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.

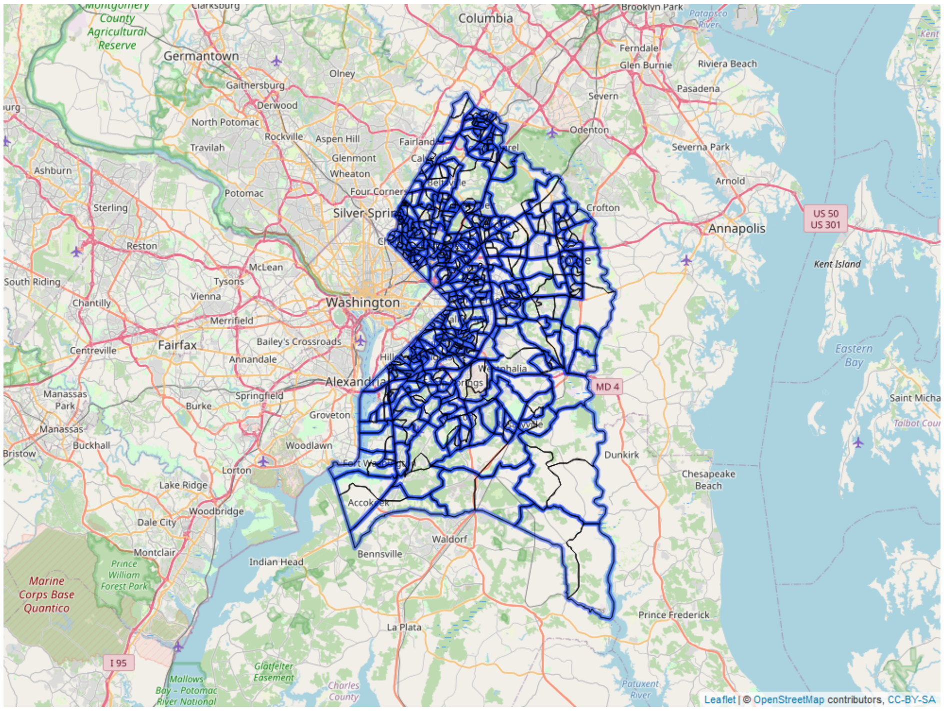


Figure 1: Map of Prince George's County, MD Tracts and Block Groups

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 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.

Take the first sampled block group as an example–there are 581 people in age group 18-24 and the sample size for this group is expected to be 13. In this case, persons in that domain would be sampled at the rate 13/581= 0.0224.

Table 4: Sample Result

NO.	Selected Block Group	Total Units	Total Households	Domain 1	Domain 2	Domain 3	Workload
1	BG 1, Tract 8001.06	1294	613	13.0	10.2	5.3	28.53
2	BG 1, Tract 8004.03	2662	930	8.0	10.5	10.0	28.53
3	BG 1, Tract 8005.11	1629	590	10.9	10.2	7.4	28.53
4	BG 1, Tract 8007.01	3434	1232	11.2	11.4	5.9	28.53
5	BG 2, Tract 8012.10	1999	725	8.4	10.9	9.3	28.53
6	BG 1, Tract 8013.11	2104	741	7.8	11.7	9.0	28.53
7	BG 2, Tract 8017.02	2867	1403	16.7	9.0	2.8	28.53
8	BG 1, Tract 8019.08	1883	797	13.3	9.6	5.6	28.53
9	BG 2, Tract 8025.01	1628	739	12.0	10.1	6.4	28.53
10	BG 1, Tract 8035.09	2011	661	18.7	7.7	2.1	28.53
11	BG 2, Tract 8036.02	824	296	8.5	7.0	13.0	28.53
12	BG 1, Tract 8041.02	1754	587	11.1	9.6	7.8	28.53
13	BG 1, Tract 8056.02	3643	952	23.8	4.1	0.6	28.53
14	BG 1, Tract 8066.02	2463	821	15.4	8.4	4.7	28.53
15	BG 1, Tract 8072	8101	753	27.9	0.4	0.3	28.53

A map of sampled block groups is shown in Figure 2, with the 15 selected block groups displayed in red.

