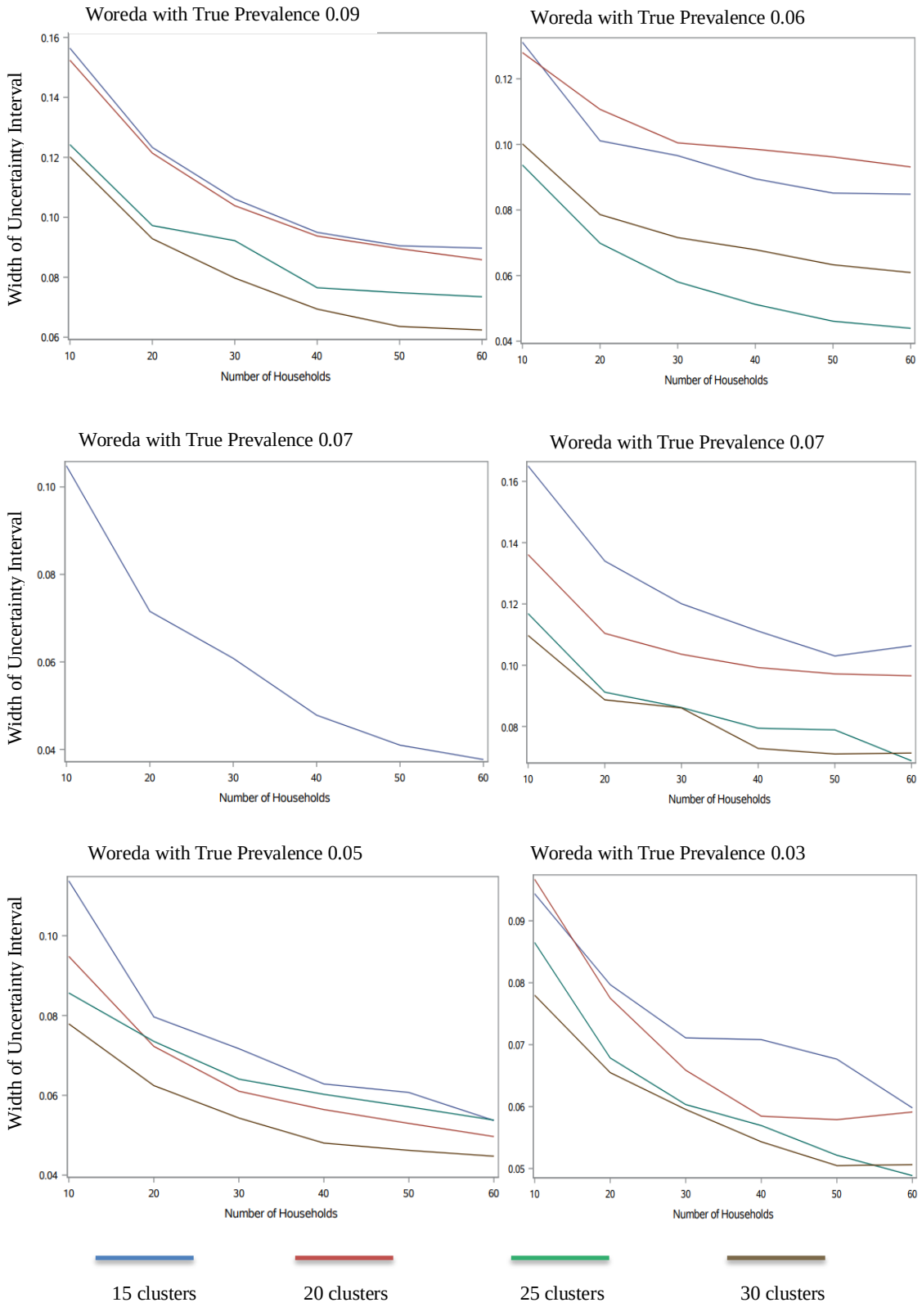
Upper Bound of 95% Uncertainty Interval

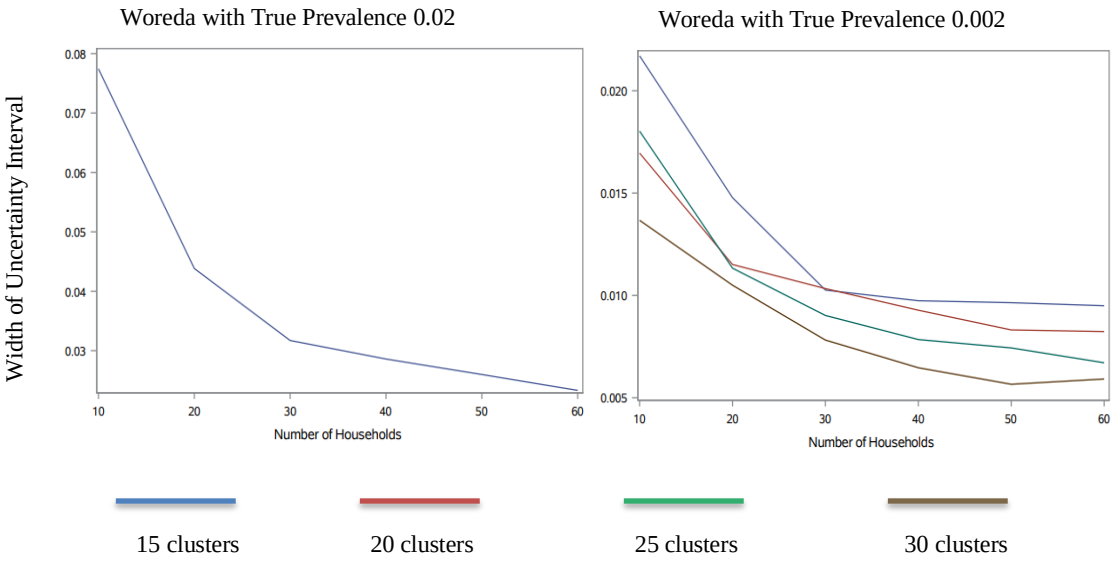
Supplemental Figures for Household Level Results



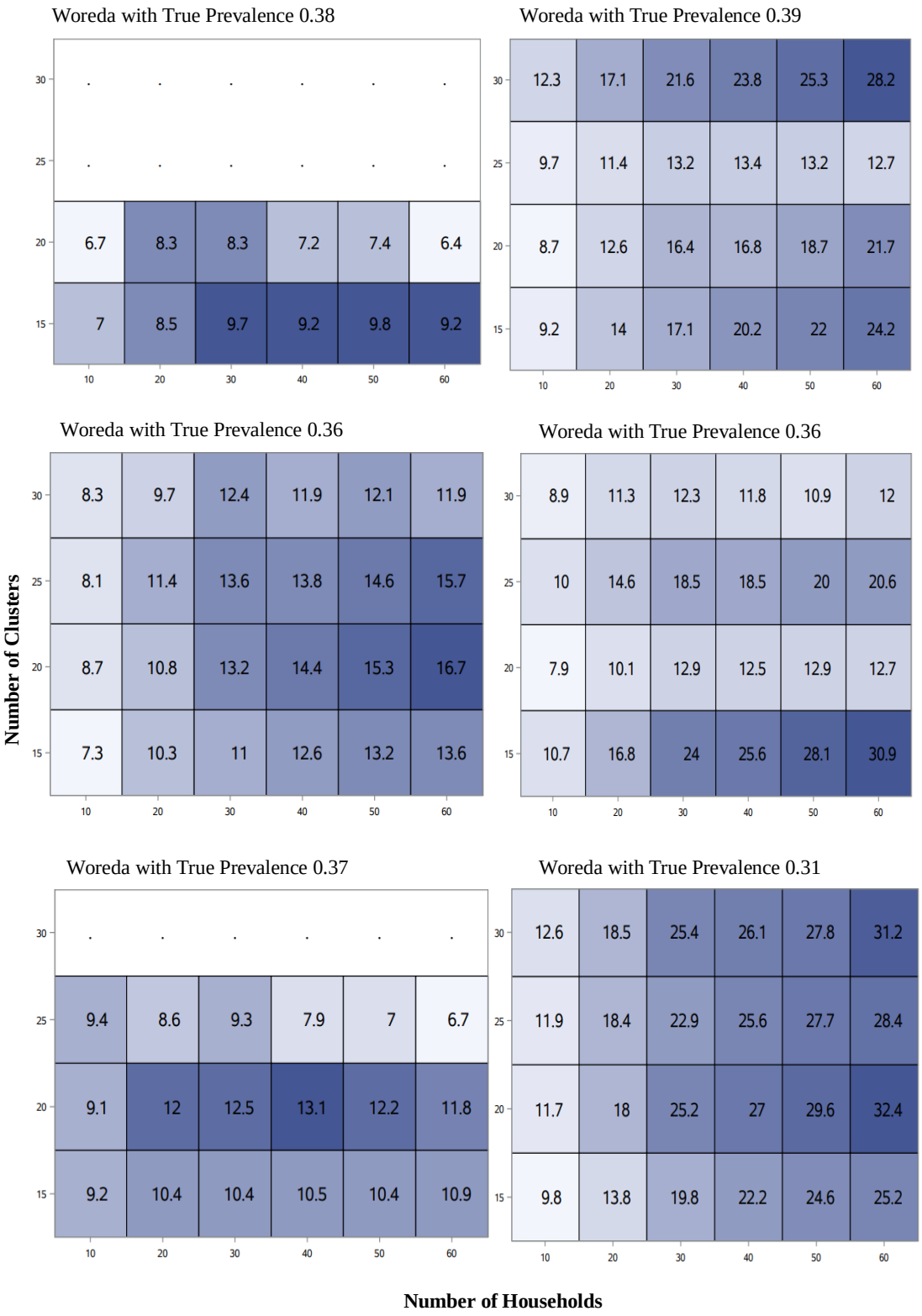

Woreda with True Prevalence 0.30
Woreda with True Prevalence 0.30
Woreda with True Prevalence 0.29
Woreda with True Prevalence 0.26
Woreda with True Prevalence 0.27
Woreda with True Prevalence 0.20
Width of Uncertainty Interval
Number of Households
15 clusters
20 clusters
25 clusters
30 clusters


Woreda with True Prevalence 0.21
Woreda with True Prevalence 0.19
Woreda with True Prevalence 0.15
Woreda with True Prevalence 0.15
Woreda with True Prevalence 0.14
Woreda with True Prevalence 0.10
Width of Uncertainty Interval
Number of Households
15 clusters
20 clusters
25 clusters
30 clusters


Woreda with True Prevalence 0.09
Woreda with True Prevalence 0.06
Woreda with True Prevalence 0.07
Woreda with True Prevalence 0.07
Woreda with True Prevalence 0.05
Woreda with True Prevalence 0.03
Width of Uncertainty Interval
Number of Households
15 clusters
20 clusters
25 clusters
30 clusters

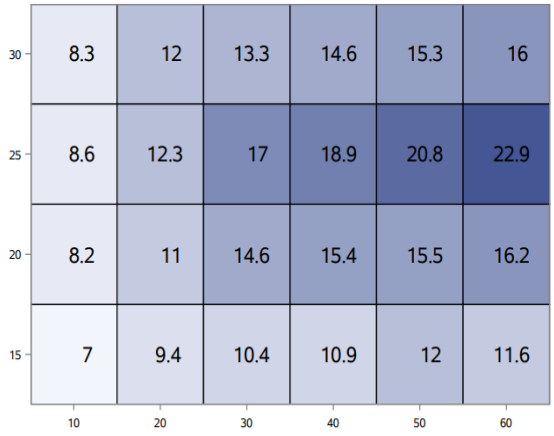

Woreda with True Prevalence 0.02
Woreda with True Prevalence 0.002
Width of Uncertainty Interval
0.08
0.07
0.06
0.05
0.04
0.03
10
20
30
40
50
60
Number of Households
0.020
0.015
0.010
0.005
10
20
30
40
50
60
Number of Households
15 clusters
20 clusters
25 clusters
30 clusters

