Essays on wealth, health, and data collection

Kools, L.

5 One-stage versus two-stage cluster sampling, a simulation study

Abstract

Two-stage cluster sampling is a widely used method to sample households for large-scale face-to-face household surveys. Developments in survey sampling methodology, such as gridded sampling, can make it easier to define smaller Primary Sampling Units and adopt a one-stage cluster sampling approach. This approach may mitigate the risk of excluding mobile populations and reduce operational costs per cluster by combining the listing and interviewing phases. However, one-stage cluster sampling may require larger sample sizes if households of the same type tend to live close to each other. Based on a synthetic population of Oshikoto, Namibia, we analyze the potential increase in the required number of clusters under a one-stage design to achieve the statistical power of a typical two-stage cluster sample. We find that under moderate assumptions sample sizes at most double. However, in some extreme cases the required number

The chapter is co-authored by Dana Thomson. The authors thank Felicity Cutts and Dhale Rhoda for their feedback at several stages of the chapter. Furthermore, the authors thank Chris Jochem for his guidance in setting up the synthetic data set. The authors would also like to thank Jeremiah J. Nieves for assembling and sharing the WorldPop geospatial datasets used in this study. The geospatial datasets were produced by David Kerr, Heather Chamberlain, Chris T. Lloyd, Maksym Bondarenko (WorldPop, University of Southampton), Gregory Yetman, and Linda Pistolesi (Center for International Earth Science Information Network, Columbia University) in the framework of the WorldPop âĂIJGlobal High Resolution Population DenominatorsâĂİ Project funded by the Bill & Melinda Gates Foundation (OPP1134076). Lieke Kools received financial support from the Leiden University Fund/Kroese-Duijsters Fonds for conducting a research visit to Southampton University during which a large part of the work for this chapter has been conducted.

of clusters can increase by up to thirteen times. The potential increase depends on both prevalence of the characteristic and the intracluster correlation at the level of Enumeration Areas. The differences between extreme and moderate scenario's fade out when segment sizes are increased.

5.1 Introduction

Large multi-topic household surveys, such as the Living Standards Measurement Study (LSMS) and Demographic and Health Surveys (DHS) in low and middle income countries (LMICs) and EU-SILC in Europe are an important tool for monitoring socio-economic progress. When deciding how to select the respondents for such surveys, one has to ensure that the resulting sample is representative of the population and large enough to estimate key characteristics at the (sub)national level with sufficient precision. On the other hand, the survey should be affordable and the approach easy to implement in the field. A two-stage cluster sampling design was historically the only available study design and is seen as the gold standard for survey sampling in LMICs, because it offers a good balance between these requirements. However, in cases where clusters contain few household and the target population is rare, recent WHO guidelines also suggest one-stage cluster sampling as a suitable sampling method (World Health Organization 2015). Thanks to novel ways of defining clusters this alternative method now even becomes feasible for general household surveys. Compared to two-stage sampling, one-stage cluster sampling may reduce operational costs per cluster and may mitigate the risk of excluding hard-to-survey populations. However, it is more sensitive to spatial clustering of household characteristics and may therefore come at the cost of larger sample size requirements. The potential usefulness of one-stage cluster sampling in the field depends on the required increase in sample size to maintain the same statistical precision as a two-stage cluster sample. Therefore, an effort to quantify this increase is needed.

Two-stage cluster sampling is a common sampling approach for face-to-face household surveys. The first stage consists of defining primary sampling units (PSUs), i.e. mutually exclusive subsets of the population,

and then selecting a subset of these PSUs.[1] PSUs can for example be defined based on administrative boundaries, such as the Enumeration Areas (EAs) used in the most recent Census, or by overlaying a map by a raster (gridded sampling, Galway et al. (2012), Thomson et al. (2017)). In the second stage a subset of households within each PSU is selected, who together form the sample of the survey. Often this is done by visiting the selected PSU, listing all the households in that region, taking a systematic or random sample from this list, and then revisiting these households for an interview. Two-stage cluster sampling has practical benefits compared to taking a Simple Random Sample (SRS) of the population: building the sampling frame does not require complete population registries and field work can be concentrated in a few areas. However, the sampling approach may result in non-random selection of household types due to various reasons. Among these is the substantial time lag between the listing phase and interviewing phase, which has as a consequence that mobile populations, such as seasonal workers, are at risk of being excluded from the sample.

One could forgo on revisits by defining PSUs in such a way that each contains only few households and interview all the households in the selected PSU, i.e. one-stage cluster sampling. For example, a grid cell of $100m^2$ is often smaller than a EA, such that gridded sampling offers opportunities for one-stage cluster sampling. Also, when taking a gridded sampling approach, PSUs of different sizes can be established by combining neighboring cells or segmenting cells. A one-stage cluster approach could potentially lead to cost savings because the area to cover in the sampled clusters is much smaller, so that listing and interviewing can be executed on the same day. Moreover, the one-stage setup does allow to capture mobile and non-standard households, as shown by Himelein et al. (2014). However, if similar types of households tend to live close to each other, each one-stage cluster adds less new information to the sample than each two-stage cluster of the same sample size, so that one should sample

[1] There also exist two-stage methods in which the PSUs are not a mutually exclusive subset of the population. For example when points on a map are sampled and the PSU is defined as the region in a specified radius around that point. In such cases survey weights should be adapted for the possibility that two selected PSUs overlap.

more clusters for a one-stage cluster sample to achieve the same precision as a two-stage cluster sample. How many more depends on the spatial clustering of characteristics and thus on the variable of interest and the context these are measured in.

There is little evidence on the difference in precision of one- and two-stage cluster samples and how different forms of spatial clustering affect these differences. However, literature on other sampling procedures may give guidance for the direction of the results, in particular the literature on two-stage EPI sampling. For this approach, the second stage consists of selecting a random starting point in the PSU and from there taking a 'random walk' through the PSU on which households are selected until the required number of households are interviewed. The most prominent critique on this sampling method is that is does not lead to a true probability sample[2], however another critique is that households living close together are more likely to be sampled than households living further apart. Therefore, a two-stage EPI sample is affected by spatial clustering in a similar way as a one-stage cluster sample. Indeed, Milligan et al. (2004) show that implementation of two-stage EPI sampling and one-stage cluster sampling[3] lead to equivalent point estimates of vaccination coverage in the Western region of Gambia. The EPI approach has been shown to be sensitive to pocketing of vaccinated individuals (Lemeshow et al. 1985) and to perform poorly for socio-economic variables (Bennett et al. 1994). However, the approach does usually lead to estimates within

[2]When selecting households using a 'random walk', the households in the EA are not listed. As most two-stage sampling methods, the EPI approach depends on the last census for its sampling frame. These sampling frames are usually outdated, so that without listing all the households currently living in the EA one cannot establish second stage sampling probabilities. Therefore, despite the ease of implementation, the current survey guidelines of the WHO recommend against the use of this approach and in favor of systematic two-stage cluster sampling or one-stage cluster sampling (World Health Organization 2015). Contrary to the EPI approach, one-stage cluster sampling does generally lead to a true probability sample.

[3]Actually Milligan et al. (2004) compare the EPI approach to an approach they call two-stage compact segment sampling. This approach was introduced by Turner et al. (1996) and consists of first sampling PSUs, dividing these PSUs in x equal segments, and subsequently interview all households in one segment per PSU. Practically this gives a similar sample to a one-stage cluster sample, with the only exception that in a one-stage cluster sample two segments belonging to one 2-stage PSU could in principle both be sampled.

10% of the population mean (the EPI criterion for a good sample) and has shown to provide similar estimates for mortality and vaccination status as more systematic sampling procedures (Luman et al. 2007, Rose et al. 2006). To our knowledge, specific guidelines with respect to the difference in the number of clusters to sample are not given in the literature.

In this chapter we aim to find out how many additional clusters need to be sampled when using a one-stage cluster design to achieve the statistical power of a typical two-stage cluster sample in LMIC household surveys. In order to answer this question we create a synthetic population of households in Oshikoto, Namibia. We argue that this population has the same properties as the true population in the sense that both population averages and the distribution of EA-level prevalences are equivalent for key characteristics.[4] Next we adopt several scenarios for the spatial distribution of individuals within each EA, while keeping EA-level prevalences fixed. For each of these different scenarios we calculate the minimal number of clusters to be sampled to achieve a given statistical precision based on bootstrapped measures of performance. We focus on three measures (1) household wealth index, (2) womens use of modern contraception, and (3) 0-5 years old children's DPT3 vaccination coverage. These measures show different distributions of EA-level prevalences and each cover a different subsample of the population.

The results show that under moderate assumptions sample size requirements at most double. However, under extreme assumption of within EA clustering sample size requirements can increase dramatically, especially for variables with low EA-level ICCs and prevalence levels near 50%. The most extreme case showed an increase in the minimal number of clusters to sample of almost thirteen times. The differences between extreme and moderate scenario's fade out when segment sizes are increased. Before implementing one-stage cluster sampling one should carefully examine the likely scenarios of within EA clustering, to assess feasibility of the approach.

[4]That is, we expect the intra cluster correlations found in the synthetic population is close to those in the true population.

5.2 Method

We evaluate the sampling procedure using a synthetic population of the region Oshikoto in Namibia. This region was selected because of both the availability of high quality data and the diversity of the region. The region covers $38,653 km^2$ including planned and unplanned city neighborhoods, rural settled agriculture, rural nomadic populations and large unpopulated areas. In this section we briefly explain how the synthetic population is generated and argue why this provides a valid testing ground for the question at hand. Next, we explain how we constructed the different scenarios of within EA clustering and how we calculate the minimal number of clusters to be sampled when using either a one- or two-stage cluster sampling approach. For a more detailed description of the data and methods used to construct the synthetic population, we refer the reader to Thomson et al. (2018).

5.2.1 Generating a realistic synthetic population

The synthetic population is constructed using the 2013 Namibian Demographic Household Survey (DHS), the 2011 Namibian census Public Use Microdata Sample (PUMS), a set of publicly available spatial covariates, and a household point location file constructed by visual inspection of satellite imagery of the region. Table 5.1 provides an overview of the datasets and variables used to generate the synthetic population. The population is created in three steps: first we predict the spatial distribution of household types, which we then use to assign a realistic set of synthetic households to realistic household locations, and finally we predict some extra characteristics of the individuals. The steps constitute of a series of random processes, which will be further explained below, so that executing them once results in one of many possible realizations of the synthetic population. For the analysis we generate five realizations of the synthetic population, run the analysis on each of these realizations, and base our conclusions on the combined results of the different analyses.