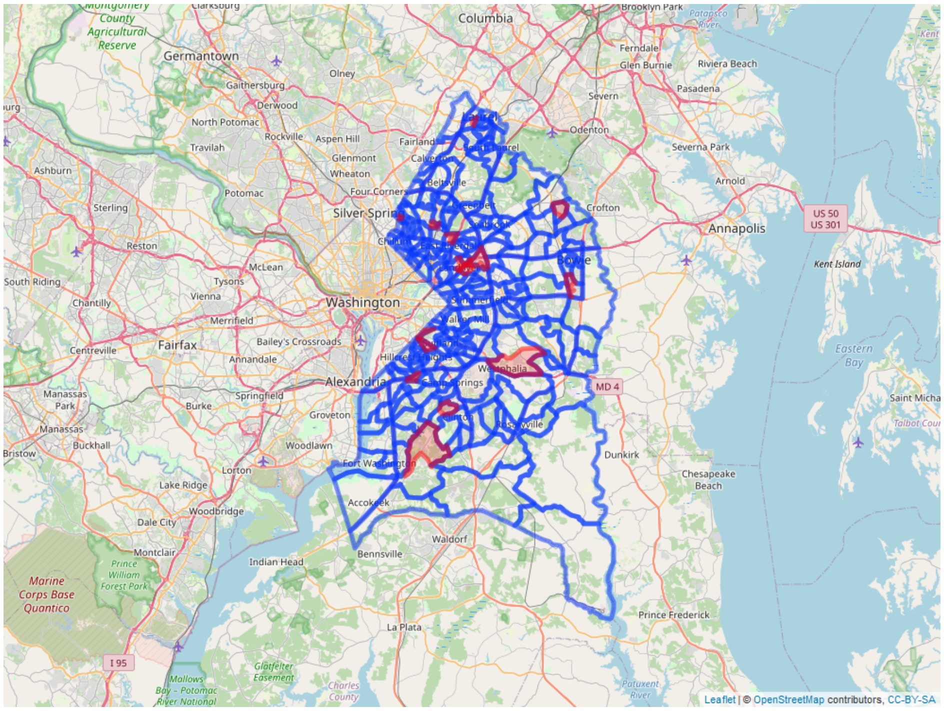


Figure 2: Map of Sampled Block Groups in Prince George's County, MD

Selection Probabilities of Units

After combining the undersized block groups, the selection probabilities for tracts and block groups are summarized in the table below.The range of selection probabilities for tracts is 0.0235 to 0.2678, while the range of weights for tracts is 3.7338 to 42.4787. The range of selection probabilities for block groups is 0.0134 to 0.1357, while the range of weights for block groups is 7.3679 to 74.165. Since we are excluding any tracts or block groups that are out of the frame, a self-weighting sample can still be achieved.

Table 5: Selection Probability and Weights

	Min	1st Quantile	Median	Mean	3rd Quantile	Max
pi_i	0.0235	0.0529	0.0659	0.0688	0.0818	0.2678
pi_ij	0.0134	0.0210	0.0271	0.0347	0.0366	0.1357
1/pi_i	3.7338	12.2210	15.1841	16.4288	18.8871	42.4787
1/pi_ij	7.3679	27.3238	36.9306	37.6532	47.6795	74.5165

Element Level Selection of Persons

This sample design calls for persons to be directly sampled using simple random sampling from within block groups at the third stage of selection. This assumes that there are block group level rosters of adult residents from which a sample of persons can be drawn. The sampling rate will differ within each selected SSU because the population size of each SSU is different. The element level selection probability within each block group was calculated by dividing q, the desired workload in each block group, by the total number of adults in each block group. Note that the total number of adults in each block group was used, rather than the total number of persons, due to the fact that individuals under the age of 18 are ineligible for this survey.

The sampling rate for each block group is calculated by taking the inverse of the element level probability of selection for each block group. The total number of adults, element level probability of selection, and element level sampling rate are shown below for each of the 15 selected block groups.

Table 6: Person Level Selection Probabilities and Sampling Rates

NO.	Selected Block Group	Total Adults	Element Selection Probability	Element Sampling Rate
1	BG 1, Tract 8001.06	999	0.029	35.012
2	BG 1, Tract 8004.03	1970	0.014	69.042
3	BG 1, Tract 8005.11	1204	0.024	42.196
4	BG 1, Tract 8007.01	2538	0.011	88.949
5	BG 2, Tract 8012.10	1539	0.019	53.937
6	BG 1, Tract 8013.11	1613	0.018	56.530
7	BG 2, Tract 8017.02	2057	0.014	72.091
8	BG 1, Tract 8019.08	1391	0.021	48.750
9	BG 2, Tract 8025.01	1218	0.023	42.687
10	BG 1, Tract 8035.09	1242	0.023	43.528
11	BG 2, Tract 8036.02	597	0.048	20.923
12	BG 1, Tract 8041.02	1261	0.023	44.194
13	BG 1, Tract 8056.02	3016	0.009	105.701
14	BG 1, Tract 8066.02	1751	0.016	61.367
15	BG 1, Tract 8072	8049	0.004	282.091