Design Effects

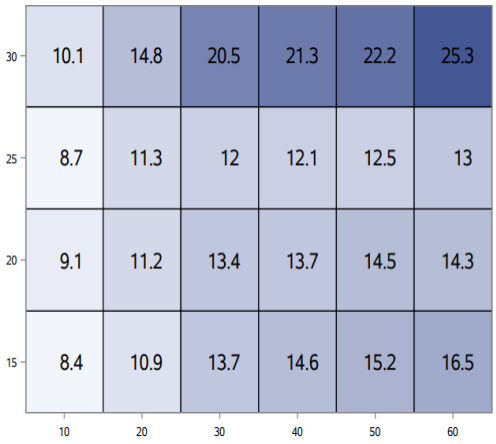


Design Effects

Woreda with True Prevalence 0.30

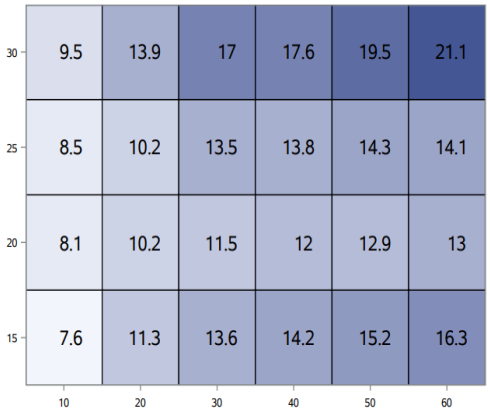


Woreda with True Prevalence 0.30

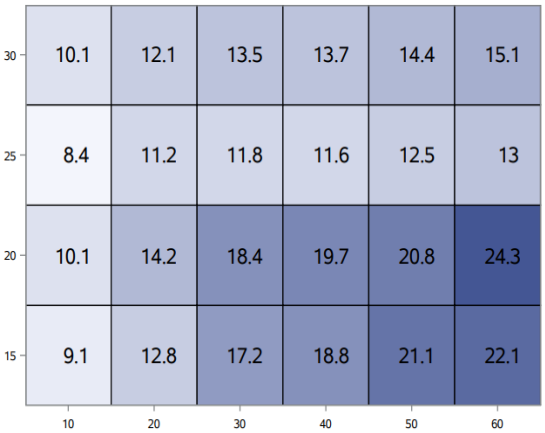


Woreda with True Prevalence 0.29

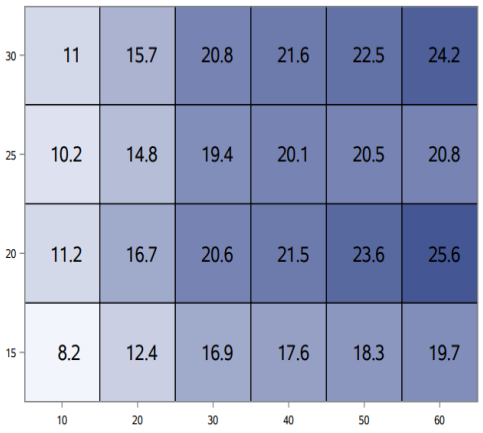
Woreda with True Prevalence 0.26



Woreda with True Prevalence 0.27

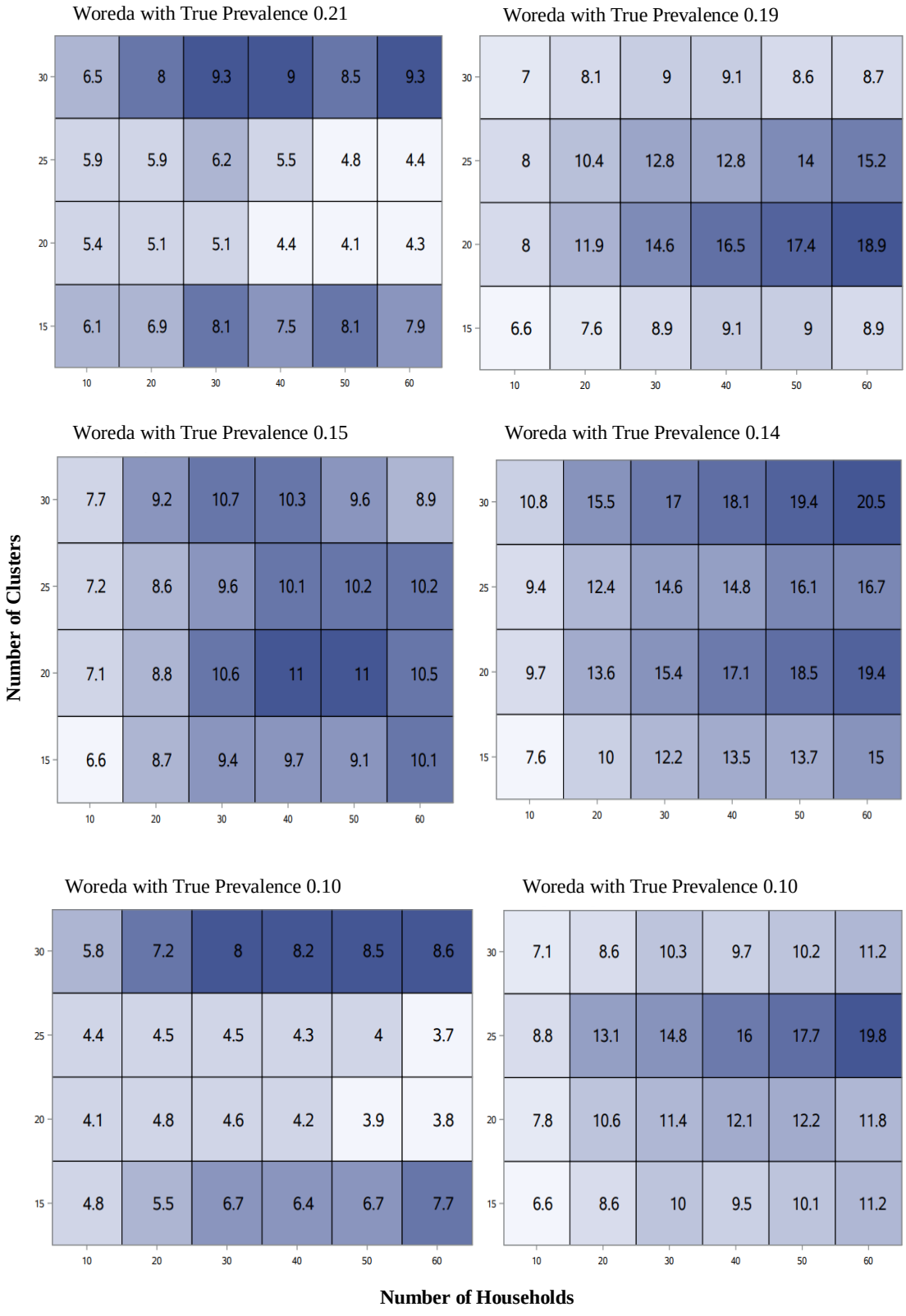


Woreda with True Prevalence 0.20

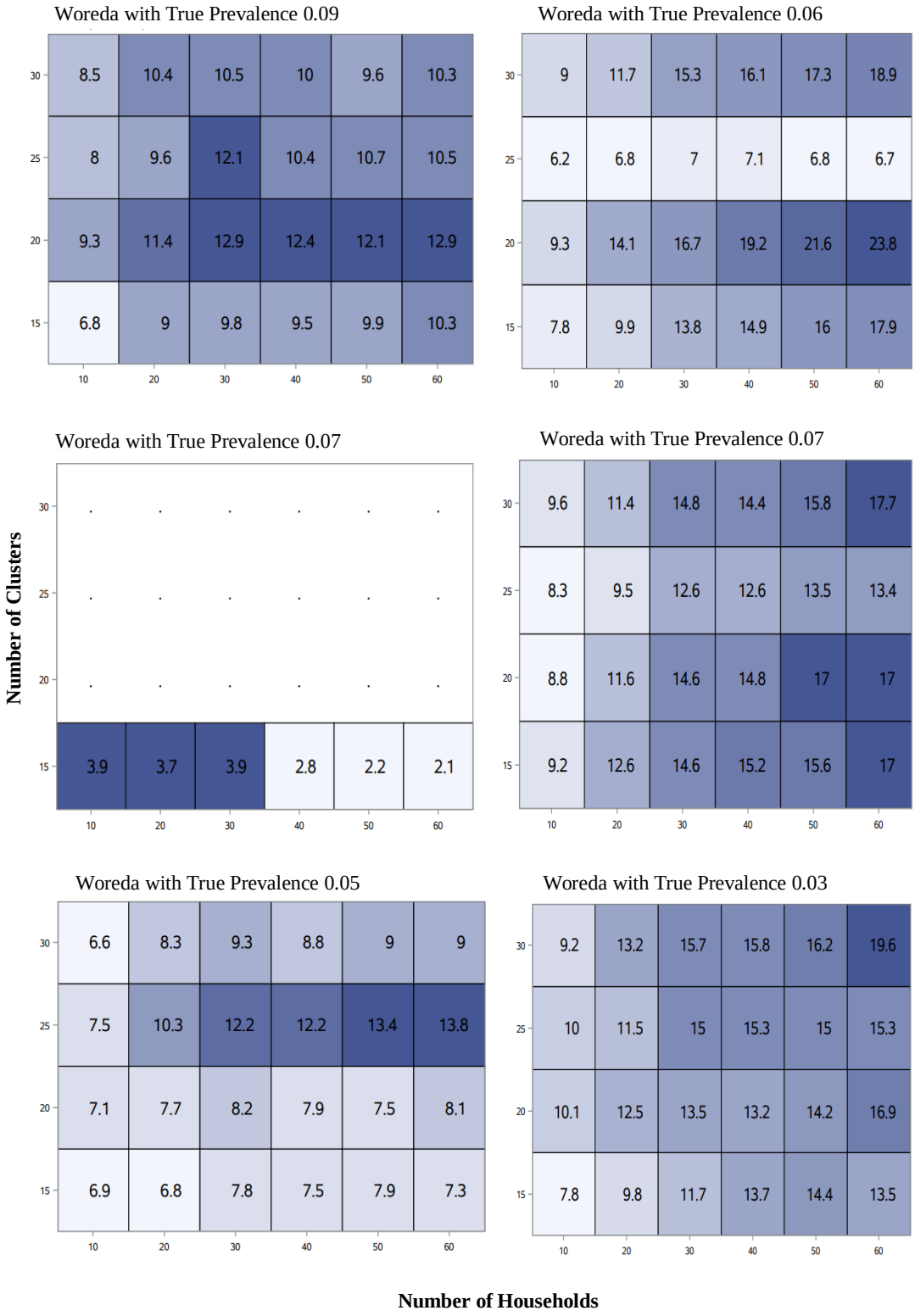


Number of Households

Design Effects

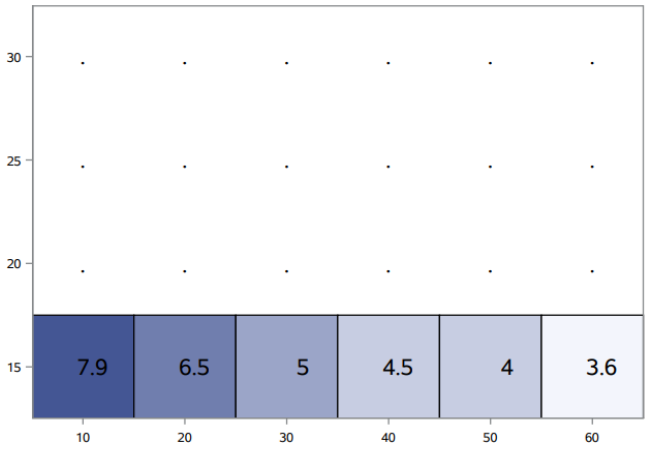


Design Effects



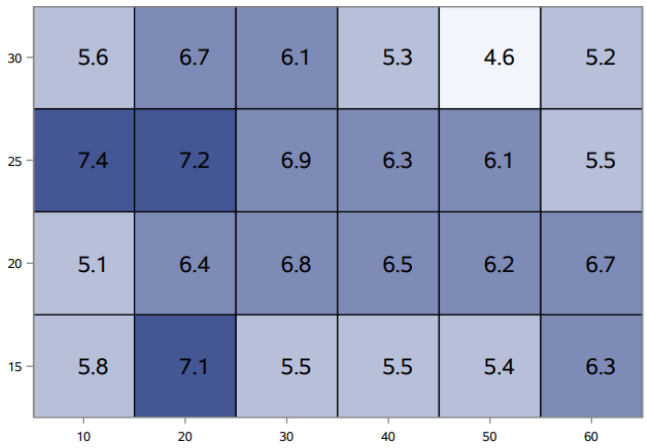
Design Effects

Woreda with True Prevalence 0.02



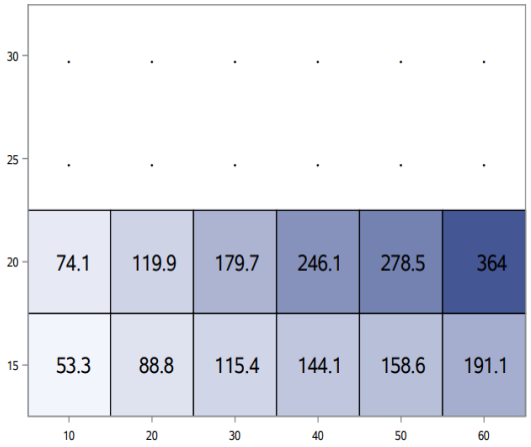
Woreda with True Prevalence 0.01

Woreda with True Prevalence 0.002

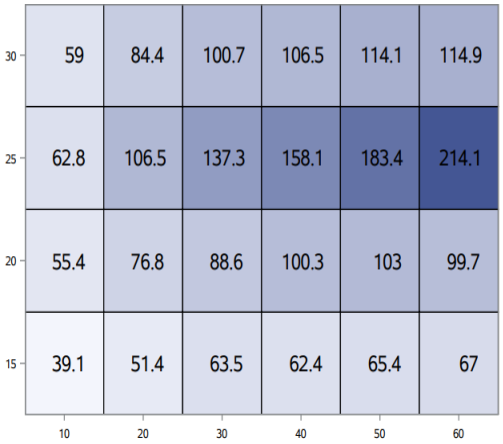


Effective Sample Sizes

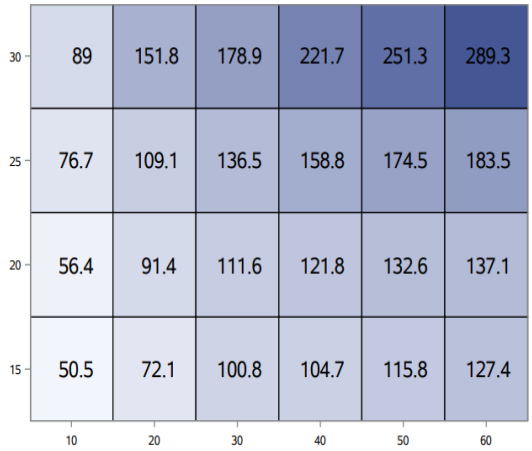
Woreda with True Prevalence 0.38



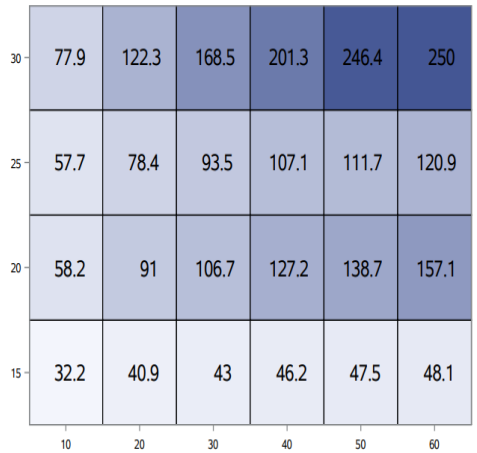
Woreda with True Prevalence 0.39



Woreda with True Prevalence 0.36

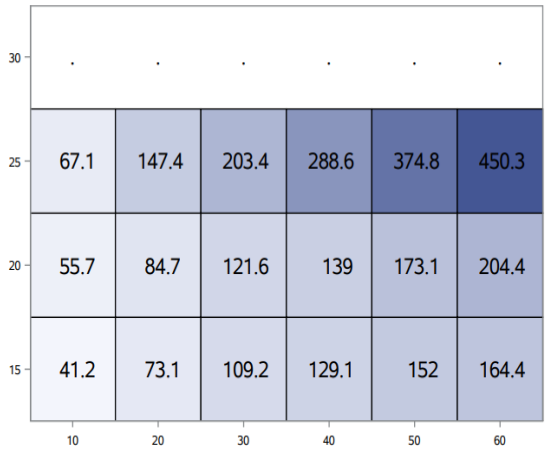


Woreda with True Prevalence 0.36

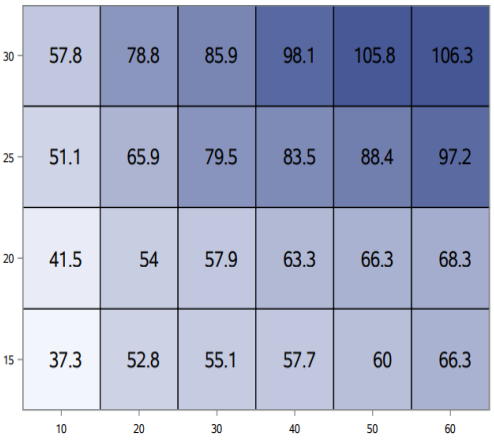


Number of Clusters

Woreda with True Prevalence 0.37



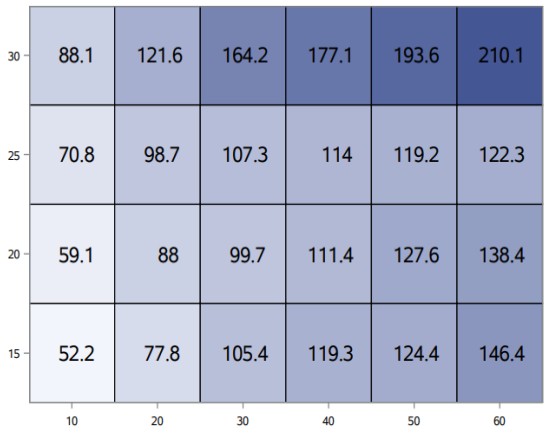
Woreda with True Prevalence 0.31



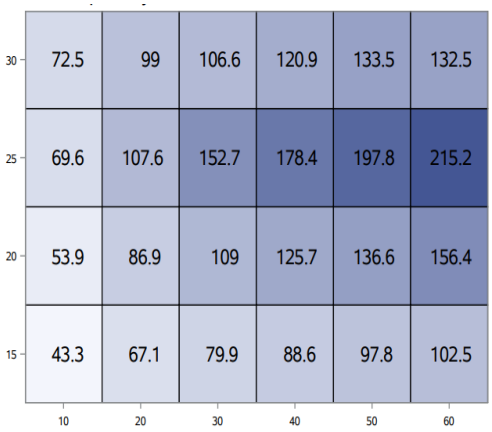
Number of Households

Effective Sample Sizes