Table 5.1: Overview of data sources used for simulations

Dataset	Information retrieved	Original source (unit)
Demographic and Health Survey 2013[a](MoHSS and ICF 2014)	geo-displaced cluster coordinates, urban/rural (hv025), cluster (v001), hhsize (derived), water source (hv201), toilet facility (hv205, hv225), space (hv216), structure (hv213), cooking fuel (hv226), relationship (hv101), age (hv105), sex (hv104), education (hv109), wealth index (hv270) , contraception (v313), DPT3 vaccination[b](h7)	
Census 2011 PUMS (NSA 2013)	admin-3 level indicator (constituency), urban/rural (urban_rural), hhsize (derived), water source (H9), toilet facility (H10), space (H4), structure (H7), cooking fuel (H8a), relationship (B3), age (B5), sex (B4), education (D3)	
2011 Census EA boundaries (NSA 2011a)	EA boundaries	
2011 Census main report (NSA 2011b)	Constituency populations totals	
2014-2016 DigitalGlobe Quickbird imagery, 50cm (DigitalGlobe 2014)	(estimated) household point locations	
ccilc_dst011_2012	Distance to cultivated terrestrial lands[c]	2012 ESA CCI annual LC maps v2.0.7 (≈300m)[d]
ccilc_dst040_2012	Distance to woody areas[c]	""
ccilc_dst130_2012	Distance to shrub areas[c]	""
ccilc_dst140_2012	Distance to herbaceous areas[c]	""
ccilc_dst150_2012	Distance to terrestrial vegetation areas[c]	""
ccilc_dst190_2012	Distance to urban area[c]	""
ccilc_dst200_2012	Distance to bare areas[c]	""
cciwat_dst	Distance to water bodies[c]	ESA CCI, Water bodies v4.0 (≈150m)[d]
dmsp_2011	Nighttime lights intensity[c]	2011 inter-calibrated version of the v4 DMSP-OLS Nighttime Lights Time Series (≈1km)[d]
gpw4coast_dst	Distance to coastline[c]	GPWv4 input administrati units (≈100m)[d]
osmint_dst	Distance to road intersections[c]	2016 OSM highways[d]
osmriv_dst	Distance to major waterways[c]	2016 OSM waterways[d]
osmroa_dst	Distance to major roads[c]	2016 OSM highways[d]
slope	Slope[c]	2000 Viewfinder Panoramas (≈100m)[d]
topo	Elevation[c]	2000 Viewfinder Panoramas (≈100m)[d]
tt50k2000	Travel time to populated places (pop more 50k)	2000 EC-JRC Travel time to major cities (≈1km)[d]
urbpx_prp_1_2012	Proportion of settlement pixels within 1 cell radius[c]	2012 DLR Global Urban Footpring (≈12.5m) & 2000 EC-JRC Global Human Settlement layer; 38m[d]
2010 MODIS (≈1km) (Running et al. 2014)	Annual net primary productivity[c]	
2001 education facilities (UN-OCHA ROSA 2001b)	Distance to schools[c]	
2001 health facilities (UN-OCHA ROSA 2001a)	Distance to health facilities[c]	

[a] The household variables are taken from the household recode file, the women and child variables come from the individual recode file.

[b] We only measure DPT3 vaccination coverage for children living with their mother. In Oshikoto 29 children under 5 (9.9%) are reported to live away from their mother in the DHS. The vaccination coverage of children living away from their mother is slightly lower than that of children living with there parents, though not statistically significantly lower (72.41% vs 78.41%, p-value of one-side t-test: 0.2312).

[c] Unit of measurement: 3 arc seconds (≈ 100m).

[d] Spatial covariate was processed by the "Global High Resolution Population Denominators" Project.

Step 1: Predict the spatial distribution of household types.

In order to generate a synthetic population which is realistically distributed over space, we need to understand which types of households are likely to live in which areas of our region. We can do this by estimating relationships between spatial covariates and household types. For this we rely on the 2013 DHS survey, which not only provides information on households but also GPS coordinates of the surveyed clusters. To establish this relationship we need to find out what the typical household looks like in each surveyed cluster. We start by summarizing the individual characteristics to the household-level, so that we have a household file with the following dummy variables: 1[is rural], 1[head has any formal education], 1[has any children under 5 years old], 1[does not have access to an improved water source][5], 1[does not have access to improved toilet facility][6], 1[lives in a non-durable structure][7], 1[lives in house with inadequate space][8], 1[cooks on solid fuel][9]. The choice for these characteristics is based on the availability of information in both the DHS and the census PUMS, which we use in step 2 to generate our synthetic population. Next, we take the cluster average of these household characteristics, resulting in one typical household per cluster. Finally, we construct a single variable

[5]i.e. the water source is labeled as well unprotected, river/dam/stream in cases of census data and the water source is labeled as unprotected well, river/dam/lake/ponds/stream/canal/irrigation, unprotected spring, tanker truck, cart with small tank (hv205), or shared (hv225) in case of DHS data (UN-HSP 2003). For both the DHS as the census the category other is set to missing.

[6]i.e. the toilet facility is labeled as uncovered pit latrine without ventilation, bucket toilet, or no facility in case of census data and the toilet facility is labeled as pit latrine without slab/open pit, flush to somewhere else, bucket toilet, hanging toilet/latrine, or no facility/bush/field in case of DHS data (UN-HSP 2003). For both the DHS as the census the category other is set to missing.

[7]i.e. the floor material is labeled as sand/earth, cement, mud/clay or wood in case of census data and the floor material is labeled as earth/sand, dung, mud/clay, wood planks, palm/bamboo in case of DHS data Fink et al. (2014). For both the DHS as the census the category other is set to missing.

[8]i.e. on average more than 3 individuals share one sleeping room. For both the census as the DHS data this was derived from household size and a variable measuring the number of sleeping rooms in the house (UN-HSP 2003).

[9]i.e. cooking fuel is labeled as wood/charcoal from wood, charcoal-coal, or animal dung in case of census data and cooking fuel is labeled as charcoal, wood, agricultural crop or animal dung in case of DHS data. For both the DHS as the census the category other is set to missing.

Table 5.2: Household types

Type	Name	Description
1	rural rich	educated and high access to facilities
2	rural poor 1	No formal education and low access to facilities, except water
3	urban rich	educated, few under fives, and high access to facilities
4	urban average	No formal education and average access to facilities
5	rural average 1	few under fives, average to high access to facilities
6	rural poor 2	many under fives and low access to facilities
7	rural average 2	average access to facilities, low access to fuel.

High:= above regional average, Low:= below regional average, Regular:= close to regional average. Summary statistics of each type are given in Thomson et al. (2018).

that summarizes the information of the different characteristics by means of k-means clustering. K-means clustering is a form of unsupervised clustering aiming to partition observations into a pre-defined number (k) of groups or clusters. The clusters are formed such that the sum of squares between the points and the cluster centroids (middle points) is minimized (Hartigan and Wong 1979). We denote the resulting variable as the 'household type'. The algorithm results in the 7 types depicted in table 5.2.

In order to describe the relationship between these household types and the spatial covariates we fit a Random Forest model predicting the cluster household type using information related to the location of the clusters, such as elevation and distance to roads. A Random Forest is a supervised learning technique that can be used for both classification as regression (Breiman 2001). It is an ensemble method meaning that it combines information from multiple fitted models so to obtain better predictive performance. In the case of a Random Forest these building blocks are called Decision Trees. A Decision Tree is a classification or regression model aiming to predict the class or value of a certain outcome measure by generating splits on the input variables.[10] Splits are chosen so to minimize a cost function, e.g. sum of squares in case of regression.

For sake of anonymity the DHS provides GPS coordinates that are displaced by up to two kilometer in urban areas, up to five kilometer

[10]Would we for example have one input variable x and output variable y a decision tree could hold the following information: if $x < 3$ then $\bar{y} = 2$, if $x \geq 3$ then $\bar{y} = 7$.

in rural areas, and up to ten kilometer in a random one percent of the rural cases (Perez-Heydrich et al. 2013). Therefore, rather than using the actual value of the spatial covariates at the given GPS locations, we extract for each cluster the average, minimum, and maximum value of the spatial covariates within a radius of five kilometers around the cluster coordinates and use those generated covariates in the Random Forest model. When applying the k-means method for clustering we selected only the clusters in Oshikoto to ensure that we only define household types which are meaningful to our region. When fitting the Random Forest model however, we use all the information available for Namibia, to avoid overfitting due to the small sample of clusters in Oshikoto (N=38). We thus apply the household type definition constructed using Oshikoto data to all the clusters in the DHS survey before running the model.

Finally, using the estimated Random Forest model, we predict for each 100m grid cell h in Oshikoto the probability p_{hk} that the average household is of household type $k, k = 1, \ldots, 7$. We combine these probabilities in seven grids, one for each household type, which can be seen as probability surfaces. Because the spatial covariates give little extra information about the possible variation within very dense areas (in our case the city Tsumeb), we inspected satellite imagery of those areas to create an extra probability layer with subjective probabilities of the presence of rich households in those areas. This layer is multiplied with the probability surfaces for urban household types, to force a more realistic assignment of household types within cities.

Step 2: Generate a realistic synthetic population.

We build a realistic synthetic base population from the Public Use Microdata Sample (PUMS) of the 2011 Namibian Census, using the R-package simPop. We first generate a set of households by sampling household ids from the Census PUMS file using the provided household weights recalibrated to the total number of observations per constituency in the household point location file. The variables age, gender, and relationship (i.e. head, child etc.) of the household members in the households corre-

sponding with these household ids are replicated from the Census PUMS. Next, the household attributes water, toilet, structure, space, and fuel and the individual attribute education are predicted using multinomial models. We thus create synthetic households with combinations of characteristics that are similar to those in the census PUMS, while allowing for combinations of characteristics not present in this sample. In this way anonymity of the real households is preserved (Templ et al. 2017).

Using the characteristics of our synthetic population and the probability surfaces we can now assign the synthetic households to a household location. We start by assigning a household type, as defined in step 1, to each household in our synthetic population. Next, we want to assign to each household location j the probability q_{jk} that it holds a household of household type k, $k = 1, \ldots, 7$. Assuming that within a single grid of 100 m^2 there is no clustering of household types[11], we can set the probabilities for the household locations (j) equal to the probabilities of the grid cell (h) that they fall in. That is,

$$q_{jk} := p_{hk} \quad \text{if } j \in h, \quad \text{for } j = 1, \ldots, N_j, h = 1, \ldots, N_h, k = 1, \ldots, 7.$$

Now, if household i is of household type k, the probability that it is located at household location point j is given by

$$r_{ij} = \sum_{k=1}^{7} 1_{[\text{hhtype}=k]} \frac{q_{jk}}{\sum_{j=1}^{N} q_{jk}} \quad \text{for } i, j = 1, \ldots, N_j.$$

Based on this information we iteratively assign individuals to locations by applying the following steps for each constituency x urban-rural group of household points/households separately:

0. Let H be a list of household ids and corresponding household types, and let L be a list of locations and corresponding probabilities q_{hk}.
1. Randomly select a household i from H.

[11] We should note here that this does not mean that we do not assume any spatial clustering, that is, each different grid cell does have a different prevalence of household types.

2. Select a location j from L by sampling with probability weights r_{ij}.
3. Remove household i from H and remove location j from L.
4. Repeat step 1-3 until all households are allocated.

Step 3: Clustered household characteristics

We now have a base synthetic household population representative of space, but it does not yet contain the characteristics of interest to us, that is household wealth index, womens use of contraception, DPT3 vaccination coverage. The PUMS does not contain any information on these variables, so that they could not be added in step 2. However, the DHS does contain information on these characteristics as well as the base characteristics of our synthetic population. Therefore, for each variable we thus fit a multinomial model[12] on the DHS using the base characteristics as dependent variables, and subsequently predict these variables for our synthetic population.