Block group 1 in tract 8072 has the largest sampling rate because it has the largest adult population of all the sampled block groups, while block group 2 in tract 8036.02 has the smallest sampling rate due to its small adult population. Note that the element level sampling rates are not round numbers. This means that a fractional interval will need to be utilized during the element level sample selection to ensure that the correct sampling rate is used in each block group and targeted number of respondents per block group is achieved. Since the targeted workload per SSU is also not a round number ($\bar{\bar{q}} \approx 28.5333$), some SSUs will have 28 respondents, while others will have 29 respondents.

If a roster of adults who live within each selected block group is not available, this design would have to be modified to include a fourth level of selection. This would involve selecting households within block group at the third level, using either a preexisting household listing, or having survey staff create one for each of the selected block groups. Once households were selected, one adult would be randomly selected from among the adult members of the selected household for the fourth level of selection.

Precision and Variance Estimation

Anticipated Precision

It is possible to calculate the anticipated precision of the estimates that will be made with this sample by creating element level dummy data. This dummy data can then be analyzed using the BW2stagePPSe function available in the PracTools R package to calculate the variances from each of the stages of sample selection.

Dummy data was created by expanding the data frame of selected SSUs to include a row for each element (person) that will ultimately be selected. The expanded element level data file had 428 total observations, with 28 or 29 elements in each selected block group. After creating the element level file, dummy analysis variables were created for the anticipated precision analysis. The dummy analysis varaibles were in the form of a binary response, with 1 indicating that the respondent correctly answered a question related to Maryland civic awareness, and 0 indicating the question was incorrectly answered.

Two synthetic dummy analysis variables were created. The first dummy variable had approximately 50% of cases responding to the civic awareness question correctly, while the second dummy variable had about 5% of cases answering the civic awareness question correctly. The dummy data was assigned at these rates within SSUs, rather than randomly assigned throughout the sample. This means that in any given SSU, about 50% of cases answered correctly for the first dummy variable, and 5% of cases answered correctly for the second. Calculated anticipated precision for both of these variables will give a good idea of the possible range of variances in the final achieved sample, since we do not currently have projections of the proportion of the Prince George's County population that will be able to answer civic awareness questions correctly.

Since this sample design includes 15 first-stage clusters, but only one second-stage cluster within each first stage cluster, variances cannot be computed for the first and second stage clusters. Therefore, for the purposes of variance estimation, we will treat this as a two-stage sample, with the first stage being a PPS selection of 15 block groups and the second stage being a simple random selection of persons within block groups. Therefore, we used the BW2stagePPSe function, which assumes a PPS selection at the first stage and an SRS selection at the second stage.

The inputs for the BW2stagePPSe function are:

1. Ni: the total number of adults within each selected block group
2. ni: the total number of elements (persons) sampled within each selected block group
3. X: the vector of data that should be analyzed. The BW2stagePPSe function was run twice on the two synthetic dummy variables that were created.
4. psuID: the block group ID
5. w: the overall sample weight. This was calculated by multiplying the inverse of the 1st and 2nd stage selection probabilities (for tract and block group from the original 3 stage design) with the inverse of the element level selection probability for each SSU (displayed above in Table 6).
6. m: the number of sampled PSUs, which is 15
7. pp: a vector of PSU selection probabilities. These were calculated by multiplying the 1st and 2nd stage (tract and block group) selection probabilities from the original 3-stage design.

The results from the BW2stagePPSe function show that the variance of estimates are expected to be smaller in the case of low levels of civic awareness in the sample, while variance will be larger if levels of civic awareness are closer to 50% in the achieved sample.

Table 7: Anticipated Precision for Proposed Sample Design

Variable	PSU Variance	SSU Variance	B^2	W^2	k	Delta
Est. 50% Correct Responses	134002617	1322375	0.017	0.005	0.021	0.777
Est. 5% Correct Responses	2417121	344614	0.015	0.062	0.006	0.194

Variance Estimation

In order to conduct formal variance calculations on the achieved sample estimates, we recommend taking the approach outlined above for calculating the anticipated precision. The variance of estimates can be calculated assuming a two stage sample, with PPS selection of block groups and simple random selection of persons.