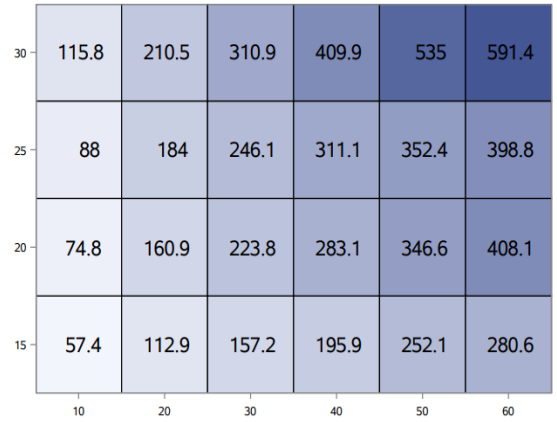
Woreda with True Prevalence 0.30



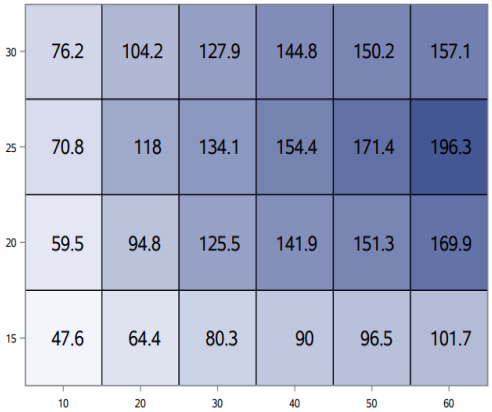
Woreda with True Prevalence 0.30



Woreda with True Prevalence 0.29

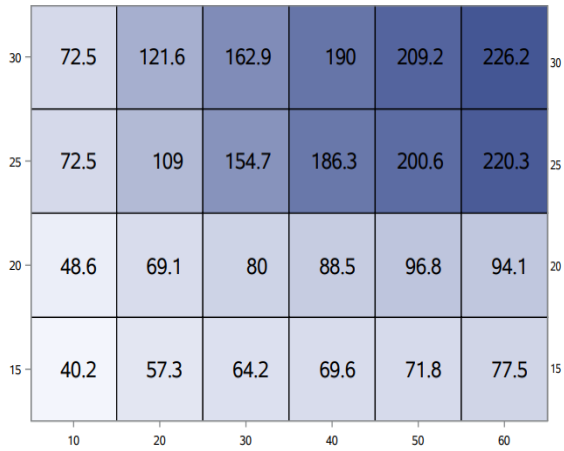


Woreda with True Prevalence 0.26

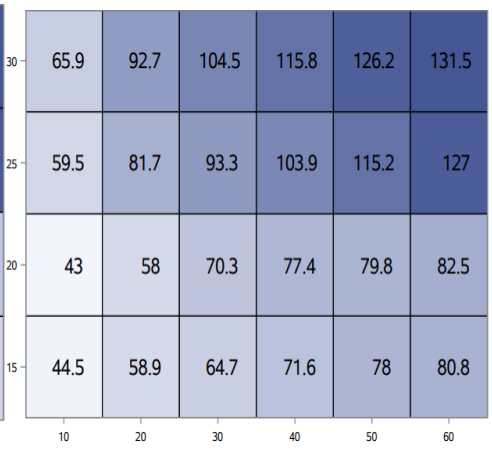


Number of Clusters

Woreda with True Prevalence 0.27



Woreda with True Prevalence 0.20



Number of Households

Effective Sample Sizes

Woreda with True Prevalence 0.21

Number of Clusters \ Number of Households	10	20	30	40	50	60
30	111.1	180	233.3	282.1	339.6	350.9
25	102.3	203.6	289.9	384.1	499.6	622.7
20	88.9	188	284.2	382.9	474.8	513.3
15	59.6	105	133.4	169.4	179	207.6

Woreda with True Prevalence 0.19

Number of Clusters \ Number of Households	10	20	30	40	50	60
30	106	182.6	245.3	290.3	353.8	399.2
25	77.6	118.2	144.8	172	181.8	191.4
20	61.6	83.2	101.3	107	117.1	123.5
15	55.7	96.6	124.7	145	169.1	195

Woreda with True Prevalence 0.15

Number of Clusters \ Number of Households	10	20	30	40	50	60
30	93.7	157	201.8	242	295.6	356.3
25	83.3	139.5	187.1	207.1	233.2	262.5
20	67.3	109.7	136.9	153.2	173.3	203.6
15	54.7	83	115.5	128.3	156.5	158

Woreda with True Prevalence 0.14

Number of Clusters \ Number of Households	10	20	30	40	50	60
30	68.7	95.4	129.6	144.1	155.4	166.8
25	65.5	99.2	126.9	147.9	157	172.2
20	50.7	72.1	95.6	101.7	109	118
15	48.3	73.7	90.8	97	110.1	114.1

