Practical Tools Sampling Project

Team Sarndal: Stacey Frank & Chendi Zhao

March 29, 2022

Contents

Introduction

This report will outline the process for sample design and selection for a sample of census tracts, block groups, and persons from Prince George's County, Maryland. This sample was designed to allow for estimates of the proportion of persons in different age groups who have civic awareness. Civic awareness will be measured in a survey by asking respondents questions about the name of their district representative in the U.S. House of Representatives, the name of their local delegate to the Maryland house of Delegates, and other indicators.

A three-stage cluster sample was drawn, with probability proportional to size (PPS) selection of 15 primary sampling units (PSUs), PPS selection of 1 secondary sampling unit (SSU) within each PSU, and a simple random sample (SRS) selection of elements within each SSU.

First, this report will explain the overall sample design and the method of assigning measure of size to PSUs and SSUs. Next, we will describe the method of sample selection and the units that were selected. Lastly, we will discuss the precision of estimates that can be anticipated from this sample, and the process for correctly measuring the variance of estimates in the achieved sample.

Sample Design

Target Population and Goal of Sample Design

The target population for this study is the adult (18+) non-institutionalized population of Prince George's County, Maryland. The sample frame is the United States 2010 decennial census.The population for this study's sampling frame includes approximately 657,421 persons.

The primary goal of this sample design is to allow the estimation of the proportion of the Prince George's County, Maryland population that has certain markers of civic awareness. The client desires to conduct this analysis within three age groups: people aged 18-44, people aged 45-64, and people aged 65 or over.

The desired total sample size is 300 persons which was split equally among the three age groups.To achieve the desired sample size - which reflects completed questionnaires - one needs to account for non-response. The response rates for the three age groups are anticipated to be 0.60, 0.70 and 0.85, respectively. After adjusting to account for the non-response, the new desired total sample size is 428, with 167, 153, and 118 persons in each age group. Thus, the new overall sampling rate f becomes 0.00065, calculated by 428/657421. The sampling rate for each age group, f_d can also be obtained using the same formula. The population, desired number of completed interviews, desired sample size, and sampling fraction per age group is listed in the table below.

Table 1: Desired Age Domain Sample Sizes

Age Group	Population	n	Expected Response Rate	Target Sample Size	Sampling Rate
18-44 years	350725	100	0.6	167	0.00048
45-64 years	225183	100	0.7	153	0.00064
65+ years	81513	100	0.8	118	0.00145
Total	657421	300		428	0.00065

Method of Selection

Given that the goal of this study is to measure civic awareness within these three age domains, a composite measure of size was used in sampling that accounted for the prevalence of persons within these age groups within each cluster. Using this method of selection should ensure that a targeted number of respondents per age group will be achieved in the final sample. Secondary goals of this sample are to achieve these domain sample sizes while also achieving a self-weighting sample within the three age groups and also creating an equal interviewer workload within each PSU. The equal workload for each tract can be calculated by $\bar{\bar{q}} = 428/(15 * 1) \approx 28.5333$.

As specified by the client, this sample design uses census tracts as PSUs, block groups as SSUs, and persons as elements. We will use the composite measure of size (MOS) method to meet the sampling goals. This method can also provide PSU selection probabilities that give "credit" for containing domains that are relatively rare in the population. To be specific, a three-stage cluster sample was drawn, with systematic sampling with probabilities proportional to size in PSUs and SSUs,and a simple random sample of persons within each block group.

The population data that was used for sample selection was pulled from the U.S. Census Bureau's website using the TidyCensus R package. The Census Bureau makes available summary-level tract and block-group data, which gives aggregate totals of the number of households and persons in each tract and block group, as well as a breakdown of the number of persons in each of the three age groups of interest. In total, there are 218 tracts and 523 block groups in the sampling frame. The map in Figure 1 shows all of the tracts (outlined in blue) and block groups (outlined in black) in Prince George's County, MD.