According to Applied Survey Data Analysis by Heeringa, West and Berglund, estimates of population proportions can be calculated using a ratio mean estimator of the prevalence in the population π with this form:

$p = \frac{\sum_{a=1}^{a}\sum_{i=1}^{n} W_{ai} I(y_i=1)}{\sum_{a=1}^{a}\sum_{i=1}^{n} W_{ai}} = \frac{\hat{N}_1}{\hat{N}}$. This is a non-linear estimator when calculated from a complex survey design, and therefore variances need to be calculated using an estimation method such as the Taylor Series Linearization. The Taylor Series Linearization variance estimator for a ratio estimate of a proportion is $v(p) = \frac{V(\hat{N}_1)+p^2*V(\hat{N})-2*p*cov(\hat{N}_1,\hat{N})}{\hat{N}^2}$. This variance estimator can be calculated in R using the svy package, which allows you to set the design and weighting variables that should be accounted for when doing variance estimation for complex samples.

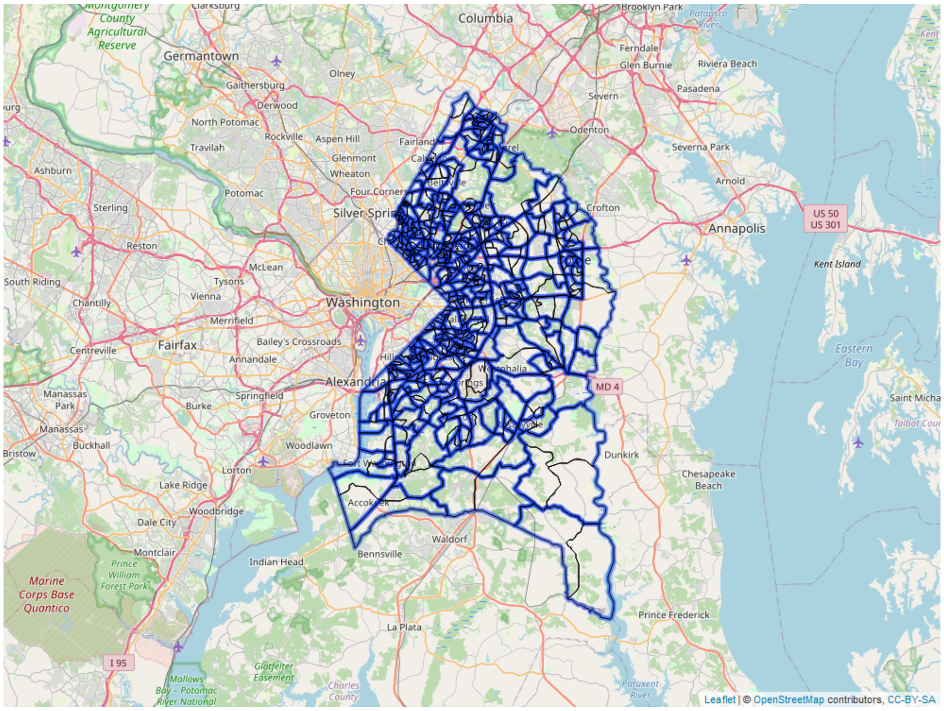
Conclusion

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.

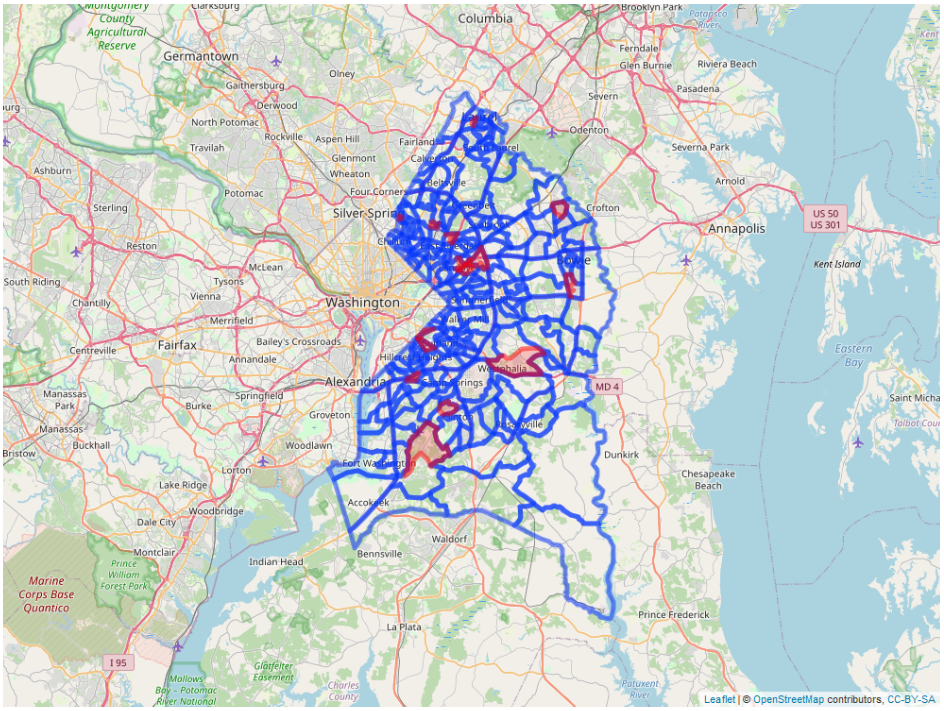
The goal of this study is to measure civic awareness within three age domains, so a composite measure of size for tracts and block groups was used in sampling to accounted for the prevalence of persons in each age groups within each cluster.Using this method of selection should ensure that the target 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