Woreda with True Prevalence 0.10

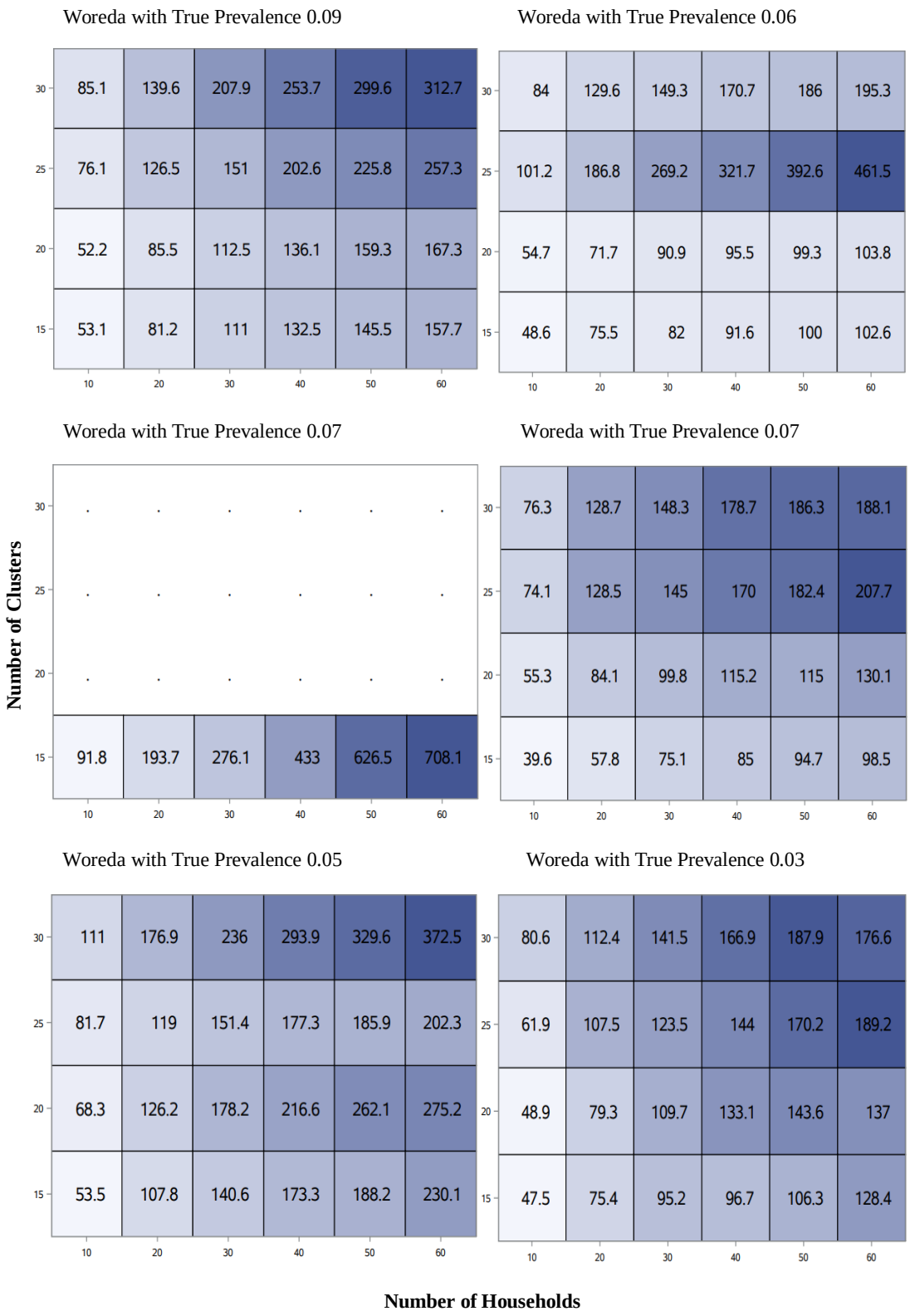
Number of Clusters \ Number of Households	10	20	30	40	50	60
30	121.6	196.5	265.8	299.1	326.9	360.6
25	135.4	262.1	394.2	475.8	585.1	692.6
20	115.7	196	305.1	387.9	478.8	541.9
15	73.5	129	159.2	191	206.8	200.4

Woreda with True Prevalence 0.10

Number of Clusters \ Number of Households	10	20	30	40	50	60
30	103.7	169.6	213.4	266.1	290.3	299.3
25	69.3	93.4	123.6	135.2	139.7	141.8
20	62.2	92.1	129.6	142.9	164.1	191.8
15	55.3	85	109.6	136.4	147.7	150.5