Degree of realism of the population

For the generated synthetic household populations to be useful for our research it should have realistic population and EA-level properties. We are confident that this is the case because (1) population means of characteristics in the synthetic populations are equivalent to the population means of the 20% census PUMS; (2) Constituency-level means of characteristics in the synthetic population are equivalent to constituency-level means of the 20% census PUMS; (3) EA-level maps of prevalences show realistic spatial distributions of characteristics; (4) density plots of EA prevalences based on the synthetic population look sufficiently similar to plots based on DHS data; (5) the DHS sample is a potential sample from our synthetic populations. For the tables and figures supporting these claims we refer the reader to Thomson et al. (2018).

[12]The SimPop package used in step 1 employs the same model to build up the synthetic population.

5.2.2 Calculating the minimal number of clusters

We will calculate the minimal number of clusters required for one- and two-stage cluster sampling by means of bootstrapped samples from our synthetic population. In this section we explain (1) the set-up of our one- and two-stage sampling procedures, (2) the different scenarios of within EA clustering, and (3) the algorithm used to search for the minimal number of clusters for each combination of scenario and sampling method.

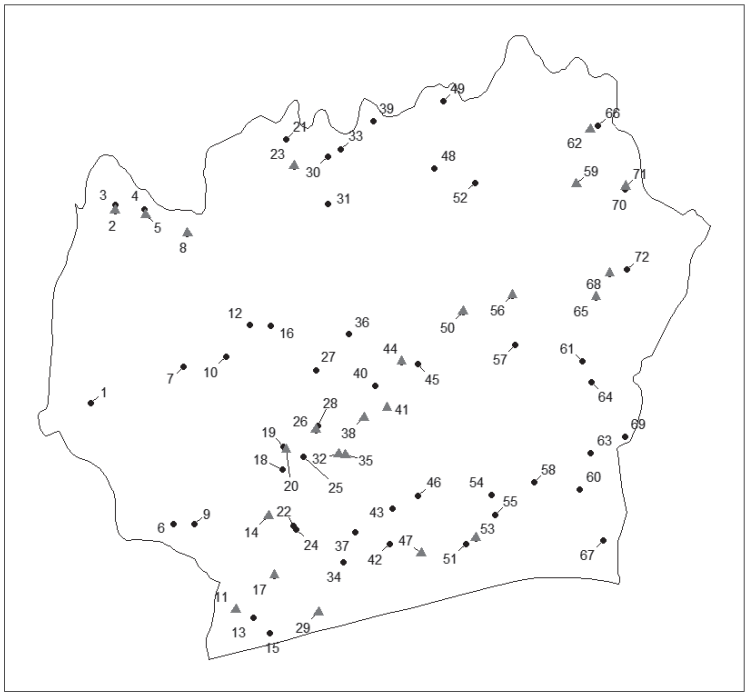
Cluster sampling set-ups

We design the two-stage cluster sampling approach such that it corresponds to the approach adopted for the 2013 DHS in Namibia. The Enumeration Areas from the 2011 Census are used as PSUs, holding on average 86 households. Since we know the coordinates of the household locations, we can easily retrieve the accompanying EA using a shapefile containing the boundaries of the EAs used in the 2011 Census. First a given number of EAs is randomly selected, after which from each of the selected EAs 25 households are systematically selected for interviews. That is, we order households within a given EA first by longitude and then by latitude. Then we randomly select one of the first $n := floor(EAsize/25)$ households on the list and from there on select every nth next household on the list.[13] An example of a two-stage sample is given in Figure 5.1

When we design our one-stage sampling set-up we aim to design PSUs of approximately 25 households, so that the size of the sample taken from one cluster is the same in the one- and two-stage cluster sampling approaches. We will call these groups of 25 households *segments*, since we define them by 'segmenting' the EAs in blocks of 25 households. That is, we order the households within a given EA first by longitude and then by latitude. Then we assign the first $m := floor(EAsize/n)$ households to one segment, the next m households to a second segment and so on. As a result the whole region will be divided into small segments of approximately 25 households. The one-stage cluster sample results from randomly selecting

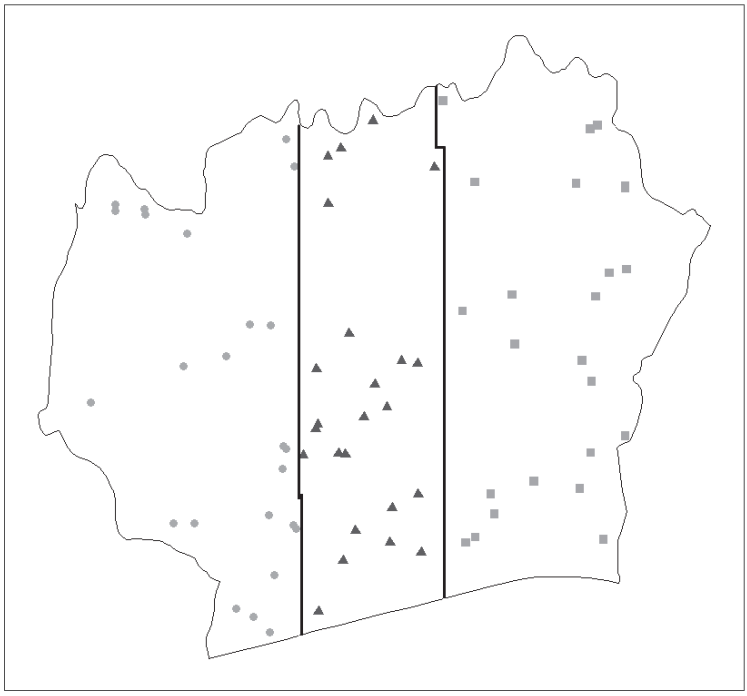
[13] We take every nth household on the list rather than a completely random set of households, with the aim to mimic as close as possible the sampling process in the field.

Figure 5.1: An example of 2-stage sampling of ordered household points and selection.



The numbers in the figure represent the ordering of households first by longitude and then by latitude. The selected households in the two-stage cluster sample are indicated by a triangle.

Figure 5.2: An example of 1-stage sampling segments.



For a one-stage cluster sample the given EA would be divided into three segments, the households indicated with dots would fall in segment 1, those indicated with triangles in segment 2, and those indicated with squares in segment 3. This is stressed by the two lines dividing the EA in three parts.

a given number of these segments. An example of segments for one-stage sampling are given in Figure 5.2.

Scenarios of spatial clustering

In order to analyze the performance of one- and two-stage cluster samples under several (extreme) cases of spatial clustering within EAs, we relocate households within EAs according to several scenarios. We start by ordering all household locations within one EA first by longitude and then by latitude. Next, we order the list of households in the same EA according to one of the scenarios given in table 5.3. For example, for scenario 3 we order the households from low wealth to high wealth. Then we attach the ordered list of households to the ordered list of coordinates and define the one-stage cluster segments as explained above. Figure 5.3 shows an example of the spatial distribution of wealth within an EA for the baseline case (scenario 1) and the case where wealth index is perfectly clustered (scenario 3).

The assumption of perfect clustering by our key outcome variables is rather extreme. A more moderate assumption would be that households are clustered by other household variables, such as access to improved toilet or water facilities, which are related to our outcome variables but do not necessarily lead to perfect clustering in these variables (scenarios 6-10). Scenarios 11-13 also represent more moderate configurations of clustering, by randomly relocating 50% of the households from the extreme scenarios 3-5 within each EA.

Calculating minimal number of clusters

A common way to express the requirements for sample estimates is by defining a bandwidth of B percentage points from the population mean in which $(100 - \alpha)\%$ of sample means should fall. For example, the EPI framework was once developed with the aim in mind that it should provide estimates of immunization coverage which are with 95% certainty within 10 percentage points from the true value (Bennett et al. 1991). In

Figure 5.3: An example of the spatial clustering of wealth within one EA for scenario 1 (top) and scenario 3 (bottom).

Table 5.3: Within EA clustering scenarios

Scenario		Description
1.	Baseline	Original ordering of households (HHs)
2.	Perfectly homogeneous	HHs randomly reordered
3.	Wealth index perfectly clustered	HHs ordered from low to high wealth
4.	Contraception perfectly clustered	HHs ordered from low to high prevalence of contraception, where HHs without women aged 15-64 positioned at random
5.	DPT3 perfectly clustered	HHs ordered from low to high prevalence of DPT3 vaccination, where HHs without children under 5 positioned at random
6-10.	Perfect clustering by underlying variables	HHs ordered by access to improved toilet facility (6), improved water (7), adequate structure (8), adequate space (9), or non-solid fuel (10).
11-13.	Moderate clustering	From scenarios 2-4 randomly replace 50% of HHs

that case $B = 10$ and $\alpha = 5$. We find the minimal number of clusters for which this condition holds in an iterative way, by partitioning the search space. First, we choose the maximum number of clusters to be included in the sample (n_{max}), for example n_{max} could be set to 15% of all clusters. Then, we calculate a bootstrapped $(100 - \alpha)\%$ confidence interval based on 10.000 samples using the maximum number of clusters and the relevant cluster sampling approach. If this confidence interval is wider than the benchmark, we set the sample size requirements equal to n_{max}. If it is smaller however, we recalculate the bootstrapped $(100 - \alpha)\%$ confidence interval using $0.5 * n_{max}$ clusters. If this confidence interval is wider than the benchmark, we next evaluate the bootstrapped confidence interval when using $0.75 * n_{max}$ clusters, if it is smaller, we next evaluate the bootstrapped confidence interval at $0.25 * n_{max}$ clusters and so on until we find a bootstrapped confidence interval that is (almost) equal to the desired size.

We repeat this process using both cluster sampling approaches and for each scenario on five realizations of our synthetic population.

5.3 Data description

We perform the analysis on a set of five synthetic populations. In this section we show the data description of one of the five synthetic populations. The tables and figures for the other populations are provided in Appendix 5.B.

Table 5.4 provides summary statistics for the whole region. Synthetic population 1 consists of 179,931 individuals living in 37,298 households. The households are mostly located in rural area's. A quarter of the household heads has had no formal education, almost a third has an incomplete primary degree and only 14.2% has completed secondary or tertiary schooling. Many are lacking access to improved toilet facilities (79.9%), adequate structure (61.4%), or non-solid fuel (83.8%). However, only few lack access to improved water facilities (26.8%) or have a house with inadequate space (7.5%). Compared to the rest of Namibia, the area is rather poor, with 62.1% of households falling in the lowest two wealth quintiles, and only 17.9% in the highest two wealth quintiles. The individuals are rather young, although compared to the average of Sub-Saharan African countries there are relatively few children under five (14.0% compared to 16.4%, UN (2017)) and relatively many individuals older than fifty (13.6% compared to 9.8%, UN (2017)). The use of modern contraception under 15-49 year old women is at 43.1% and 80.3% of children under five have received all three DPT vaccinations.

Figure 5.4 shows a map of the EA-level prevalences as measured in synthetic population 1 for each of the three key characteristics: individuals in the poorest wealth quintile, women 15-49 using modern contraception, and children under 5 who have received three DPT vaccinations.[14] All three maps show substantial spatial variation. The top graph shows the prevalence of individuals in the poorest wealth category. There is a clear relationship between the prevalence of poverty measured by asset indexes and accessibility of areas. The poorest never live in the urban areas and rarely live in the regions near a large road or in more densely populated

[14]Note, these maps do not depend on the chosen scenarios, as these scenarios only result in within EA relocations of households. The EA-level prevalences are thus not affected by the choice of scenario.