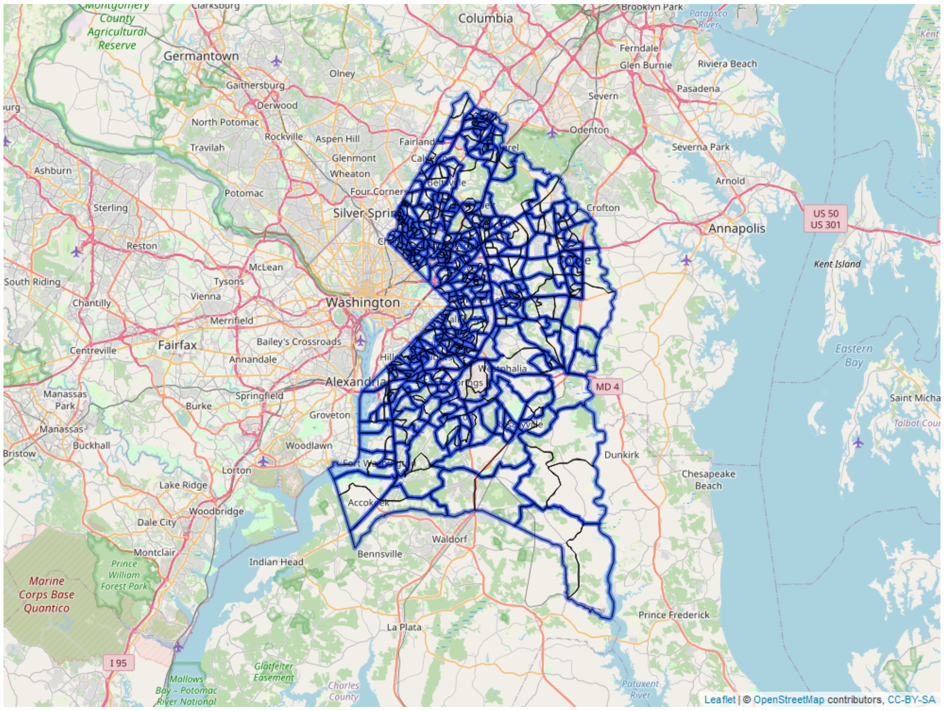


Figure 1: Map of Prince George's County, MD Tracts and Block Groups

Sample Selection

Composite Measure of Size and Selection Probability

The composite MOS for each PSU_{ij}, $S_i = \sum_{j \epsilon U_i} S_{ij} = \sum_d f_d Q_i(d)$, where S_{ij} is the composite MOS for SSU_j in PSU_i and Q_i is defined as number of elements in PSU i that are in domain d. Summing the S_i will give us the total composite MOS, which should be equal to the total desired sample size, 428.

Given that both PSUs and SSUs are sampled with probabilities proportional to the composite MOS, the selection probability of SSU_{ij} is defined as $\pi_i \pi_{k|ij} = mnS_{ij}/S$, where m is the number of sample PSUs and n is the number of sample SSUs in each PSU.Then, we are able to calculate the desired number to be selected from domain d in each SSU with $q^*_{ij}(d) = \bar{\bar{q}} f_d / S_{ij}$. It is worth to mentioning that $\bar{\bar{q}}$ is constant in each sampling stage.

The goal of this sample is to achieve an equal number of interviews within each of the three age groups of interest, meaning that the percentage of cases in each age group in the final sample should be about 33%. However, this does not match the distribution of age groups in the population. Among the adult population of Prince George's County, MD, about 53% are aged 18-44 and 12% are aged 65 or older. Only the population proportion of people aged 45-64 is approximately equivalent to the desired sample proportion for this age group. In essence, this means that people in the youngest age group need to be under-sampled, while people in the oldest age group need to be over-sampled.