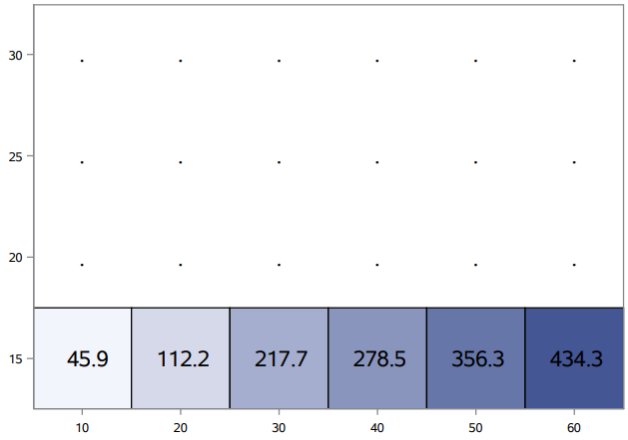
Number of Households

Effective Sample Sizes

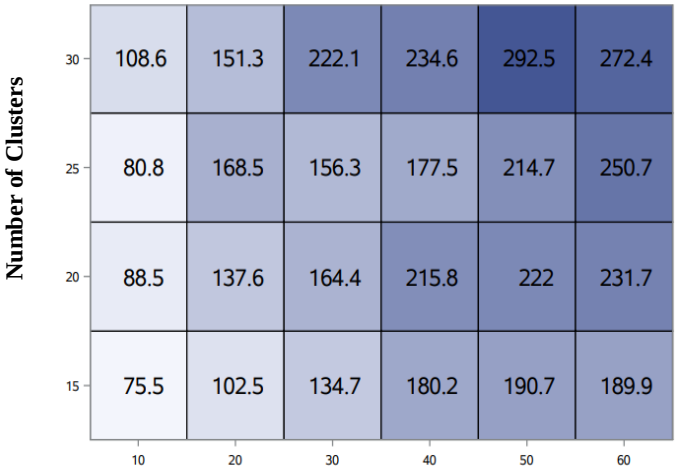


Effective Sample Sizes

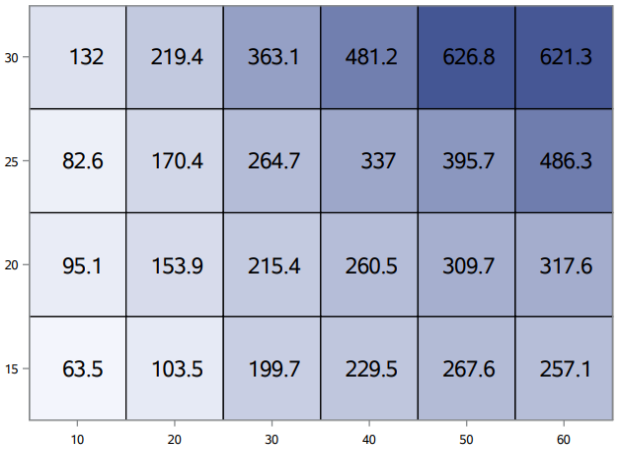
Woreda with True Prevalence 0.02



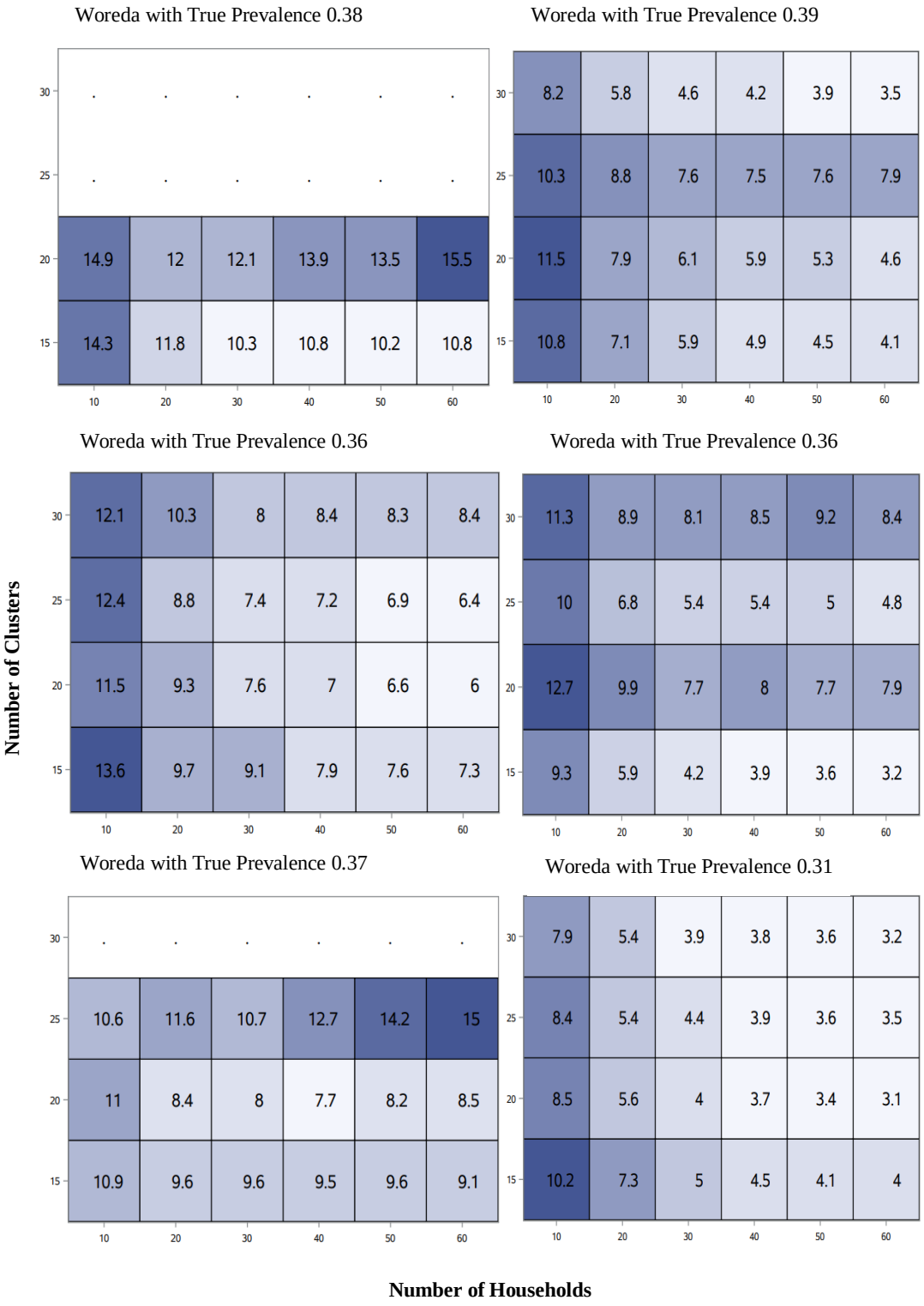
Woreda with True Prevalence 0.01



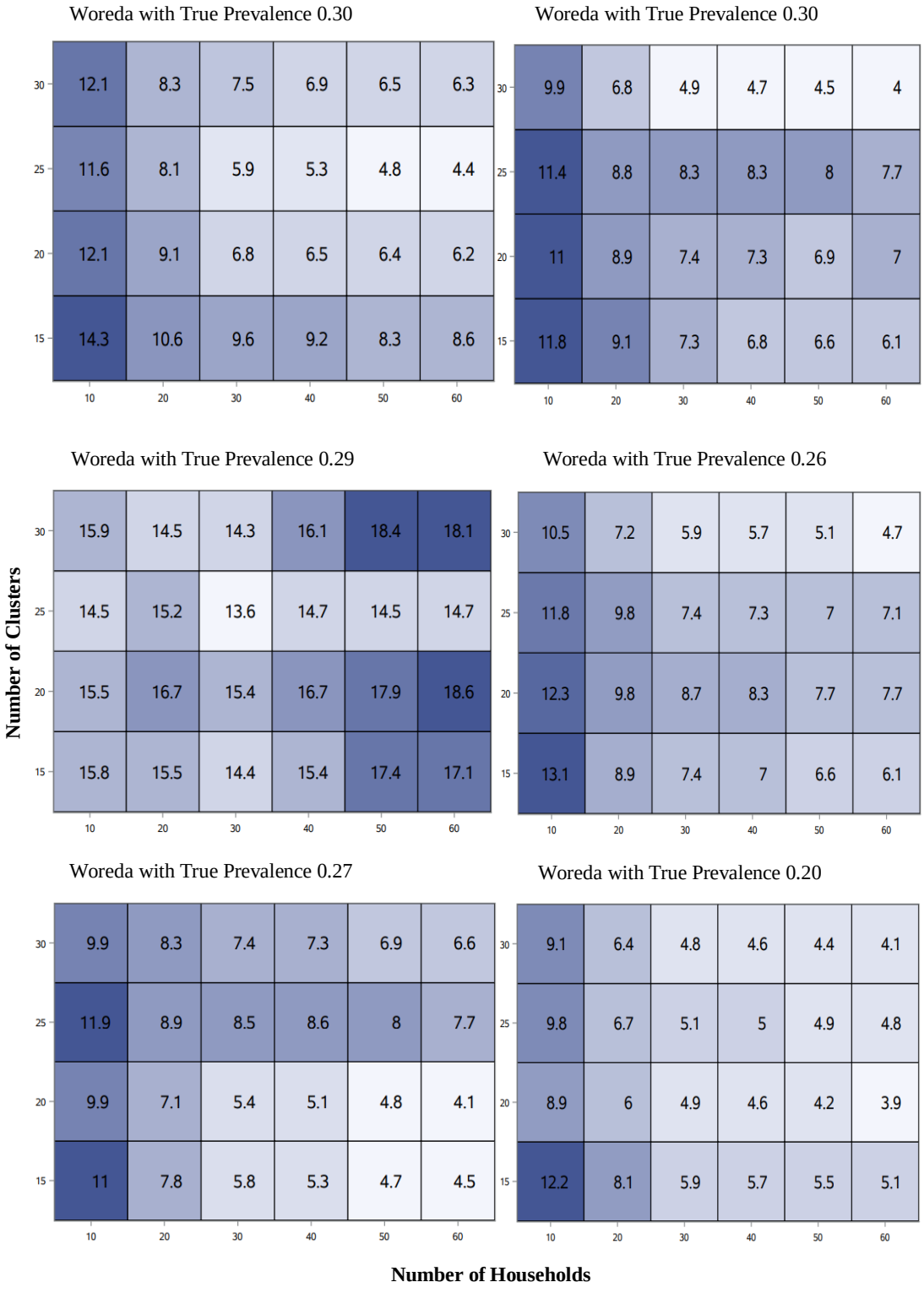
Woreda with True Prevalence 0.002



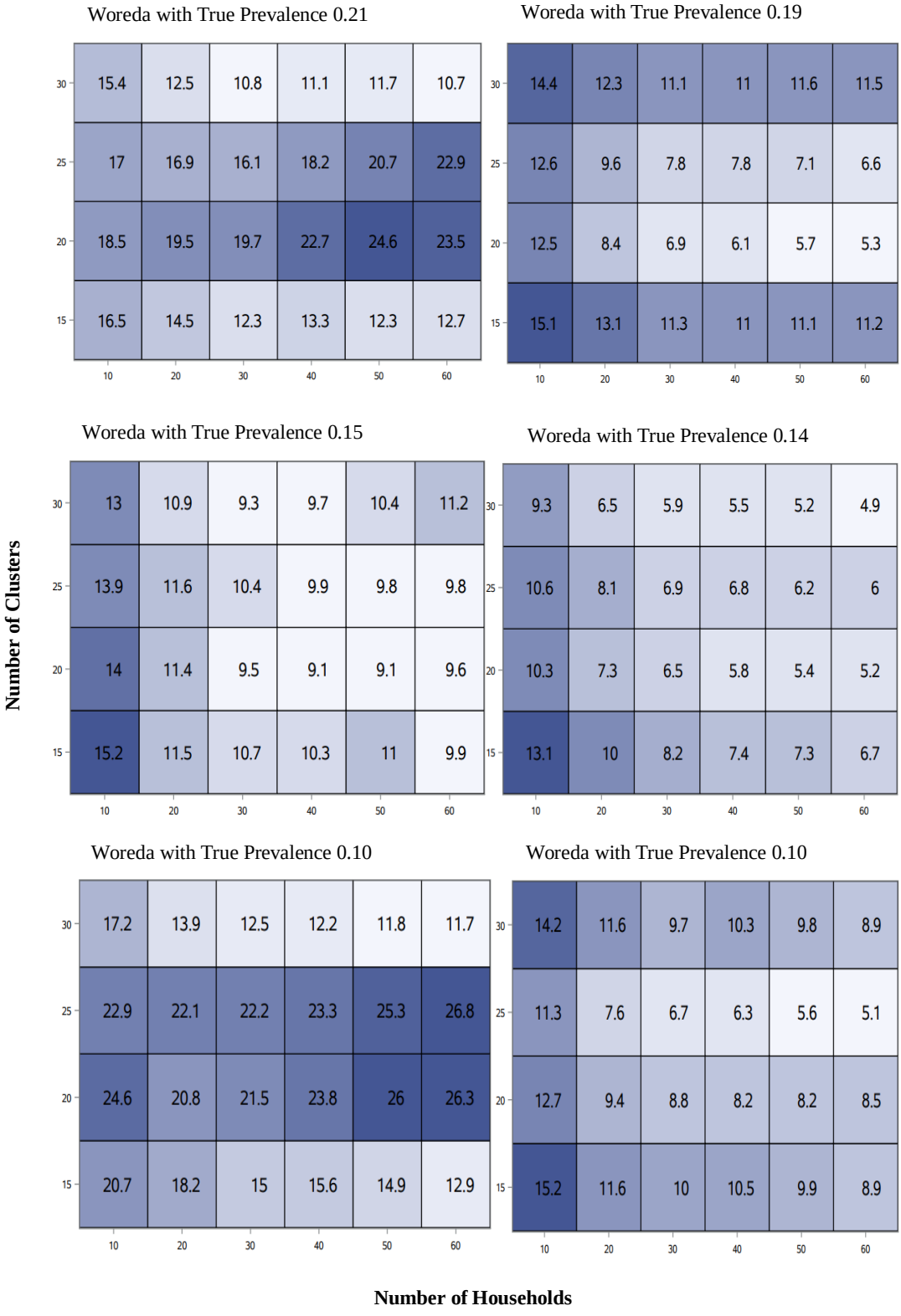
Percent of Sample Size Efficiently Used



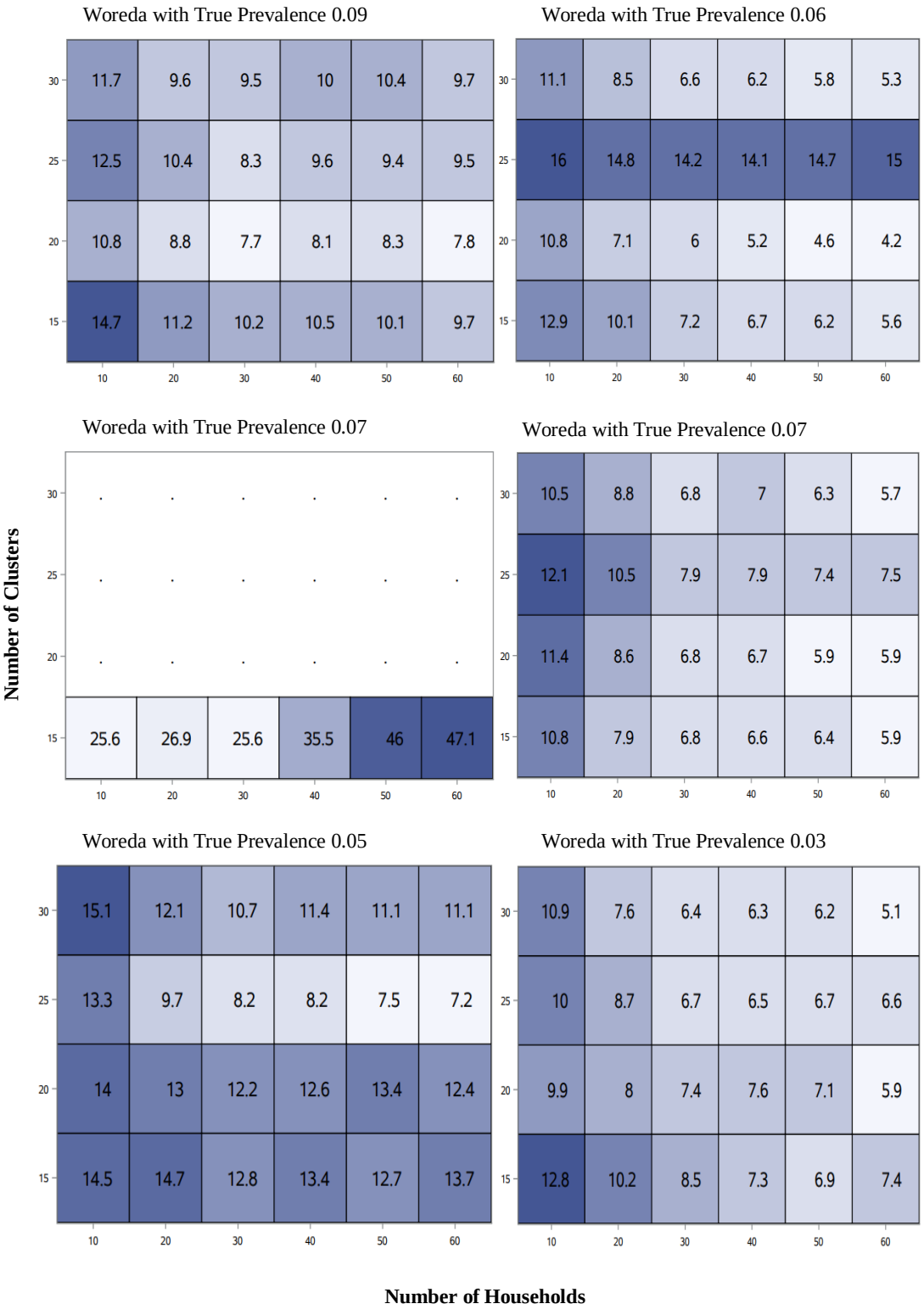
Percent of Sample Size Efficiently Used



Percent of Sample Size Efficiently Used

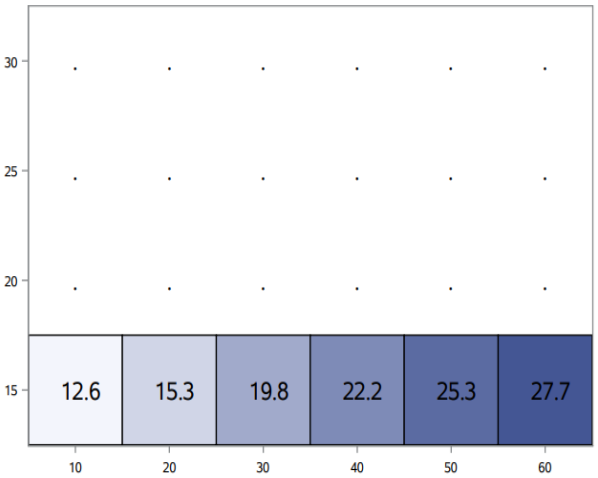


Percent of Sample Size Efficiently Used

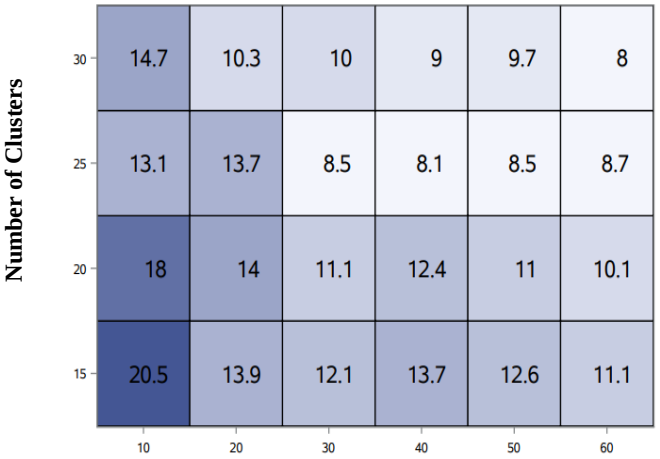


Percent of Sample Size Efficiently Used

Woreda with True Prevalence 0.02

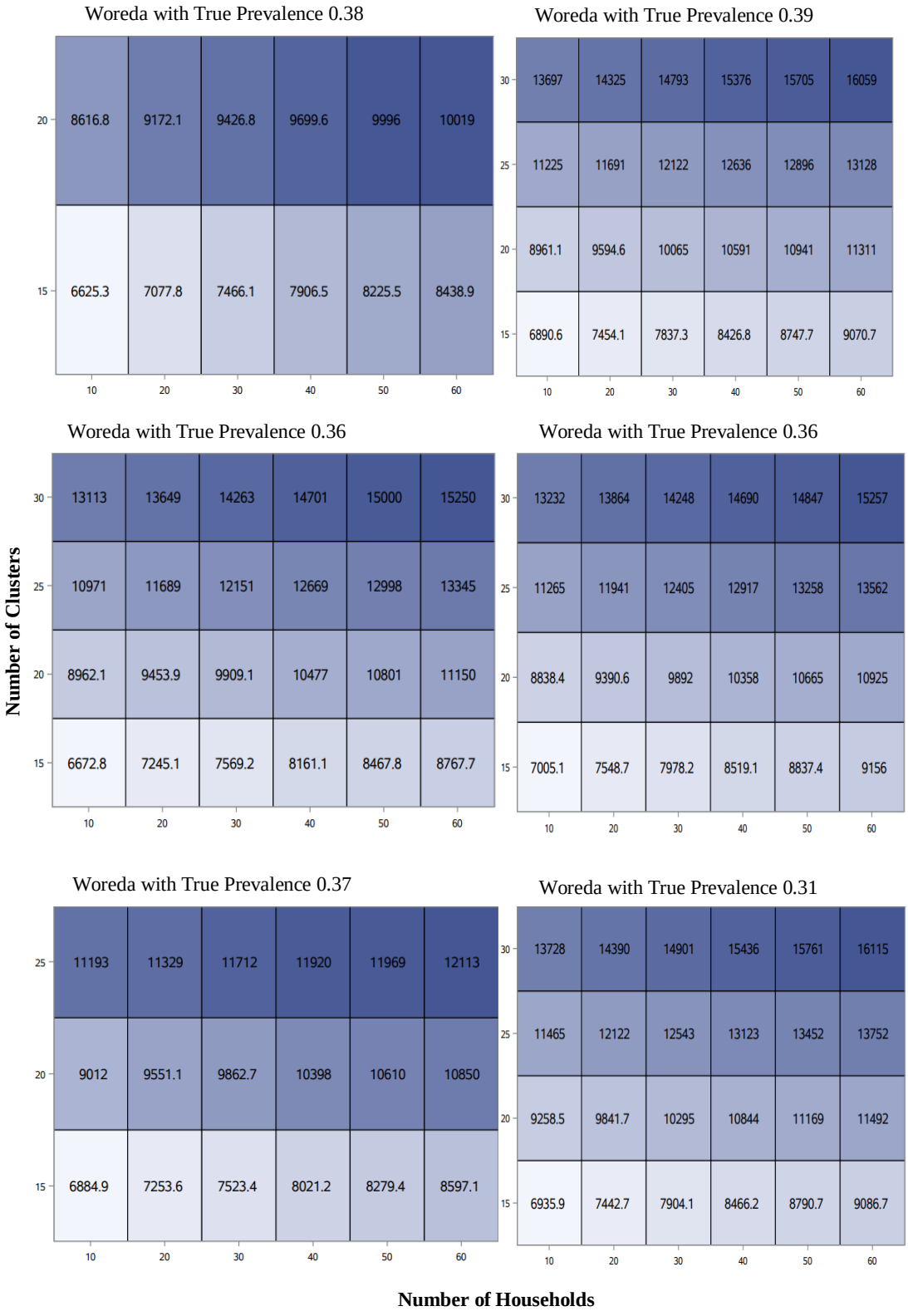


Woreda with True Prevalence 0.01

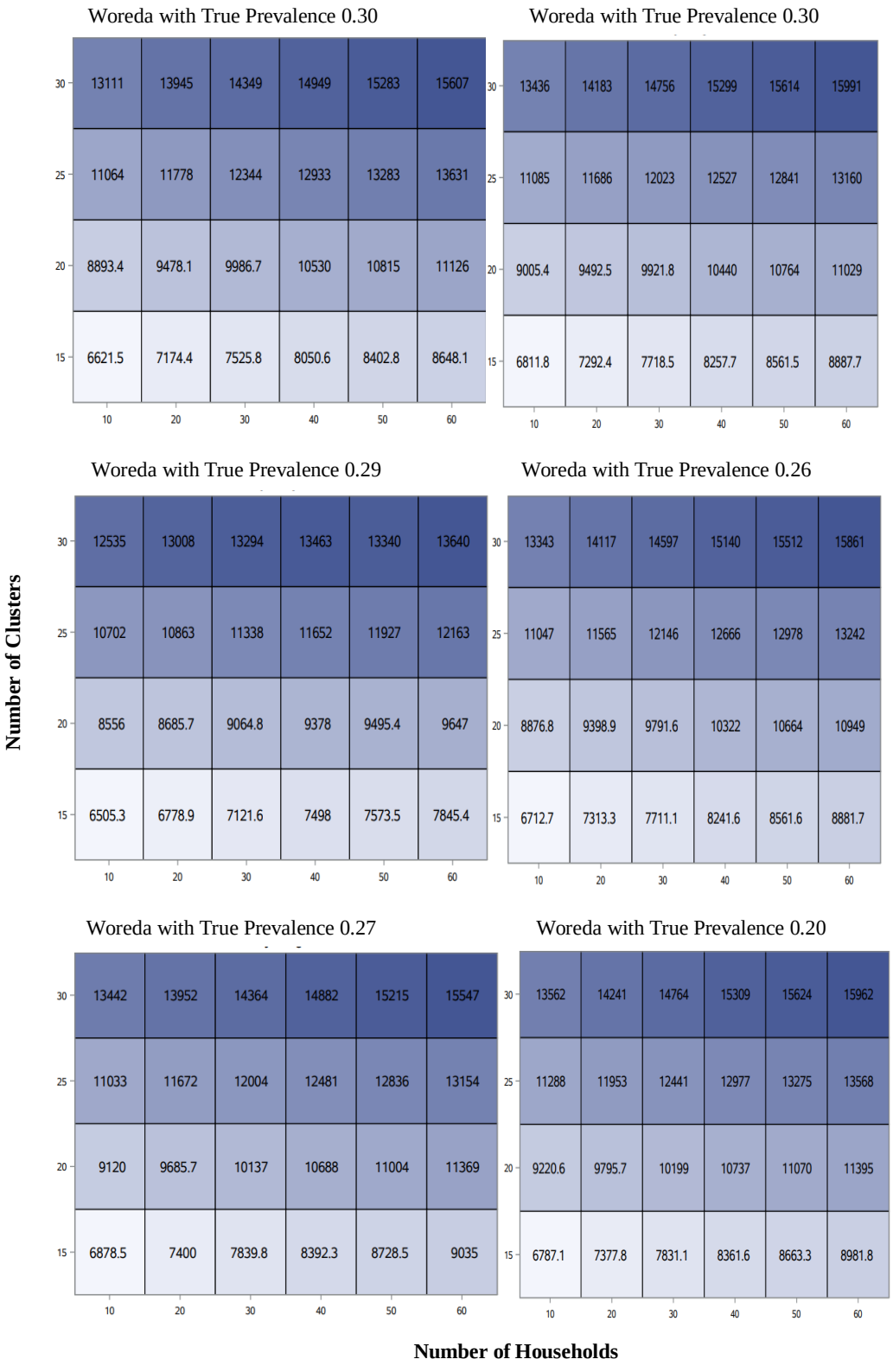


Woreda with True Prevalence 0.002

Cost Wasted

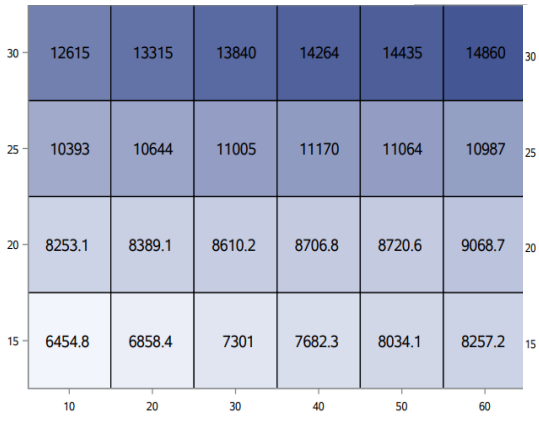

Woreda with True Prevalence 0.38
20 8616.8 9172.1 9426.8 9699.6 9996 10019
15 6625.3 7077.8 7466.1 7906.5 8225.5 8438.9
10 20 30 40 50 60
Woreda with True Prevalence 0.39
30 13697 14325 14793 15376 15705 16059
25 11225 11691 12122 12636 12896 13128
20 8961.1 9594.6 10065 10591 10941 11311
15 6890.6 7454.1 7837.3 8426.8 8747.7 9070.7
10 20 30 40 50 60
Woreda with True Prevalence 0.36
30 13113 13649 14263 14701 15000 15250
25 10971 11689 12151 12669 12998 13345
20 8962.1 9453.9 9909.1 10477 10801 11150
15 6672.8 7245.1 7569.2 8161.1 8467.8 8767.7
10 20 30 40 50 60
Woreda with True Prevalence 0.36
30 13232 13864 14248 14690 14847 15257
25 11265 11941 12405 12917 13258 13562
20 8838.4 9390.6 9892 10358 10665 10925