Figure 5.4: Prevalence of characteristics per EA - Synthetic population 1

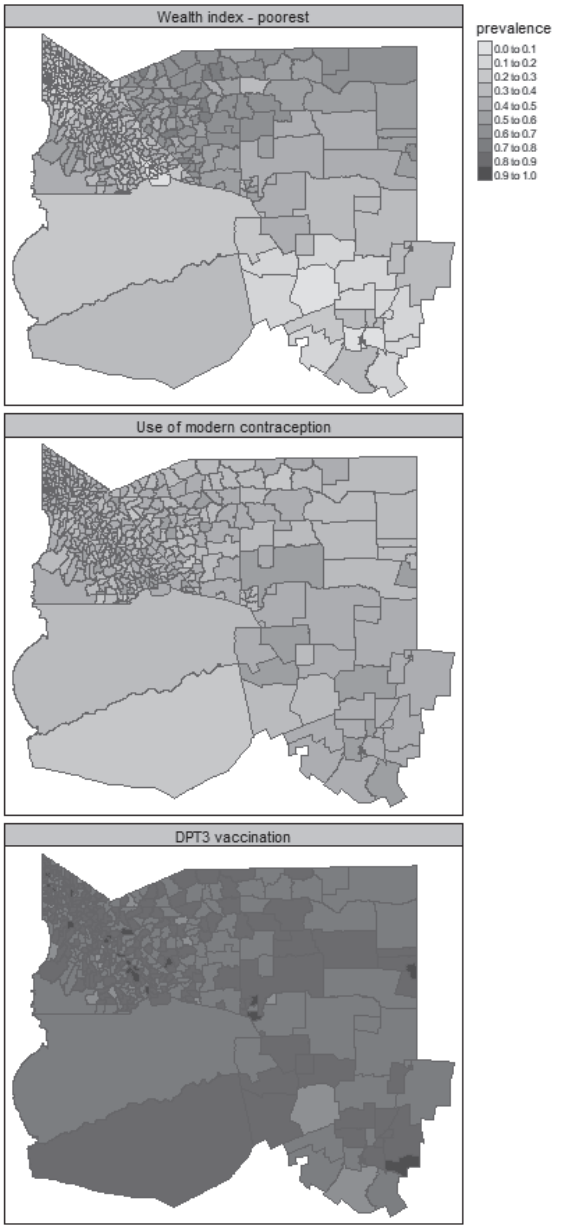


Table 5.4: Summary statistics - Synthetic population 1

household-level variables		individual-level variables	
nr households	37298	nr individuals	179931
average household size	4.82	male	47.8%
urban	15.7%	age:	
education head		- 0 - 4	14.0%
- no formal	25.1%	- 5 - 14	26.3%
- incomplete primary	30.1%	- 15 - 49	46.1%
- complete primary	30.6%	- 50 plus	13.6%
- complete secondary	10.7%		
- complete tertiary	3.5%	nr women 15-49	42785
unimproved water	26.8%	modern contraception	43.1%
unimproved toilet	79.9%		
inadequate space	7.5%	nr children under 5	25249
inadequate structure	61.4%	DPT3 vaccination	80.3%
solid fuel	83.8%		
wealth index			
- poorest	33.9%		
- poorer	28.2%		
- middle	19.7%		
- richer	13.5%		
- richest	4.6%		

areas.[15] The prevalence of the use of modern contraception shows a less clear spatial pattern. Prediction models also show that the use of modern contraception is only weakly related to indicators like education and access to improved water facilities (which do show a distinct spatial pattern) and more so to variables such as age (which is more uniformly distributed over space). The prevalence of DPT3 vaccination seems to be somewhat higher in more densely populated areas. However, also in the case of DPT3 vaccination there is substantial variation unrelated to the observed factors that show distinct spatial patterns.

The different scenarios defined in table 5.3 are likely to lead to fairly different segment-level ICCs. Figure 5.5 shows the segment-level ICCs together with boxplots of the segment-level prevalences. We can compare these to the EA-level ICC and boxplots reported in the same figure. The most left boxplot shows the spread of EA prevalence and confirms the image painted in Figure 5.4. There is substantial spread in the prevalence

[15] Figures of population density and distance to roads can be found in Appendix 5.A.

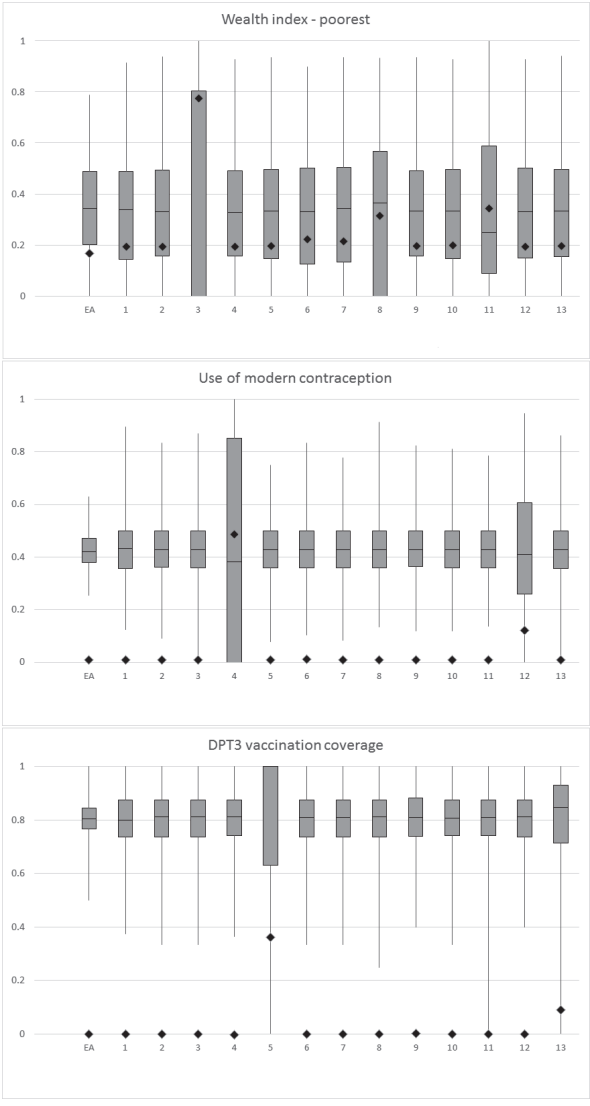
of being in the poorest wealth category, with half of EA prevalence levels falling within about 14 percentage point of the median EA prevalence. The spread of prevalence of modern contraception and DPT3 vaccination is more centered, with half of EAs having a prevalence within about 4 percentage point of the median EA prevalence. This is also reflected in the ICC, which is 0.17 for the poorest wealth category, but respectively 0.01 and 0.001 for use of modern contraception and DPT3 vaccination. The EA-level will be the basis of our two-stage cluster sampling procedure. The ICCs found in our synthetic population are close to the observed ICCs in the DHS 2013 sample, which are 0.23 for the poorest wealth category, -0.01 for contraception, and 0.02 for DPT3.[16]

For the one-stage cluster sampling procedure we cut the EAs into segments holding about 25 households (based on a list ordered by the (x,y) coordinates of households), relocated within their EA according to different scenarios. In the baseline case households are kept at their original location (scenario 1, second boxplot from the left). The spread of segment prevalences is somewhat larger than the spread of EA prevalences. ICC levels are similar in case of contraception and DPT3 and only slightly higher for the wealth indicator, such that we do not expect large increases in sample size requirements when moving from a two-stage cluster sampling design to a one-stage cluster sampling design.

The next four scenarios depict four possible extreme cases of spatial clustering within EAs: (2) complete homogeneity, (3) perfect clustering by wealth, (4) perfect clustering by contraception, (5) perfect clustering by DPT3 vaccination. The spread and ICC in scenario 2 are almost equivalent to the baseline scenario. When we assume perfect (within EA) clustering by wealth index (scenario 3) or contraception (scenario 4) the spread on the respective variable increases substantially. This is what one would expect, since within every EA the poorest households, those with high fractions of women using modern contraception, or those with high fractions of

[16]These figures are slightly lower than the national-level ICCs for Namibia, which are respectively 0.41, 0.04, and, 0.09. There is quite some variation in regional-level ICCs for the different regions in Namibia and Oshikoto represents a rather average regional case for Namibia. The national-level ICCs and regional-level ICCs for all regions in Namibia are reported in Appendix 5.D.

Figure 5.5: ICC (♦) and boxplots of prevalences per EA/segment - Synthetic population 1



children with a DPT3 vaccination were placed close together, so that they end up in the same segment. Segments are thus likely to have either very high prevalence rates on the respective variables or very low prevalence rates. In scenario 4, the prevalence of use of modern contraception falls within 40 percentage point of the median prevalence level for 50% of the segments. The difference between the minimum and maximum prevalence of DPT3 also increases under the assumption of perfect clustering (scenario 5). In this case however the bulk of prevalences is centered at the top, as the median prevalence is equal to 100%. The average prevalence of DPT3 vaccination is quite high, such that clustering and segmenting will likely lead to many segments with 100% prevalence and only few with slightly lower prevalences.[17] ICC levels increase substantially after clustering, most extremely for wealth (from 0.19 under the baseline scenario to 0.77 under scenario 3) and contraception (from 0.01 to 0.49 under scenario 4), and somewhat more moderately for DPT3 (from 0.00 to 0.36 under scenario 5). Clustering by wealth also slightly affects the spread of prevalences of contraception and DPT3, though the ICC levels are not affected. Based on the reported increases in ICC we expect that there may be large differences in the sample size requirements of one- and two-stage cluster sampling, under the assumption of perfect clustering.

Clustering by underlying factors (scenarios 6-10) leads to a slight increase in the spread of the prevalences. The increase is substantially larger for the wealth indicator when we assume households are perfectly clustered by structure (scenario 8). This may be because the percentage of households with an inadequate structure is closer to 50%, so that there is more potential for clustering or because the wealth and structure are more closely related. Also when replacing 50% of households from scenarios 3-5 (scenarios 11-13) ICC levels only increase slightly.

[17] For example suppose we have an EA with 75 households and a household characteristic with a prevalence rate of 50%. If we would order the households by prevalence and then segment into three groups, the segments would have prevalence rates of 0%, 50%, and 100%. However, would the EA prevalence have been equal to 90%, the resulting segment-level prevalences would equal 70%, 100%, and 100%.

Results 5.4

Baseline analysis 5.4.1

For each of the five synthetic populations we calculated the minimum number of clusters necessary to obtain a sample estimate which is with 95% certainty within 5 percentage points from the population mean. Table 5.5 reports the averages of the minimum number of clusters found in the five synthetic populations. The results for the five synthetic populations separately can be found in Appendix 5.C. There is only little variation between the results for the different synthetic populations.

The minimum number of clusters under a two-stage cluster sample are given in row 1. In case of a two-stage cluster sample we need to sample sixty-five clusters to obtain a sufficiently precise estimate of the proportion of individuals in the poorest wealth quintile (column 2), fifteen clusters for the proportion of women using modern contraception (column 4), and also fifteen clusters for the proportion of under five year old children with a DPT3 vaccination (column 6). The relatively large requirements for the wealth indicator stem from the fact that this measure has a higher ICC than the other variables. Moreover, the variable is perfectly correlated within households, so that each extra household virtually only adds one extra data point (like in the case of perfectly clustered EAs/segments). Would we want to estimate all three variables sufficiently precise in one survey, we would need sixty-five clusters (column 8).