Table 2: Age Distribution in Population and Sample

Age Group	Population Proportion	Desired Sample Proportion
18-44 years	0.533	0.333
45-64 years	0.343	0.333
65+ years	0.124	0.333

Using the composite MOS allows researchers a greater measure of control over the probable age distribution in the final sample by assigning larger selection probabilities to clusters that contain a disproportionate number of units that are members of a domain of interest. This means that in the current sample, tracts and block groups that contain a disproportionate number of people aged 65 or older are given a larger measure of size than their unadjusted population proportion would indicate.

Given the lack of balance in the age distribution of Prince George's County residents, a sample that was drawn with probabilities proportional to overall population size without accounting for age would be unlikely to produce an equal distribution of respondents across the three age groups in the final sample. The primary advantage of sample design with the composite measure of size is that it allows for self-weighting samples from each of the domains of interest. This means that variances of the final survey estimates will be smaller, because there will not be large differences in the sizes of weights across the sample, which would contribute to the variance of the estimates.

Quality Control Checks

After obtaining the information above, we did quality control checks to ensure that the desired sample size is possible for each SSU.The four criterion include:

(1). $q^*_{ij}(d) \leq Q_{ij}(d)$ for every SSU and domain,$q^*_{ij}(d)$ where is the expected number of sample persons in SSU_{ij} from domain d.

(2). $\bar{\bar{q}} \leq Q_{ij}$ for each SSU.

(3). $\bar{n}\bar{\bar{q}} \leq Q_i$ for each PSU.

(4). π_i,$\pi_{j|i}$,$\pi_{k|ij}$ less or equal to 1.

In the current sample frame, the seven block groups listed in the below table were detected to be undersized. Based on the map of Prince George's County, these unqualified areas include an air base, golf course, park land, and a university campus. Therefore, we combined them with the nearest block group within the tract to ensure each cluster met the minimum criteria for selection.

The first three block groups in Table 3 above are the only SSUs within that tract. After combing them, the new block group still had a desired sample size larger than the actual population in domain 3. We decided to keep the new group in the frame, since there were no other block groups within the tract that it could be combined with. If this new block group is sampled, we would sample more persons in domain 3 in the next sampled block group to achieve the expected sample size. Block Group 240338024082 and 240338035192 were combined with 240338024082 and 240338035191, respectively. Block Group 240338072002 and 240338072003 are a university campus so there are mainly young adults living there. If we combine them together, there will still be insufficient sample for domain 2 and 3. Also, the population will be very disproportionately contributed across the domains. Therefore, we combined 240338072002 with 240338072001 and 240338072003 with 240338072004.

Table 3: Unqualified Block Groups

NO.	Block Group	Total Units	Domain 1	Domain 2	Domain 3
1	240338011041	0	0	0	0
2	240338011042	0	0	0	0
3	240338011043	2973	1734	183	8
4	240338024082	8	5	2	0
5	240338035192	55	24	2	1
6	240338072002	5219	5200	0	0
7	240338072003	6585	6551	9	9

Selected Units and Their Characteristics

The sampled block groups are listed in the below table with the information for households and overall population in each domain. We noticed that the workloads are not integers, which means that when the samples of persons within a sample block groups are selected, the sampling needs to be done using fixed rates not fixed sample sizes.

Table 4: Sample Result

NO.	Selected Block Group	Total Units	Total Households	Domain 1	Domain 2	Domain 3	Workload
1	Block Group 1, Census Tract 8001.06	1294	613	581	340	78	28.53
2	Block Group 1, Census Tract 8004.03	2662	930	824	810	336	28.53
3	Block Group 1, Census Tract 8005.11	1629	590	625	440	139	28.53
4	Block Group 1, Census Tract 8007.01	3434	1232	1308	1003	227	28.53
5	Block Group 2, Census Tract 8012.10	1999	725	658	641	240	28.53
6	Block Group 1, Census Tract 8013.11	2104	741	643	725	245	28.53
7	Block Group 2, Census Tract 8017.02	2867	1403	1407	571	79	28.53
8	Block Group 1, Census Tract 8019.08	1883	797	828	448	115	28.53
9	Block Group 2, Census Tract 8025.01	1628	739	674	426	118	28.53
10	Block Group 1, Census Tract 8035.09	2011	661	923	285	34	28.53
11	Block Group 2, Census Tract 8036.02	824	296	282	173	142	28.53