15 7005.1 7548.7 7978.2 8519.1 8837.4 9156
10 20 30 40 50 60
Number of Clusters
Woreda with True Prevalence 0.37
25 11193 11329 11712 11920 11969 12113
20 9012 9551.1 9862.7 10398 10610 10850
15 6884.9 7253.6 7523.4 8021.2 8279.4 8597.1
10 20 30 40 50 60
Woreda with True Prevalence 0.31
30 13728 14390 14901 15436 15761 16115
25 11465 12122 12543 13123 13452 13752
20 9258.5 9841.7 10295 10844 11169 11492
15 6935.9 7442.7 7904.1 8466.2 8790.7 9086.7
10 20 30 40 50 60
Number of Households

Cost Wasted

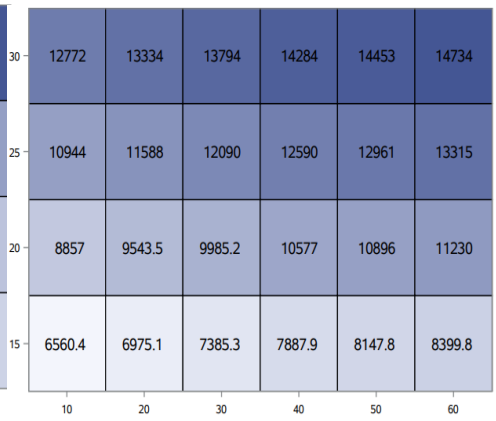

Woreda with True Prevalence 0.30
13111 13945 14349 14949 15283 15607
11064 11778 12344 12933 13283 13631
8893.4 9478.1 9986.7 10530 10815 11126
6621.5 7174.4 7525.8 8050.6 8402.8 8648.1
Woreda with True Prevalence 0.30
13436 14183 14756 15299 15614 15991
11085 11686 12023 12527 12841 13160
9005.4 9492.5 9921.8 10440 10764 11029
6811.8 7292.4 7718.5 8257.7 8561.5 8887.7
Woreda with True Prevalence 0.29
12535 13008 13294 13463 13340 13640
10702 10863 11338 11652 11927 12163
8556 8685.7 9064.8 9378 9495.4 9647
6505.3 6778.9 7121.6 7498 7573.5 7845.4
Woreda with True Prevalence 0.26
13343 14117 14597 15140 15512 15861
11047 11565 12146 12666 12978 13242
8876.8 9398.9 9791.6 10322 10664 10949
6712.7 7313.3 7711.1 8241.6 8561.6 8881.7
Woreda with True Prevalence 0.27
13442 13952 14364 14882 15215 15547
11033 11672 12004 12481 12836 13154
9120 9685.7 10137 10688 11004 11369
6878.5 7400 7839.8 8392.3 8728.5 9035
Woreda with True Prevalence 0.20
13562 14241 14764 15309 15624 15962
11288 11953 12441 12977 13275 13568
9220.6 9795.7 10199 10737 11070 11395
6787.1 7377.8 7831.1 8361.6 8663.3 8981.8
30
25
20
15
10 20 30 40 50 60
Number of Clusters
Number of Households