The minimum number of clusters to obtain a precise enough one-stage cluster sample estimate are given in rows 2 to 14. Under the baseline scenario (scenario 1) these are slightly higher than the two-stage cluster sampling requirements. The average required number of clusters increase by 1.1 times to seventy clusters for the wealth indicator, increases by 1.2 times to eighteen clusters for contraception, and remain fifteen clusters for DPT3 vaccination status. This is not too different from the results for the case of perfect homogeneity within EAs (scenario 2), indicating that our baseline scenario is likely very similar to a perfect homogeneous within EA setting. In our model we assumed perfect within grid cell

Table 5.5: Required number of clusters - Baseline

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	65	-	15	-	15	-	65	-
1. Baseline	70	1.1	18	1.2	15	1.0	70	1.1
b. Extreme scenario's								
2. Homogeneous	70	1.1	18	1.2	16	1.0	70	1.1
3. Clustered by wealth index	282	4.3	19	1.3	17	1.2	282	4.3
4. Clustered by contraception	70	1.1	188	12.5	16	1.1	188	2.9
5. Clustered by DPT3	75	1.1	19	1.3	99	6.6	99	1.5
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	89	1.4	18	1.2	16	1.0	89	1.4
7. Clustered by improved water	75	1.1	18	1.2	16	1.1	75	1.1
8. Clustered by adequate structure	94	1.4	20	1.4	17	1.2	94	1.4
9. Clustered by adequate space	70	1.1	18	1.2	15	1.0	70	1.1
10. Clustered by non-solid fuel	70	1.1	18	1.2	16	1.1	70	1.1
d. Moderate clustering								
11. 50% replaced from 3.	141	2.2	18	1.2	17	1.1	141	2.2
12. 50% replaced from 4.	70	1.1	58	3.9	16	1.1	70	1.1
13. 50% replaced from 5.	70	1.1	19	1.3	43	2.8	70	1.1

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.
The minimum number of clusters is the average of the minimum number of clusters found in the five synthetic populations. The results for each of the five synthetic populations seperately can be found in Appendix 5.C.

homogeneity when assigning households to household locations. Since an EA usually consists of more than one grid cell this assumption does not need to lead to perfect within EA homogeneity. However, the different cells within one EA likely have rather similar spatial characteristics, so that they also get assigned a similar mix of households. Given this model, it is thus reasonable to expect moderate within EA homogeneity. However, households may also be located at certain locations for reasons that cannot be captured by observables in a model, so that also perfect within EA clustering is still a reasonable assumption.

Assuming perfect clustering by wealth index (scenario 3) increases the one-stage sample size requirements to on average 282 clusters. This is 4.3 times higher than the two-stage sample size requirements. Perfect clustering by contraception (scenario 4) increases the the sample size requirements to 188 clusters, an increase of 12.5 times the two-stage

sample size. Perfect clustering by DPT3 vaccination increases the sample size requirements from 15 to 99. In all cases the sample size requirements for the other variables are hardly affected. Although the sample size requirements are the largest for the wealth indicator, the increase in sample size requirements is starkest for contraception. The EA-level ICC on this variable is very low, so that clustering and segmenting within the EA can have large impacts. Also DPT3 vaccination has a low EA-level ICC, but contrary to contraception it has a high overall prevalence rate, so that even after within EA clustering and segmenting, the different EAs look pretty similar.[18] Even though perfect clustering leads to large increases in the sample size requirements for contraception and DPT3, the number of clusters needed for a complete household survey are affected only moderately. Clustering makes these variables look more similar to the wealth index in terms of ICC. Sample size requirements for the complete survey are at most 4.3 times the requirements for a two-stage cluster sample.

It may be more reasonable to assume that not wealth, contraception, and DPT3 are perfectly clustered, but that the underlying characteristics are perfectly clustered, i.e. if one does not have access to improved water, his or her neighbor probably does not either. Under that assumption the differences between one- and two-stage clustering are less extreme. DPT3 vaccination is hardly correlated with water, toilet, structure, space, or fuel, so that clustering on these variables is 'as if' there is complete homogeneity for the variable of interest. The wealth index is more strongly correlated with the underlying variables so that an increase in clusters is required to estimate the prevalence of poorest sufficiently precise, from 70 to at most 94 clusters. The largest difference is found when clustering by the adequacy of structure. When we assume that there is only moderate

[18]For the scenario 4 and 5 we assumed that the households without respectively women aged 15-49 or children under 5 (those with missing values on use of contraception/DPT3) were randomly located in between the households ordered by use of modern contraception/DPT3 vaccination. We could also assume that not only families with vaccinated children (women using modern contraception) live close together, but also families with young children (women aged 15-49) in general. When we would apply this assumption, the sample size requirements for contraception remain at 188 clusters, but the sample size requirements for DPT3 decrease to 65.

within EA clustering (scenario 11-13), the number of clusters increase slightly: a doubling of clusters in case of wealth, almost four times the amount in case of contraception, and almost tripling for DPT3.

The results for the different populations are comparable, though variation exist. For example, for the scenario of moderate clustering by contraception, increases in sample sizes lie between 3.1 and 4.7 times the sample size of two-stage cluster sampling. But the differences in panel c. are at most a factor 0.7.

5.4.2 Increasing sample sizes per cluster

The situation described above fits well to the case of multi-topic LMIC household surveys. However, in case of a topic specific survey focusing on a specific subsample of the population, for example immunization surveys, one may be tempted to enroll more households per cluster, because not every household will have an eligible household member.

Table 5.6 shows the required number of clusters for DPT3 vaccination under the assumption of (1) a two-stage cluster sample enrolling 25 households, (2) a one-stage cluster sample enrolling all households per cluster, (3) a one-stage cluster sample enrolling 50 households per cluster for the scenario's as described above, (4) a one-stage cluster sample enrolling 75 households per cluster for the scenario's as described above. Would one opt for generating segments of on average 50 households, rather than 25, eight clusters should be enrolled to achieve an accurate coverage estimate under the baseline scenario. In case of 75 households per segment this reduces to six clusters and would we enroll all households per cluster it reduces further to five clusters. The larger the segments, the less relevant the different scenario's become. As we have on average 86 households per EA, segments of size 75 are as if we are sampling the whole EA in most cases.

Extreme clustering by DPT3 status still results in a rather large sample size in the case where segments hold 50 households: 24 segments should be sampled, which is on average equal to 1200 households, three times

Table 5.6: Required number of clusters - increasing cluster size

	50 households per segment			75 households per segment		
	nr.	rel. diff		nr.	rel. diff	
Scenario	clusters	clusters	HHs[a]	clusters	clusters	HHs[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	5	0.4	1.2	-	-	-
1. Baseline	8	0.5	1.0	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	7	0.5	1.0	6	0.4	1.2
3. Clustered by wealth index	7	0.5	0.9	6	0.4	1.2
4. Clustered by contraception	7	0.5	1.0	6	0.4	1.2
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	8	0.5	1.0	6	0.4	1.2
7. Clustered by improved water	8	0.5	1.1	6	0.4	1.2
8. Clustered by adequate structure	7	0.5	1.0	6	0.4	1.2
9. Clustered by adequate space	7	0.5	0.9	6	0.4	1.2
10. Clustered by non-solid fuel	7	0.5	0.9	6	0.4	1.2
d. Moderate clustering						
11. 50% replaced from 3.	6	0.4	0.9	6	0.4	1.2
12. 50% replaced from 4.	6	0.4	0.9	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.
The minimum number of clusters is the average of the minimum number of clusters found in the five synthetic populations. The results for each of the five synthetic populations seperately can be found in Appendix 5.C.

more than in the baseline two-stage sample. The number of households at most doubles when taking segments of 75 households.

5.5 Discussion

The results provide us with guidelines of the possible requirements of one-stage cluster sampling under extreme situations. Whereas the variety within the modeled region (Oshikoto, Namibia) makes it an interesting case to look at, there are also some limitations to the region, which may result in more moderate outcomes then we would find elsewhere. First of all, the EAs in Oshikoto are relatively small, holding on average 86 households. Typical EAs in other countries consist of 200 or even 400/500 households. In larger EAs, the effects of clustering on the sample size requirements of one-stage cluster sampling could potentially be larger. Secondly, whereas there are relatively many households in Oshikoto falling in the poorest wealth category, none of these seem to live in the urban areas. In many other LMIC settings, you would expect to find the poorest in cities, possibly leading to more extreme clustering effects.

Another limitation of the analysis is that it does inform us about what could happen in different extreme situations, but not how likely it is to encounter such situations. Although literature gives guidance on likely levels of ICCs and more general spatial patterns, little is written about how households are located within EAs. Intuitively, the amount of within EA clustering will depend on the type of EA. One may for example expect more clustering in an EA that covers a rural village, then in an EA covering multiple farms or a city block. Similarly, it would not be unreasonable to assume that wealth levels vary from street to street, due to the difference in types of houses, whereas contraception and vaccination levels are less likely to be linked to such specific locations. There does exist a line of research devoted to the effect of scale on measurements of racial segregations in large metropolitan cities in industrial countries where there is a higher availability of geo-coded micro-data. This research indicates that there is some variation between egocentric measurements based on a 100m radius versus a 1000m radius around households or individuals (Petrović et al. 2018), and slight variations in egocentric measurements including the 50 nearest neighbors compared to 200 nearest neighbors of households or individuals (Östh et al. 2015). However, it is not clear

how these results would generalize to (rural) LMIC settings or to socio-economic and health variables.

5.6 Conclusion

In this chapter we compared the commonly used two-stage cluster sampling approach to the alternative one-stage cluster sampling approach. We generated a synthetic population of Oshikoto, Namibia to provide as a testing ground for both sampling approaches. The households in this population were assigned to realistic (x,y) coordinates, in order to simulate realistic spatial patterns of the different household characteristics. To facilitate one-stage cluster sampling we created smaller Primary Sampling Units by segmenting Enumeration Areas (EAs) under different scenarios of within EA clustering. We searched for the minimum number of clusters to obtain an adequate sample in an iterative way based on bootstrapped confidence intervals of the sample means.

The results show that in most moderate scenario's the required number of clusters for one-stage cluster sampling is fewer than twice the required number of clusters of two-stage sampling, under the assumption that the same number of households per cluster are sampled with both methods. Under extreme clustering scenarios, the required number of clusters can increase by up to thirteen times. Especially when the EA-level intracluster correlation is moderate and prevalence is close to 50%, extreme assumptions about within EA clustering have large impact on the required number of clusters. When measuring variables focusing on small subsamples of the population, it can be beneficial to apply one-stage cluster sampling with larger segment sizes. This can lower the required number of clusters to visit, while enrolling the same number of households in the survey. By increasing segment sizes, clustering scenarios also become less relevant.