NO.	Selected Block Group	Total Units	Total Households	Domain 1	Domain 2	Domain 3	Workload
12	Block Group 1, Census Tract 8041.02	1754	587	669	436	156	28.53
13	Block Group 1, Census Tract 8056.02	3643	952	2652	341	23	28.53
14	Block Group 1, Census Tract 8066.02	2463	821	1161	474	116	28.53
15	Block Group 1, Census Tract 8072	8101	753	7940	81	28	28.53

A map of sampled block groups is shown in Figure 2, with the selected block groups displayed in red.

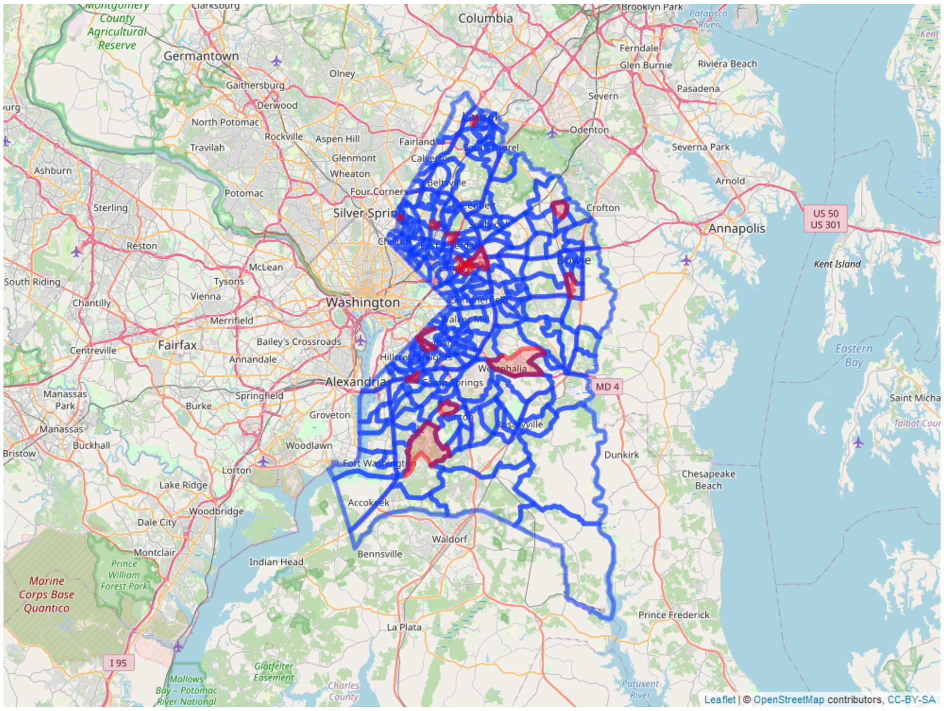


Figure 2: Map of Sampled Block Groups in Prince George's County, MD

Selection Probabilities of Units

Table 5: Selection Probability and Weights

	Min	1st Quantile	Median	Mean	3st Quantile	Max
pi_i	0.0235412	0.0529464	0.0658597	0.0688073	0.0818276	0.2678272
pi_ij	0.0134199	0.0209762	0.0270778	0.0346544	0.0366160	0.1357239
1/pi_i	3.7337504	12.2210103	15.1841080	16.4287669	18.8870896	42.4786709
1/pi_ij	74.5164532	47.6730626	36.9305666	28.8563399	27.3104968	7.3679007