Cost Wasted

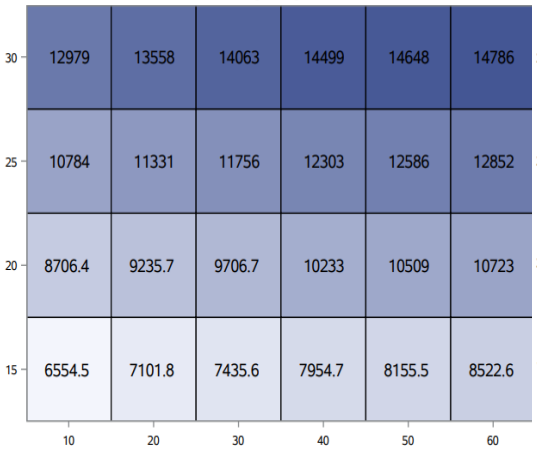
Woreda with True Prevalence 0.21



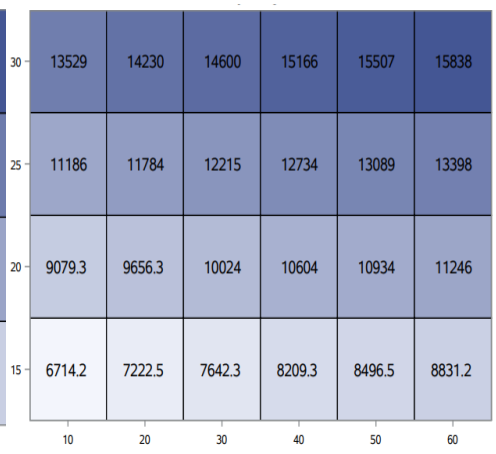
Woreda with True Prevalence 0.19



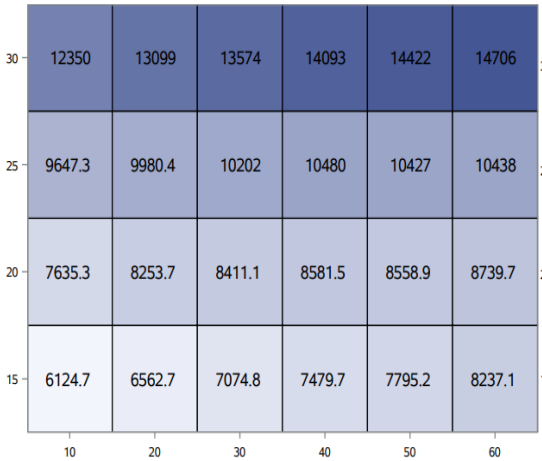
Woreda with True Prevalence 0.15



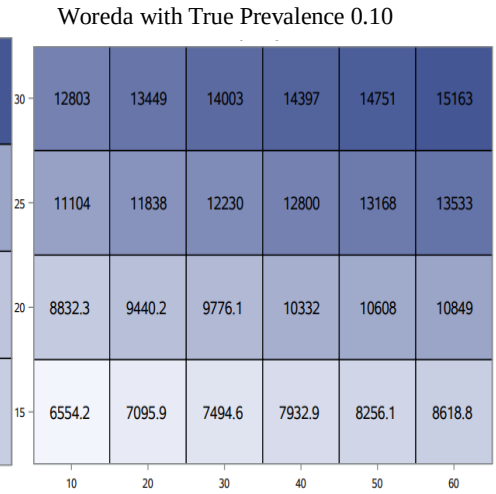
Woreda with True Prevalence 0.14



Woreda with True Prevalence 0.10



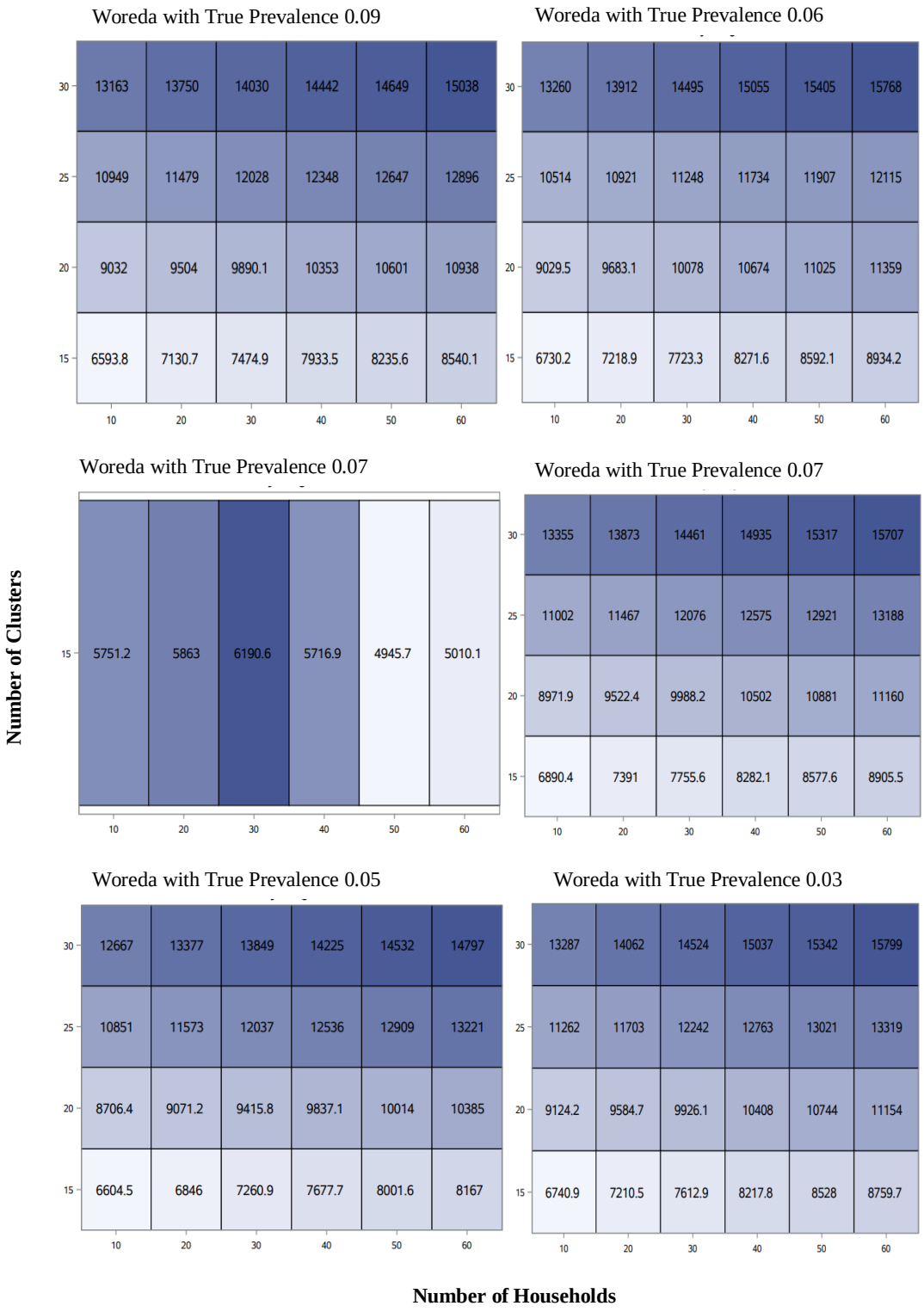
Woreda with True Prevalence 0.10



Number of Clusters

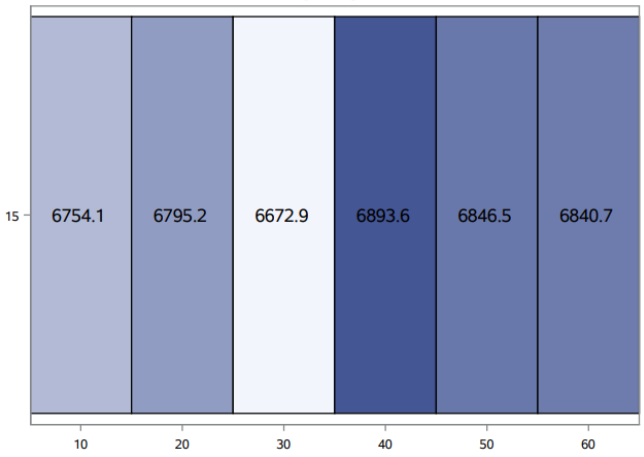
Number of Households

Cost Wasted

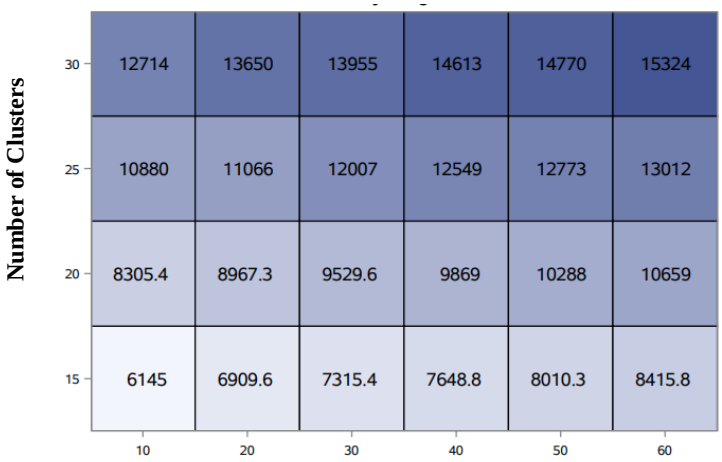

Woreda with True Prevalence 0.09
30 13163 13750 14030 14442 14649 15038
25 10949 11479 12028 12348 12647 12896
20 9032 9504 9890.1 10353 10601 10938
15 6593.8 7130.7 7474.9 7933.5 8235.6 8540.1
10 20 30 40 50 60
Woreda with True Prevalence 0.06
30 13260 13912 14495 15055 15405 15768
25 10514 10921 11248 11734 11907 12115
20 9029.5 9683.1 10078 10674 11025 11359
15 6730.2 7218.9 7723.3 8271.6 8592.1 8934.2
10 20 30 40 50 60
Woreda with True Prevalence 0.07
15 5751.2 5863 6190.6 5716.9 4945.7 5010.1
10 20 30 40 50 60
Woreda with True Prevalence 0.07
30 13355 13873 14461 14935 15317 15707
25 11002 11467 12076 12575 12921 13188
20 8971.9 9522.4 9988.2 10502 10881 11160
15 6890.4 7391 7755.6 8282.1 8577.6 8905.5
10 20 30 40 50 60
Woreda with True Prevalence 0.05
30 12667 13377 13849 14225 14532 14797
25 10851 11573 12037 12536 12909 13221
20 8706.4 9071.2 9415.8 9837.1 10014 10385
15 6604.5 6846 7260.9 7677.7 8001.6 8167
10 20 30 40 50 60
Woreda with True Prevalence 0.03
30 13287 14062 14524 15037 15342 15799
25 11262 11703 12242 12763 13021 13319
20 9124.2 9584.7 9926.1 10408 10744 11154
15 6740.9 7210.5 7612.9 8217.8 8528 8759.7
10 20 30 40 50 60
Number of Clusters
Number of Households