Whether one-stage or two-stage clustering is more cost-effective will depend on the type of survey and the regional context. Aspects such as the length of the survey, the type of survey (does it only include a questionnaire or also the collection of biomarkers?), and the accessibility of the regions will determine the relative costs of the two types of sampling

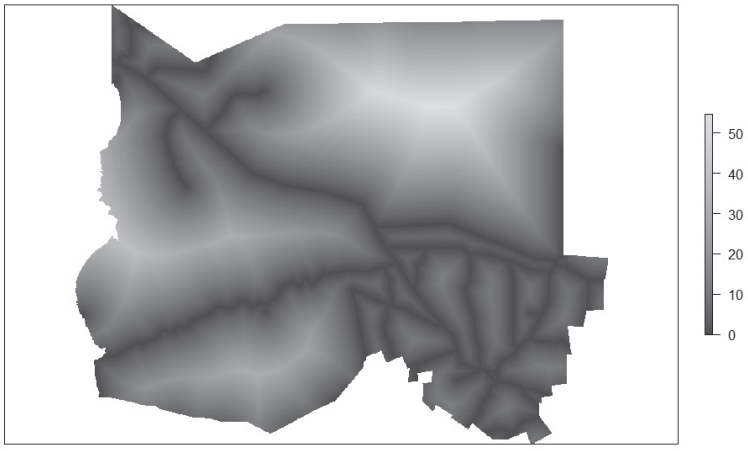
methods. The numbers provided in this chapter can serve as input when estimating these costs.

Maps of Oshikoto 5.A

Figure 5.6: Population density in Oshikoto expressed in people per pixel (roughly $100m^2$).

source: www.worldpop.org, Linard et al. (2012)

Figure 5.7: Distance to major roads (km).



source: Spatial covariate processed by the "Global High Resolution Population Denominators" Project (original source: 2016 OSM highways).

5.B Data description for synthetic populations 2-5

Table 5.7: summary statistics for synthetic populations 2-5

	pop 2	pop 3	pop 4	pop 5
nr households	37298	37298	37298	37298
average household size	4.82	4.83	4.83	4.83
urban	15.7%	15.7%	15.7%	15.7%
education head				
- no formal	24.9%	25.1%	24.8%	25.0%
- incomplete primary	30.0%	30.5%	30.3%	30.2%
- complete primary	31.2%	30.5%	31.1%	31.1%
- complete secondary	10.2%	10.6%	10.2%	10.5%
- complete tertiary	3.6%	3.4%	3.5%	3.3%
unimproved water	26.8%	27.1%	26.9%	27.3%
unimproved toilet	79.9%	80.1%	80.3%	80.5%
inadequate space	7.8%	7.7%	7.5%	7.7%
inadequate structure	61.3%	61.4%	61.5%	62.1%
solid fuel	83.6%	84.0%	84.0%	84.1%
wealth index				
- poorest	33.6%	34.0%	34.1%	34.4%
- poorer	28.7%	28.5%	28.6%	28.4%
- middle	19.6%	19.6%	19.2%	19.3%
- richer	13.4%	13.1%	13.7%	13.4%
- richest	4.7%	4.6%	4.3%	4.4%
nr individuals	179854	180233	180164	180111
male	48.0%	48.1%	48.2%	48.0%
age:				
- 0 - 4	14.2%	14.3%	14.0%	14.1%
- 5 - 14	26.2%	26.1%	26.1%	26.3%
- 15 - 49	46.2%	46.1%	46.4%	46.1%
- 50 plus	13.5%	13.5%	13.5%	13.5%
nr women 15-49	43007	42930	42903	42708
modern contraception	0.4350222	0.4329839	0.438967	0.436007
nr children under 5	25466	25742	25258	25463
DPT3 vaccination	80.4%	80.6%	80.2%	80.2%

Figure 5.8: Prevalence of characteristics per EA - Synthetic population 2

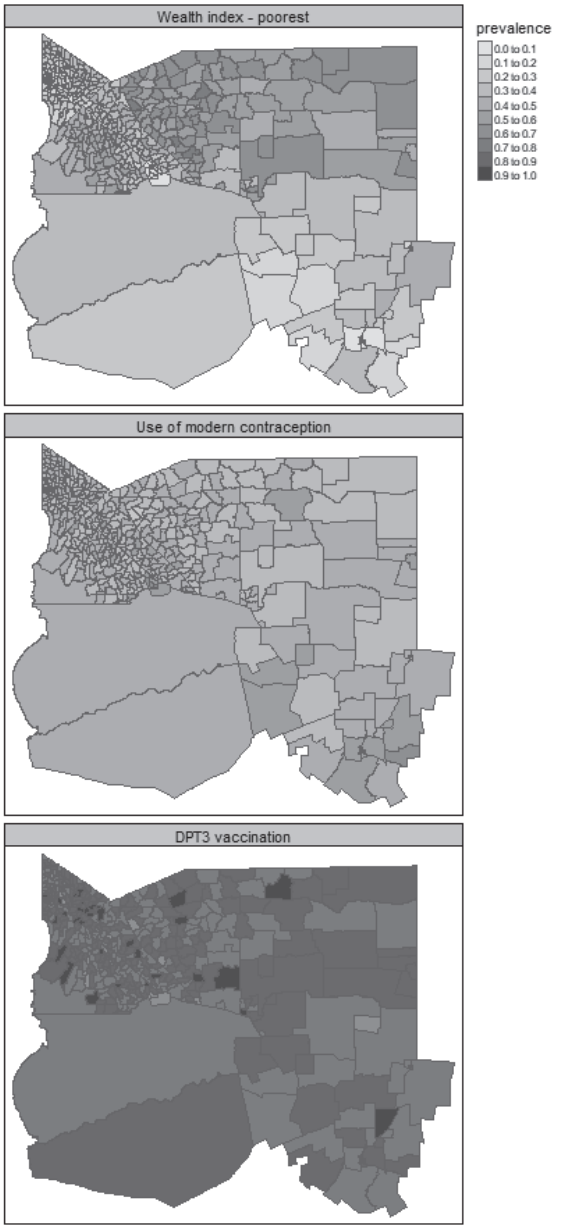


Figure 5.9: Prevalence of characteristics per EA - Synthetic population 3

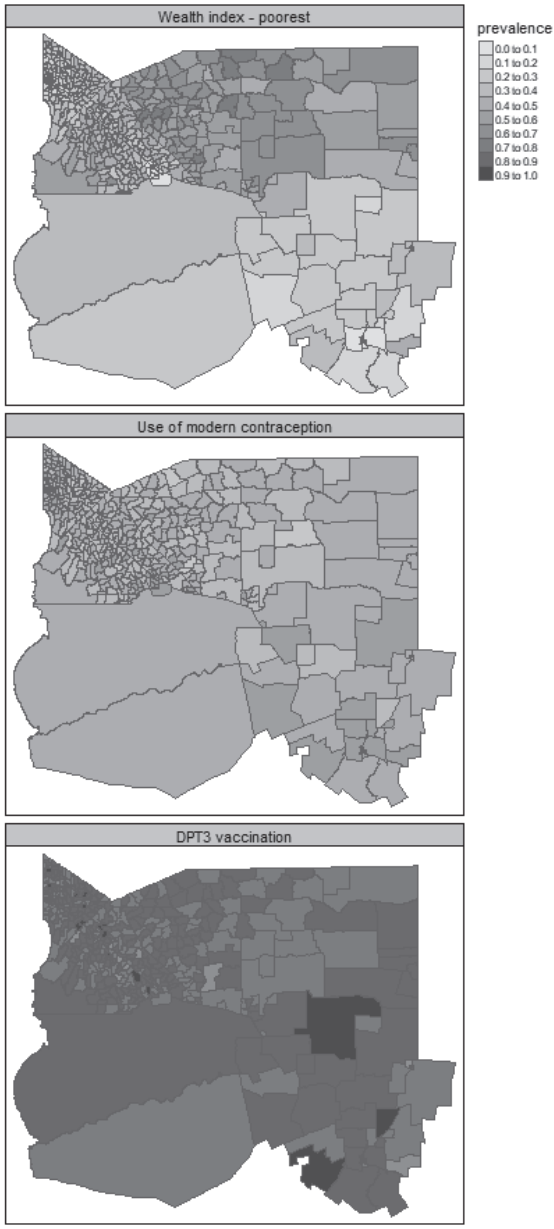


Figure 5.10: Prevalence of characteristics per EA - Synthetic population 4

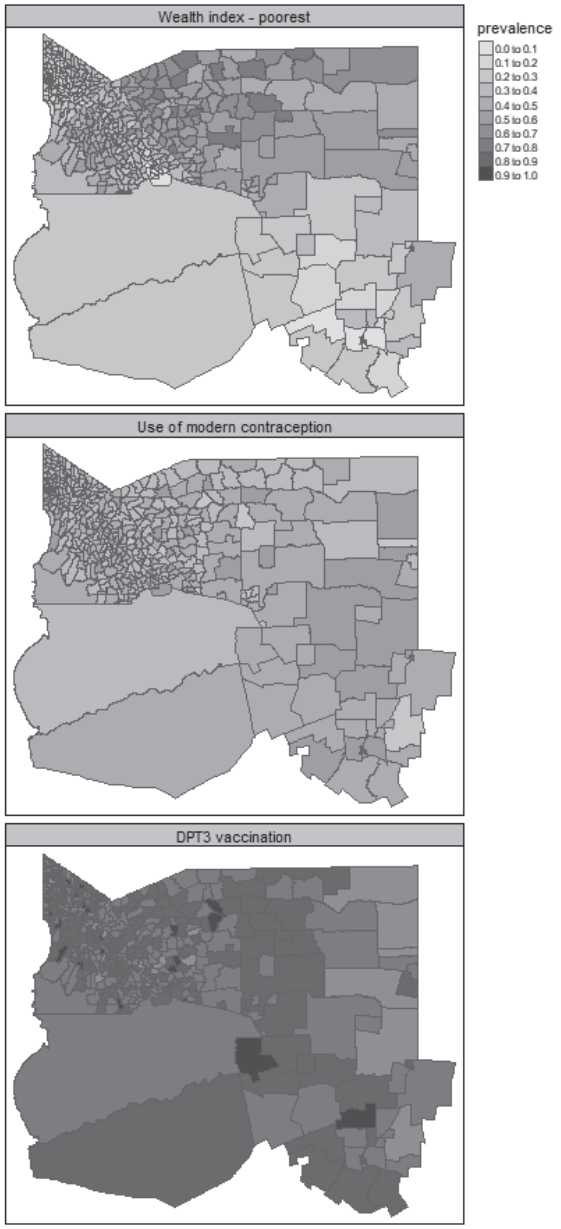


Figure 5.11: Prevalence of characteristics per EA - Synthetic population 5

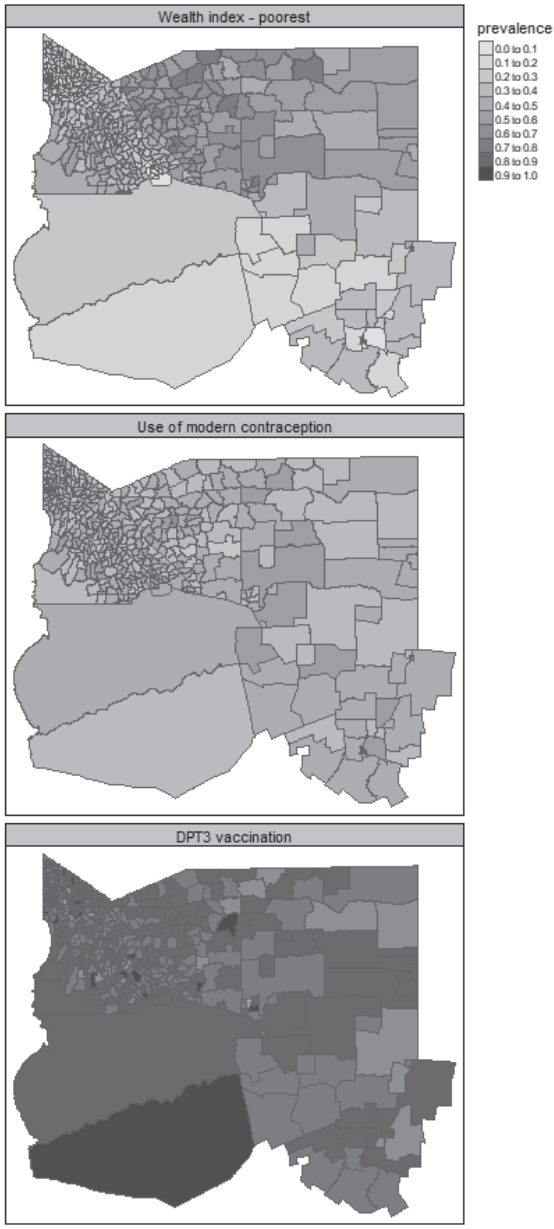


Figure 5.12: ICC (♦) and boxplots of prevalences per EA/segment - Synthetic population 2