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

Table 8: Sampled PSUs with Selection Probabilities

Selected Tract	Total Population	Total Adults	Total Aged 18-44	Total Aged 45-64	Total Aged 65+	Composite MOS	1st Stage Selection Probability
Census Tract 8001.06	2651	2028	1115	751	162	1.242	0.044
Census Tract 8004.03	3683	2790	1125	1111	554	2.043	0.072
Census Tract 8005.11	5111	3850	1822	1454	574	2.622	0.092
Census Tract 8007.01	5394	3990	1958	1590	442	2.582	0.090
Census Tract 8012.10	4103	3216	1337	1396	483	2.222	0.078

Selected Tract	Total Population	Total Adults	Total Aged 18-44	Total Aged 45-64	Total Aged 65+	Composite MOS	1st Stage Selection Probability
Census Tract 8013.11	5721	4296	1947	1840	509	2.832	0.099
Census Tract 8017.02	3737	2724	1733	774	217	1.631	0.057
Census Tract 8019.08	3363	2491	1499	812	180	1.490	0.052
Census Tract 8025.01	3167	2436	1290	891	255	1.549	0.054
Census Tract 8035.09	3079	1996	1386	482	128	1.151	0.040
Census Tract 8036.02	1866	1341	659	392	290	0.983	0.034
Census Tract 8041.02	5974	4213	2487	1378	348	2.563	0.090
Census Tract 8056.02	4918	3963	3325	551	87	2.059	0.072
Census Tract 8066.02	4578	3318	2104	939	275	1.996	0.070
Census Tract 8072	15934	15803	15525	188	90	7.642	0.268

Table 9: Sampled SSUs with Selection Probabilities

Selected BG	Total Population	Total Adults	Total Aged 18-44	Total Aged 45-64	Total Aged 65+	Composite MOS	2nd Stage Selection Probability	Element Sampling Rate
BG 1, Tract 8001.06	1294	999	581	340	78	0.605	0.487	35.012
BG 1, Tract 8004.03	2662	1970	824	810	336	1.393	0.682	69.042
BG 1, Tract 8005.11	1629	1204	625	440	139	0.778	0.297	42.196
BG 1, Tract 8007.01	3434	2538	1308	1003	227	1.588	0.615	88.949
BG 2, Tract 8012.10	1999	1539	658	641	240	1.068	0.480	53.937
BG 1, Tract 8013.11	2104	1613	643	725	245	1.121	0.396	56.530
BG 2, Tract 8017.02	2867	2057	1407	571	79	1.147	0.703	72.091

Selected BG	Total Population	Total Adults	Total Aged 18-44	Total Aged 45-64	Total Aged 65+	Composite MOS	2nd Stage Selection Probability	Element Sampling Rate
BG 1, Tract 8019.08	1883	1391	828	448	115	0.845	0.567	48.750
BG 2, Tract 8025.01	1628	1218	674	426	118	0.762	0.492	42.687
BG 1, Tract 8035.09	2011	1242	923	285	34	0.670	0.582	43.528
BG 2, Tract 8036.02	824	597	282	173	142	0.450	0.458	20.923
BG 1, Tract 8041.02	1754	1261	669	436	156	0.821	0.320	44.194
BG 1, Tract 8056.02	3643	3016	2652	341	23	1.513	0.735	105.701
BG 1, Tract 8066.02	2463	1751	1161	474	116	1.022	0.512	61.367
BG 1, Tract 8072	8101	8049	7940	81	28	3.873	0.507	282.091