Cost Wasted

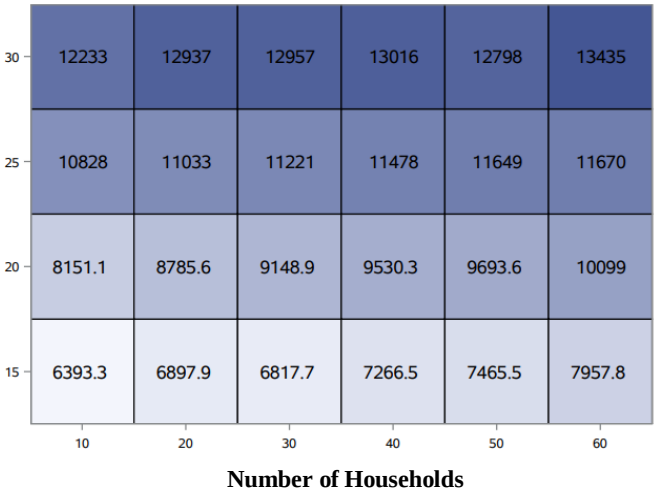
Woreda with True Prevalence 0.02



Woreda with True Prevalence 0.01



Woreda with True Prevalence 0.002



SAS Macros

Population Simulation Macro

```
/*The following macro simulates a TF dataset at the individual level for Amhara,
Ethiopia

User inputs: seed -> randomizes data, can be any integer (default=983758932)
  numgott -> randomized number of gotts (villages) per woreda, can be any positive
  integer (default=100)
  high1-3,mid1-4,low1-7 -> average prevalence levels in each gott (randomized), 14
  levels in total, can be any number in range (0,1)
  (defaults=0.4,0.35,0.3,0.25,0.2,0.15,0.1,0.08,0.05,0.04,0.03,0.02,0.01,0.005)

  A segment prevalence is randomized for each segment based on the gott prevalence.
  There are exactly 30 households per segment, with an average of 1.1 children aged 1-9
per household (# of children randomized).*/


libname prac "H:\Thesis_Practicum";

%macro
tfdata(seed=983758932,/*numgott=100*/high1=0.4,high2=0.35,high3=0.3,mid1=0.25,m
id2=0.2,mid3=0.15,mid4=0.10,low1=0.08,
low2=0.05,low3=0.04,low4=0.03,low5=0.02,low6=0.01,low7=0.005);

data prac.population;

call streaminit(&seed); /*setting seed so data is reproduceable*/

do woreda=1 to 30; /*30 woredas*/

do gott=1 to rand("negbinomial",0.01,1.51);  /* distributed around average number
(156) of specified gotts per woreda*/
unique_gott=(woreda*1000)+gott; /*creating unique id for each gott*/

/*high prevalence districts*/
if 1<=woreda<=3 then do;
tf_gott=rand("beta",20*&high1,20*(1-&high1)); end; /*default mean of 0.4*/
else if 4<=woreda=<6 then do;
tf_gott=rand("beta",20*&high2,20*(1-&high2)); end; /*default mean of 0.35*/
```

```
else if 7<=woreda<=9 then do;
tf_gott=rand("beta",20*&high3,20*(1-&high3)); end; /*default mean of 0.3*/

/*middle prevalence districts*/
else if 10<=woreda<=12 then do;
tf_gott=rand("beta",25*&mid1,25*(1-&mid1)); end; /*default mean of 0.25*/
else if 13=<woreda=<15 then do;
tf_gott=rand("beta",25*&mid2,25*(1-&mid2)); end; /*default mean of 0.2*/
else if 16<=woreda<=18 then do;
tf_gott=rand("beta",25*&mid3,25*(1-&mid3)); end; /*default mean of 0.15*/
else if 19<=woreda<=21 then do;
tf_gott=rand("beta",25*&mid4,25*(1-&mid4)); end; /*default mean of 0.10*/

/*low prevalence districts*/
else if 22<=woreda<=24 then do;
tf_gott=rand("beta",25*&low1,25*(1-&low1)); end; /*default mean of 0.08*/
else if woreda=25 then do;
tf_gott=rand("beta",25*&low2,25*(1-&low2)); end; /*default mean of 0.05*/
else if woreda=26 then do;
tf_gott=rand("beta",25*&low3,25*(1-&low3)); end; /*default mean of 0.04*/
else if woreda=27 then do;
tf_gott=rand("beta",25*&low4,25*(1-&low4)); end; /*default mean of 0.03*/
else if woreda=28 then do;
tf_gott=rand("beta",25*&low5,25*(1-&low5)); end; /*default mean of 0.02*/
else if woreda=29 then do;
tf_gott=rand("beta",25*&low6,25*(1-&low6)); end; /*default mean of 0.01*/
else if woreda=30 then do;
tf_gott=rand("beta",25*&low7,25*(1-&low7)); end; /*default mean of 0.005*/

 do segment=1 to (rand("negbinomial",0.1,0.25)+1); /*average number of segments per
gott is 3.24, this gives expected value of about 3.24 with signficant variance*/
 unique_seg=(unique_gott*10)+segment; /*creating unique value for each segment*/

 tf_segment=rand("beta?",(200*tf_gott),(200*(1-tf_gott))); /*creating segment level
prevalence to account for correlation within segments, want small variance so use 200
as multiplier*/

 randnum=rand("uniform");

  do household=1 to 30; /*30 houses per segment*/
  unique_house=(unique_seg*100)+household; /*creating unique value for each
household*/
```

```
  do member_1to9=0 to rand("Poisson",1.107); /*Poisson distributed number of kids 1-
9 per household, mean calculated as 1.107=(.27*4.1)*/
  unique_mem=(unique_house*10)+member_1to9;

  if member_1to9>0 then do;
   tf_ind=rand("Bernoulli",tf_segment); /*Indicator variable for TF based on prevalence
of TF in the gott*/
  end;
  else do;
   tf_ind=.;
  end;

  output;
  end; end; end; end; end;

run;

%mend tfdata;
```

Drawing Samples of 30 Households (Using Segment Structure) Macro

```
libname prac "H:\Thesis_Practicum";

/*creating variable for number of gotts per woreda*/
proc sql;
create table t1
as select *, max(gott) as numgott
from prac.population
group by woreda;

/*creating variable for number of segment in each gott*/
proc sql;
create table t2
as select *, max(segment) as numseg
from t1
group by unique_gott;

/*tabulating woreda by gott by segment*/
proc freq data=t2 noprint;
  tables woreda*gott*segment / out=x;
run;
```

```
proc sort data=x;
  by woreda gott segment;
run;

/*creating data set with one observation for each gott, with variable that is total
number of segments per gott*/
data frame;
  set x;
  by woreda gott segment;
  drop count percent;
  if last.gott;
run;


/*The following macro draws 1000 samples from the population data set created by the
previous macro. It also calculates 95% uncertainty intervals. It uses the segment
structure, so only one segment or 30 households can be selected using this macro. The
user inputs the cluster number desired.*/


%macro sample(cluster=14);

/*selecting user input number of clusters (gotts) within each of the 17 woredas*/
proc surveyselect data=frame method=sys N=&cluster out=sample_&cluster
seed=6212018 rep=1000;
samplingunit gott;
strata woreda;
run;

proc sort data=sample_&cluster;
  by woreda gott segment;
run;

/*expanding data set to the segment level so that we can randomly select a segment*/
data sampseg_&cluster;
  set sample_&cluster;
  do segment=1 to segment; output;
  end;
run;

/*creating random number for each segment*/
data sseg_&cluster;
set sampseg_&cluster;
rand=ranuni(123456);
```

```
run;

/*sorting by random number within replicate, woreda, and gott*/
proc sort data=sseg_&cluster;
  by replicate woreda gott rand;
run;

/*selecting only the first segment within each replicate, woreda, and gott to get 1
segment per cluster*/
data full_&cluster;
  set sseg_&cluster;
  by replicate woreda gott;
  if first.gott;
  do household=1 to 30; /*expanding to the household level for future merging*/
    output;
  end;
  drop samplingweight;
run;

proc sort data=full_&cluster;
  by woreda gott segment household;
run;

proc sort data=t2;
  by woreda gott segment household;
run;

/*merging population clinical data to sample data from above*/
proc sql;
create table sql_sample_&cluster as
select f.replicate, f.woreda, p.woreda, f.gott, p.gott, f.segment, p.segment, f.household,
p.household,
p.member_1to9, p.tf_ind, f.selectionprob, p.numseg, p.numgott
from work.full_&cluster f
left join work.t2 p
on f.woreda=p.woreda AND f.gott=p.gott AND f.segment=p.segment AND
f.household=p.household
;
quit;

proc sort data=sql_sample_&cluster;
by replicate woreda gott segment household member_1to9;
run;
```

```
/*calculating true weights by multiplying selection prob from both stages*/
data prac.sample_&cluster;
set sql_sample_&cluster;
finalprob=(1/numseg)*selectionprob;
final_weight=1/finalprob;
run;

/*getting bootstrap confidence intervals for each woreda*/

/*calculating prevalence for each replicate and each woreda weighted based on
sampling weights*/
proc means data=prac.sample_&cluster noprint;
weight final_weight;
class replicate woreda;
var tf_ind;
output out=means_&cluster;
run;

proc sort data=prac.sample_&cluster;
by gott segment;
run;

proc descript data=prac.sample_&cluster filetype=SAS design=uneqwor conf_lim=95;
/*won't calculate variance with only 1 segment available - need to change sample
design?*/
nest gott segment;
totcnt numgott numseg;
weight final_weight;
jointprob _one_;
var tf_ind;
subgroup woreda;
levels     30;
print mean semean lowmean upmean;
run;

/*manipulating data set of prevalences*/
data mn_&cluster;
set means_&cluster;
if _STAT_ ne "MEAN" then delete;
drop _TYPE_ _FREQ_ _STAT_;
rename tf_ind=Mean_prev;
if woreda=. then delete;
if replicate=. then delete;
run;
```

```
/*calculating percentiles within replicates for 95% confidence intervals*/
proc univariate data=mn_&cluster noprint;
class woreda;
var mean_prev;
output out=percentile_&cluster mean=avg_est_mean pctlpre=replicate
pctlpts=2.5,97.5;
run;

/*creating dataset with lower and upper bounds, and length on confidence interval*/
data prac.ci_one_&cluster;
set percentile_&cluster;
rename replicate2_5=lower replicate97_5=upper;
label replicate2_5=" " replicate97_5=" ";
clusters=&cluster;
length=replicate97_5-replicate2_5;
run;

%mend sample;
```

Drawing Samples of Varying Numbers of Households (Ignoring Segment Structure) Macro

```
proc sql;
create table t12
as select *, max(gott) as numgott
from prac.population
group by woreda;
proc sort data=t12;
by woreda gott segment household;
run;
/*creating variable for number of segment in each gott*/
proc sql;
create table t22
as select *, max(segment) as numseg
from t12
group by unique_gott;

/*creating max houses variable per gott*/
data house;
set t22;
maxh=30*numseg;
```

```
run;

/*tabulating woreda by gott by segment*/
proc freq data=house noprint;
  tables woreda*gott*maxh / out=x;
run;

proc sort data=x;
  by woreda gott maxh;
run;

/*creating data set with one observation for each gott, with variable that is total
number of segments per gott*/
data frame;
  set x;
  by woreda gott maxh;
  drop count percent;
  if last.gott;
run;

/*macro to draw a bunch of samples*/

/*The following macro draws 1000 samples of a user input number of clusters and
households. The segment structure is ignored.*/

%macro sampleh(cluster=30,house=10);

/*selecting user input number of clusters (gotts) within each of the 30 woredas*/
proc surveyselect data=frame method=sys N=&cluster out=sample_&cluster&&house
seed=6212018 rep=1000;
samplingunit gott;
strata woreda;
run;

proc sort data=sample_&cluster&&house;
  by woreda gott maxh;
run;

/*expanding data set to the household level so that we can randomly select a number of
households*/
data samph_&cluster&&house;
  set sample_&cluster&&house;
  do house=1 to maxh; output;
  end;
```

```
  /*adding this line to try to fix issues, may need to remove*/
  rename house=household;
run;

/*sorting by random number within replicate, woreda, and house*/
proc sort data=samph_&cluster&&house;
  by replicate woreda gott;
run;

/*selecting households*/
proc surveyselect data=samph_&cluster&&house method=srsN=&house selectall
out=samp_&cluster&&house seed=6212018;
samplingunit household;
strata replicate woreda gott;
run;

proc sort data=samp_&cluster&&house;
  by woreda gott household;/*this was changed trying to fix the error*/
run;

proc sort data=t2;
  by woreda gott household;
run;

/*merging population clinical data to sample data from above*/
proc sql;
create table sql_sample_&cluster&&house as
select f.replicate, f.woreda, p.woreda, f.gott, p.gott, f.household, p.household,
p.member_1to9, p.tf_ind, f.selectionprob, p.numgott, f.maxh
from work.samp_&cluster&&house f
left join work.t2 p
on f.woreda=p.woreda AND f.gott=p.gott AND f.household=p.household
;
quit;

proc sort data=sql_sample_&cluster&&house;
by replicate woreda gott household member_1to9;
run;

/*calculating true weights by multiplying selection prob from both stages*/
data prac.sample_&cluster&&house;
set sql_sample_&cluster&&house;
sprob1=&house/maxh;
if sprob1>1 then sprob2=1;
```

```
else sprob2=sprob1;
finalprob=sprob2*selectionprob;
final_weight=1/finalprob;
run;

/*getting bootstrap confidence intervals for each woreda*/

/*calculating prevalence for each replicate and each woreda weighted based on
sampling weights*/
proc means data=prac.sample_&cluster&&house noprint;
weight final_weight;
class replicate woreda;
var tf_ind;
output out=means_&cluster&&house;
run;

proc sort data=prac.sample_&cluster&&house;
by gott;
run;

/*manipulating data set of prevalences*/
data mn_&cluster&&house;
set means_&cluster&&house;
if _STAT_ ne "MEAN" then delete;
drop _TYPE_ _FREQ_ _STAT_;
rename tf_ind=Mean_prev;
if woreda=. then delete;
if replicate=. then delete;
run;

/*calculating percentiles within replicates for 95% confidence intervals*/
proc univariate data=mn_&cluster&&house noprint;
class woreda;
var mean_prev;
output out=percentile_&cluster&&house mean=avg_est_mean pctlpre=replicate
pctlpts=2.5,97.5;
run;

/*creating dataset with lower and upper bounds, and length on confidence interval*/
data prac.ci_&cluster&&house;
set percentile_&cluster&&house ;
rename replicate2_5=lower replicate97_5=upper;
label replicate2_5=" " replicate97_5=" ";
clusters=&cluster;
```

```
households=&house;
length=replicate97_5-replicate2_5;
run;

%mend sampleh;
```