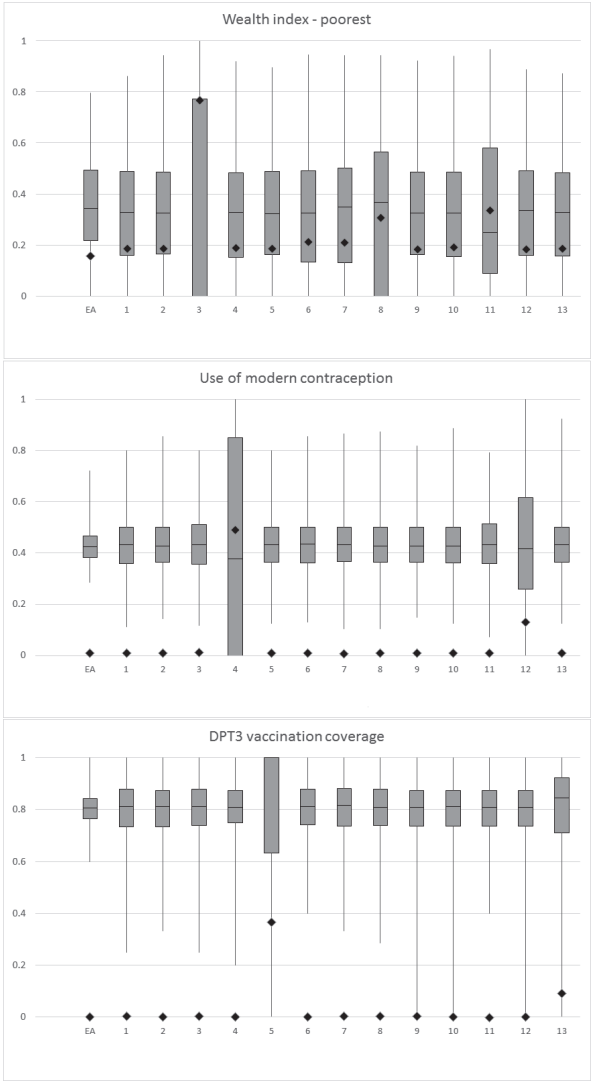


Figure 5.13: ICC (♦) and boxplots of prevalences per EA/segment -Synthetic population 3

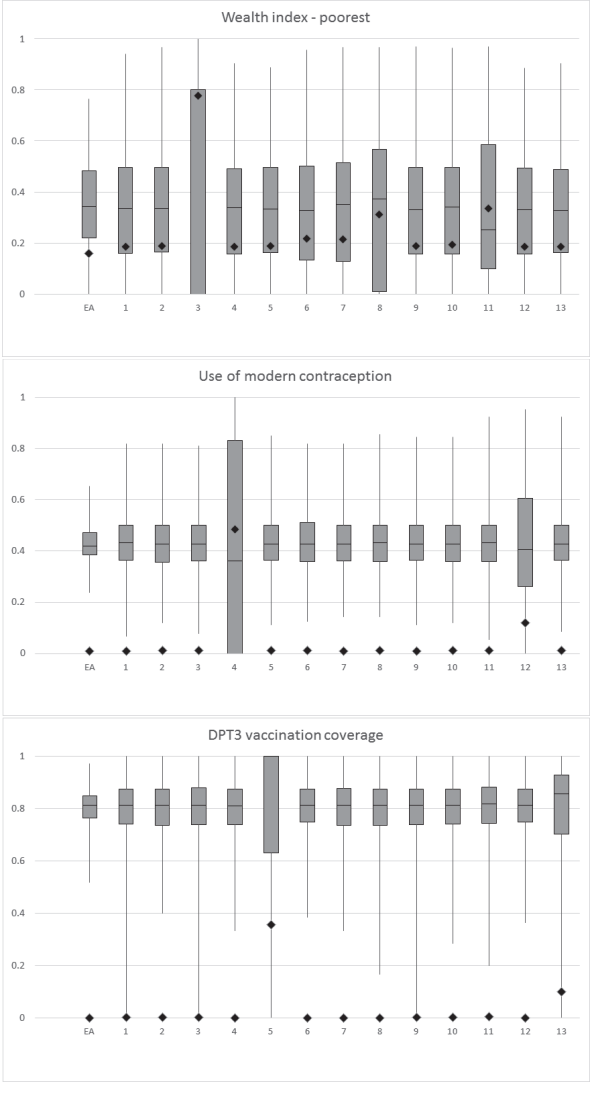


Figure 5.14: ICC (♦) and boxplots of prevalences per EA/segment - Synthetic population 4

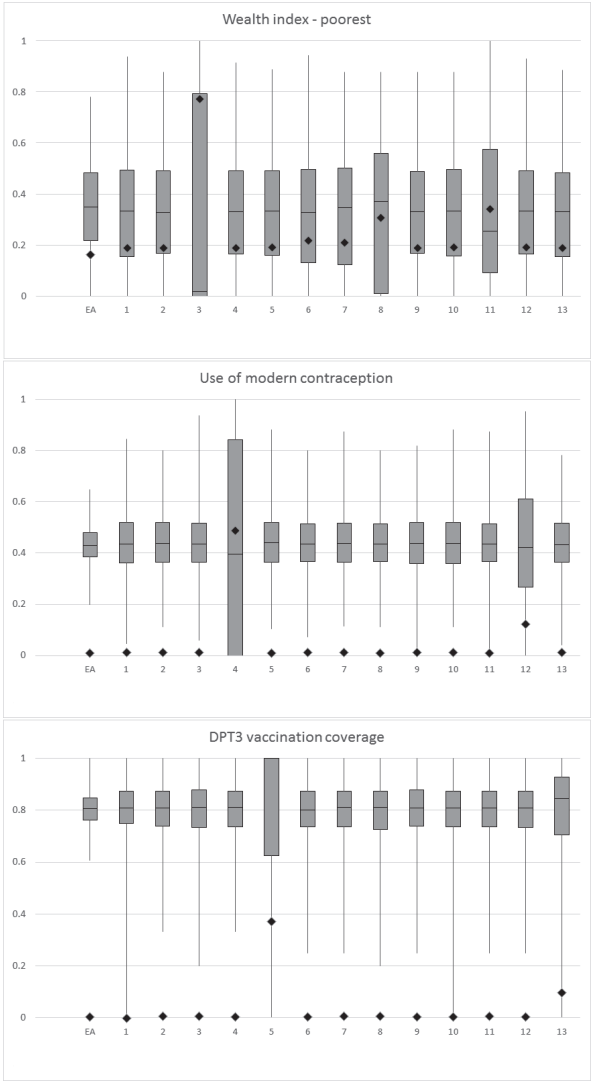
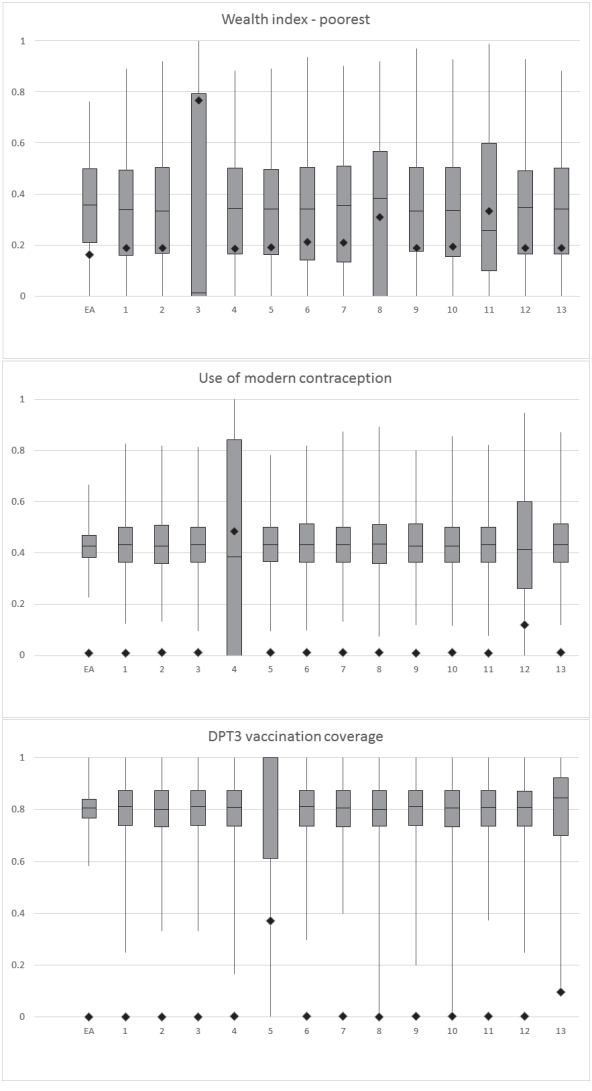


Figure 5.15: ICC (♦) and boxplots of prevalences per EA/segment - Synthetic population 5



5.C Results for each of the 5 synthetic populations

5.C.1 Baseline results

Table 5.8: Required number of clusters - Baseline - Population 1

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	70	-	15	-	15	-	70	-
1. Baseline	70	1.0	18	1.2	12	0.8	70	1.0
b. Extreme scenario's								
2. Homogeneous	70	1.0	18	1.2	18	1.2	70	1.0
3. Clustered by wealth index	282	4.0	18	1.2	18	1.2	282	4.0
4. Clustered by contraception	70	1.0	188	12.5	15	1.0	188	2.7
5. Clustered by DPT3	94	1.3	18	1.2	94	6.3	94	1.3
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	94	1.3	18	1.2	18	1.2	94	1.3
7. Clustered by improved water	94	1.3	18	1.2	15	1.0	94	1.3
8. Clustered by adequate structure	94	1.3	24	1.6	18	1.2	94	1.3
9. Clustered by adequate space	70	1.0	18	1.2	18	1.2	70	1.0
10. Clustered by non-solid fuel	70	1.0	18	1.2	15	1.0	70	1.0
d. Moderate clustering								
11. 50% replaced from 3.	141	2.0	18	1.2	18	1.2	141	2.0
12. 50% replaced from 4.	70	1.0	58	3.9	18	1.2	70	1.0
13. 50% replaced from 5.	70	1.0	18	1.2	36	2.4	70	1.0