Element Level Selection of Persons

Precision and Variance Estimation

Anticipated Precision

The fact that only 1 BG is selected per tract might raise the question of whether variances can be estimated with this design. We can still estimate design-variances because the number of first-stage units is 15, the number of sample tracts. See Textbook 9.2.1

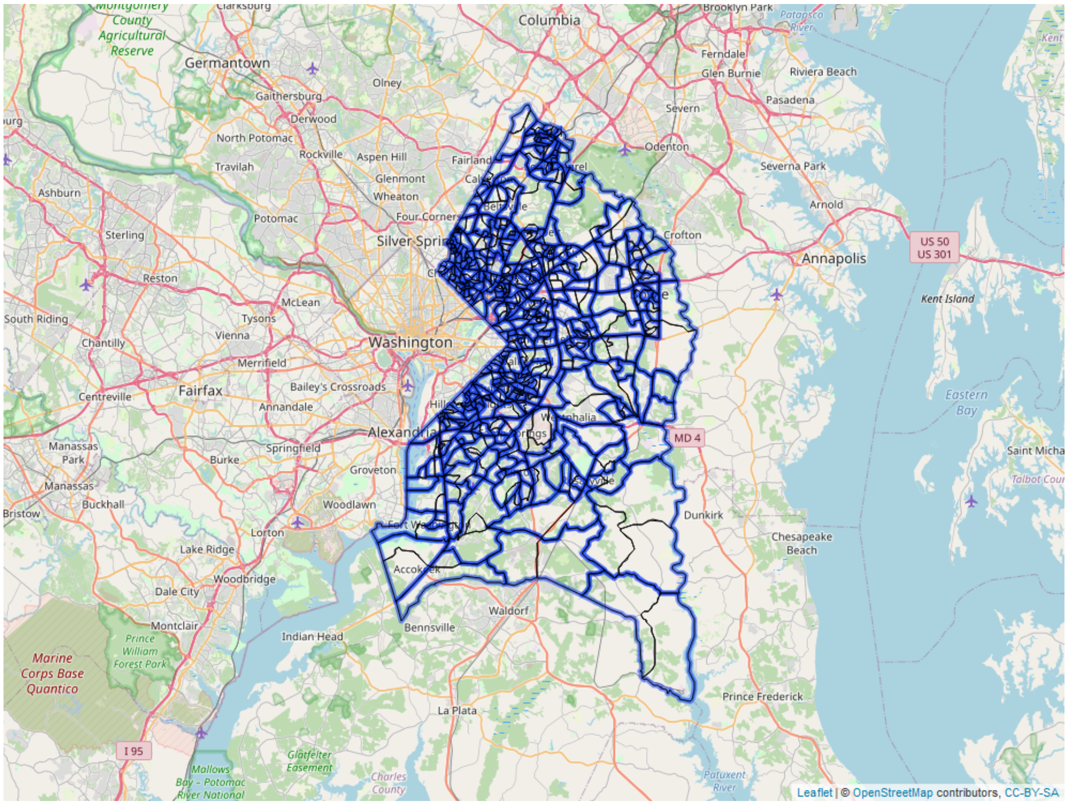
Variance Estimation

Conclusion

This report outlines our proposed sample design and sample draw of of census tracts, block groups, and persons from Prince George's County, Maryland for a survey of civic awareness among Prince George's County adults. This sample was designed to allow for estimates of the proportion of persons in different age groups who have civic awareness.A three-stage cluster sample was drawn, with probability proportional to size selection of 15 tracts, probability proportional to size selection of 1 block group within each tract, and a simple random sample of persons within each SSU. The goal of this study is to measure civic awareness within three age domains, so a composite measure of size for tracts and block groups was used in sampling to accounted for the prevalence of persons in each age groups within each cluster.Using this method of selection should ensure that the target number of interviews, 100 within each age group, is achieved achieved in the final sample. This sample was also designed to be self-weighting within the three age groups and also to have an equal interviewer workload within each PSU. Anticipated precision and a suggested approach to variance calculation for the final sample ar also discussed.