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.9: Required number of clusters - Baseline - Population 2

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	70	-	15	-	15	-	70	-
1. Baseline	70	1.0	18	1.2	18	1.2	70	1.0
b. Extreme scenario's								
2. Homogeneous	70	1.0	18	1.2	12	0.8	70	1.0
3. Clustered by wealth index	282	4.0	18	1.2	18	1.2	282	4.0
4. Clustered by contraception	70	1.0	188	12.5	18	1.2	188	2.7
5. Clustered by DPT3	70	1.0	18	1.2	94	6.3	94	1.3
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	94	1.3	18	1.2	18	1.2	94	1.3
7. Clustered by improved water	70	1.0	18	1.2	18	1.2	70	1.0
8. Clustered by adequate structure	94	1.3	18	1.2	18	1.2	94	1.3
9. Clustered by adequate space	70	1.0	18	1.2	15	1.0	70	1.0
10. Clustered by non-solid fuel	70	1.0	18	1.2	15	1.0	70	1.0
d. Moderate clustering								
11. 50% replaced from 3.	141	2.0	18	1.2	15	1.0	141	2.0
12. 50% replaced from 4.	70	1.0	70	4.7	15	1.0	70	1.0
13. 50% replaced from 5.	70	1.0	18	1.2	47	3.1	70	1.0

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.10: Required number of clusters - Baseline - Population 3

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	47	-	15	-	15	-	47	-
1. Baseline	70	1.5	18	1.2	12	0.8	70	1.5
b. Extreme scenario's								
2. Homogeneous	70	1.5	18	1.2	12	0.8	70	1.5
3. Clustered by wealth index	282	6.0	18	1.2	15	1.0	282	6.0
4. Clustered by contraception	70	1.5	188	12.5	12	0.8	188	4.0
5. Clustered by DPT3	70	1.5	18	1.2	94	6.3	94	2.0
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	94	2.0	18	1.2	12	0.8	94	2.0
7. Clustered by improved water	70	1.5	18	1.2	12	0.8	70	1.5
8. Clustered by adequate structure	94	2.0	24	1.6	18	1.2	94	2.0
9. Clustered by adequate space	70	1.5	18	1.2	12	0.8	70	1.5
10. Clustered by non-solid fuel	70	1.5	18	1.2	15	1.0	70	1.5
d. Moderate clustering								
11. 50% replaced from 3.	141	3.0	18	1.2	15	1.0	141	3.0
12. 50% replaced from 4.	70	1.5	47	3.1	12	0.8	70	1.5
13. 50% replaced from 5.	70	1.5	18	1.2	47	3.1	70	1.5

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.11: Required number of clusters - Baseline - Population 4

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	70	-	15	-	15	-	70	-
1. Baseline	70	1.0	18	1.2	15	1.0	70	1.0
b. Extreme scenario's								
2. Homogeneous	70	1.0	18	1.2	18	1.2	70	1.0
3. Clustered by wealth index	282	4.0	18	1.2	18	1.2	282	4.0
4. Clustered by contraception	70	1.0	188	12.5	18	1.2	188	2.7
5. Clustered by DPT3	70	1.0	18	1.2	94	6.3	94	1.3
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	94	1.3	18	1.2	18	1.2	94	1.3
7. Clustered by improved water	70	1.0	18	1.2	18	1.2	70	1.0
8. Clustered by adequate structure	94	1.3	18	1.2	18	1.2	94	1.3
9. Clustered by adequate space	70	1.0	18	1.2	12	0.8	70	1.0
10. Clustered by non-solid fuel	70	1.0	18	1.2	18	1.2	70	1.0
d. Moderate clustering								
11. 50% replaced from 3.	141	2.0	18	1.2	18	1.2	141	2.0
12. 50% replaced from 4.	70	1.0	70	4.7	18	1.2	70	1.0
13. 50% replaced from 5.	70	1.0	24	1.6	47	3.1	70	1.0

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.12: Required number of clusters - Baseline - Population 5

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	70	-	15	-	15	-	70	-
1. Baseline	70	1.0	18	1.2	18	1.2	70	1.0
b. Extreme scenario's								
2. Homogeneous	70	1.0	18	1.2	18	1.2	70	1.0
3. Clustered by wealth index	282	4.0	24	1.6	18	1.2	282	4.0
4. Clustered by contraception	70	1.0	188	12.5	18	1.2	188	2.7
5. Clustered by DPT3	70	1.0	24	1.6	118	7.9	118	1.7
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	70	1.0	18	1.2	12	0.8	70	1.0
7. Clustered by improved water	70	1.0	18	1.2	18	1.2	70	1.0
8. Clustered by adequate structure	94	1.3	18	1.2	15	1.0	94	1.3
9. Clustered by adequate space	70	1.0	18	1.2	18	1.2	70	1.0
10. Clustered by non-solid fuel	70	1.0	18	1.2	18	1.2	70	1.0
d. Moderate clustering								
11. 50% replaced from 3.	141	2.0	18	1.2	18	1.2	141	2.0
12. 50% replaced from 4.	70	1.0	47	3.1	18	1.2	70	1.0
13. 50% replaced from 5.	70	1.0	18	1.2	36	2.4	70	1.0

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Additional results 5.C.2

Table 5.13: Required number of clusters - Increasing cluster size - Population 1

	50 households per segment			75 households per segment		
		rel. diff			rel. diff	
Scenario	nr. clusters	clusters	households[a]	nr. clusters	clusters	households[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	4	0.3	0.9	-	-	-
1. Baseline	9	0.6	1.2	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	6	0.4	0.8	6	0.4	1.2
3. Clustered by wealth index	8	0.5	1.1	5	0.3	1.0
4. Clustered by contraception	6	0.4	0.8	6	0.4	1.2
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	8	0.5	1.1	6	0.4	1.2
7. Clustered by improved water	6	0.4	0.8	6	0.4	1.2
8. Clustered by adequate structure	6	0.4	0.8	6	0.4	1.2
9. Clustered by adequate space	6	0.4	0.8	6	0.4	1.2
10. Clustered by non-solid fuel	6	0.4	0.8	6	0.4	1.2
d. Moderate clustering						
11. 50% replaced from 3.	6	0.4	0.8	6	0.4	1.2
12. 50% replaced from 4.	6	0.4	0.8	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.14: Required number of clusters - Increasing cluster size - Population 2

	50 households per segment			75 households per segment		
	nr.	rel. diff		nr.	rel. diff	
Scenario	clusters	clusters	HHs[a]	clusters	clusters	HHs[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	5	0.3	1.1	-	-	-
1. Baseline	9	0.6	1.2	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	9	0.6	1.2	6	0.4	1.2
3. Clustered by wealth index	9	0.6	1.2	6	0.4	1.2
4. Clustered by contraception	6	0.4	0.8	6	0.4	1.2
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	9	0.6	1.2	6	0.4	1.2
7. Clustered by improved water	9	0.6	1.2	6	0.4	1.2
8. Clustered by adequate structure	8	0.5	1.1	6	0.4	1.2
9. Clustered by adequate space	8	0.5	1.1	6	0.4	1.2
10. Clustered by non-solid fuel	6	0.4	0.8	5	0.3	1.0
d. Moderate clustering						
11. 50% replaced from 3.	6	0.4	0.8	6	0.4	1.2
12. 50% replaced from 4.	8	0.5	1.1	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.15: Required number of clusters - Increasing cluster size - Population 3

	50 households per segment			75 households per segment		
	nr.	rel. diff		nr.	rel. diff	
Scenario	clusters	clusters	HHs[a]	clusters	clusters	HHs[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	6	0.4	1.3	-	-	-
1. Baseline	6	0.4	0.8	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	6	0.4	0.8	6	0.4	1.2
3. Clustered by wealth index	6	0.4	0.8	6	0.4	1.2
4. Clustered by contraception	9	0.6	1.2	6	0.4	1.2
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	6	0.4	0.8	6	0.4	1.2
7. Clustered by improved water	8	0.5	1.1	6	0.4	1.2
8. Clustered by adequate structure	8	0.5	1.1	6	0.4	1.2
9. Clustered by adequate space	6	0.4	0.8	5	0.3	1.0
10. Clustered by non-solid fuel	8	0.5	1.1	6	0.4	1.2
d. Moderate clustering						
11. 50% replaced from 3.	8	0.5	1.1	6	0.4	1.2
12. 50% replaced from 4.	6	0.4	0.8	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.16: Required number of clusters - Increasing cluster size - Population 4

	50 households per segment			75 households per segment		
	nr.	rel. diff		nr.	rel. diff	
Scenario	clusters	clusters	HHs[a]	clusters	clusters	HHs[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	6	0.4	1.3	-	-	-
1. Baseline	9	0.6	1.2	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	9	0.6	1.2	6	0.4	1.2
3. Clustered by wealth index	6	0.4	0.8	6	0.4	1.2
4. Clustered by contraception	9	0.6	1.2	6	0.4	1.2
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	9	0.6	1.2	6	0.4	1.2
7. Clustered by improved water	9	0.6	1.2	6	0.4	1.2
8. Clustered by adequate structure	9	0.6	1.2	6	0.4	1.2
9. Clustered by adequate space	9	0.6	1.2	6	0.4	1.2
10. Clustered by non-solid fuel	9	0.6	1.2	6	0.4	1.2
d. Moderate clustering						
11. 50% replaced from 3.	6	0.4	0.8	6	0.4	1.2
12. 50% replaced from 4.	6	0.4	0.8	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.17: Required number of clusters - Increasing cluster size - Population 5

	50 households per segment			75 households per segment		
	nr.	rel. diff		nr.	rel. diff	
Scenario	clusters	clusters	HHs[a]	clusters	clusters	HHs[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	6	0.4	1.3	-	-	-
1. Baseline	6	0.4	0.8	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	6	0.4	0.8	6	0.4	1.2
3. Clustered by wealth index	6	0.4	0.8	6	0.4	1.2
4. Clustered by contraception	7	0.5	0.9	5	0.3	1.0
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	6	0.4	0.8	6	0.4	1.2
7. Clustered by improved water	8	0.5	1.1	6	0.4	1.2
8. Clustered by adequate structure	6	0.4	0.8	6	0.4	1.2
9. Clustered by adequate space	6	0.4	0.8	6	0.4	1.2
10. Clustered by non-solid fuel	6	0.4	0.8	6	0.4	1.2
d. Moderate clustering						
11. 50% replaced from 3.	6	0.4	0.8	6	0.4	1.2
12. 50% replaced from 4.	6	0.4	0.8	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

5.D Additional statistics

Table 5.18: Observed ICCs in DHS Namibia 2013

	Poorest	Contraception	DPT3
Namibia	0.4142	0.0396	0.0885
Caprivi	0.4015	0.0008	0.0221
Erongo	0.0302	0.0380	0.0959
Hardap	0.2799	0.0194	0.0030
Karas	0.3624	0.0191	0.1197
Kavango	0.3842	0.0379	0.0724
Khomas	0.1324	0.0053	0.0486
Kunene	0.3179	0.0683	0.2407
Ohangwena	0.2694	0.0371	0.0677
Omaheke	0.1394	-0.0055	0.0504
Omusati	0.1487	0.0457	0.0116
Oshana	0.1772	0.0541	-0.0460
Oshikoto	0.2314	-0.0062	0.0237
Otjozondjupa	0.2132	-0.0043	0.1021