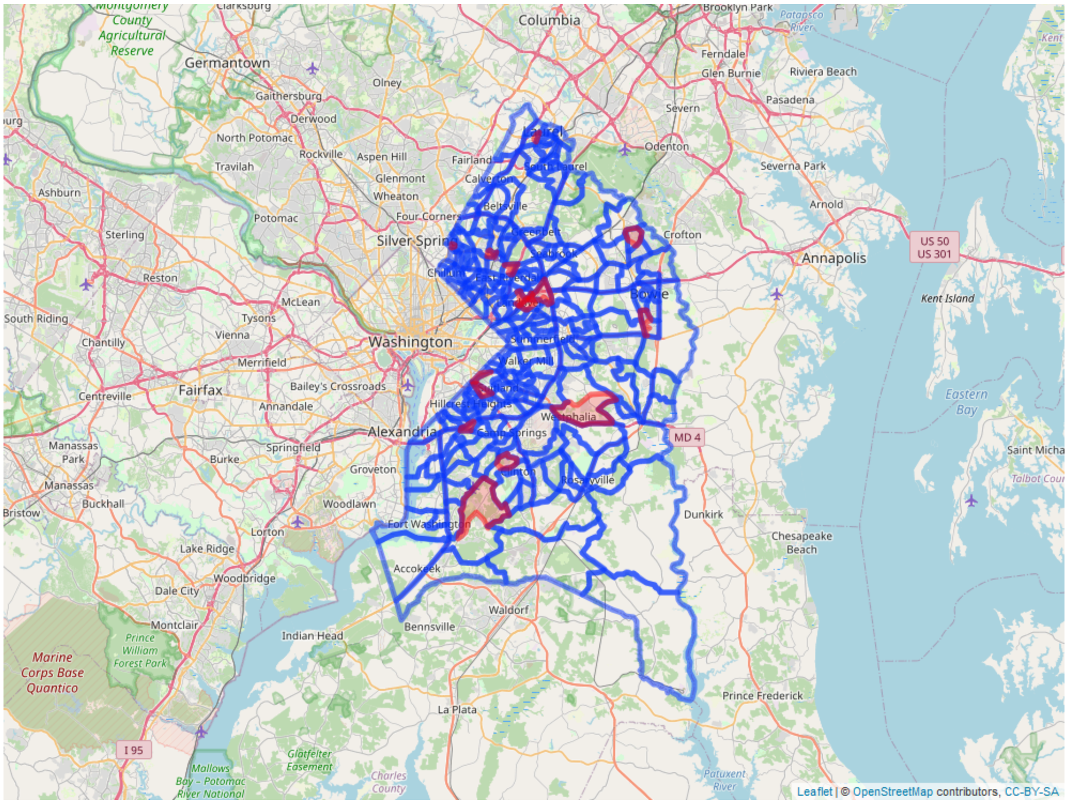
Maps

Map 1: Prince George’s County, MD Tracts and Block Groups



Leaflet | © OpenStreetMap contributors, CC-BY-SA

Map 2: Sampled Block Groups in Prince George's County, MD



Appendix

Codebook for Sample Frame and Sample File

Chendi will add:

- Codebook of frame and sample files, i.e. a list of the variables on the text files and a description of each variable

Sample Listing with Selection Probabilities

Chendi will add:

- Listing of the sample PSUs and sample SSUs with their selection probabilities and census data. On each sample SSU, list the sampling rate you will use to select persons in each domain.

In all we need these deliverables:

The deliverables for the project will be

1. A sampling report (details of the report below) [filename: REPORT_GROUPNAME.pdf];
2. Text files giving the units used for the area frame and relevant census counts and measures of size [filename: FRAME#_GROUPNAME.pdf]; and
3. Text file for the selected sample along with relevant census counts, measures of size, selection probabilities, and weights [filename: SAMPLE_GROUPNAME.pdf]