

South Dakota State University
**Open PRAIRIE: Open Public Research Access Institutional
Repository and Information Exchange**

Electronic Theses and Dissertations

2018

Extending the Utility of Public Use Microdata

Eric A. Guthrie
South Dakota State University

Follow this and additional works at: <https://openprairie.sdstate.edu/etd>

 Part of the [Demography, Population, and Ecology Commons](#)

Recommended Citation

Guthrie, Eric A., "Extending the Utility of Public Use Microdata" (2018). *Electronic Theses and Dissertations*. 2439.
<https://openprairie.sdstate.edu/etd/2439>

This Dissertation - Open Access is brought to you for free and open access by Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. For more information, please contact michael.biondo@sdstate.edu.

EXTENDING THE UTILITY OF PUBLIC USE MICRODATA

BY

ERIC A. GUTHRIE

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Sociology

South Dakota State University

2018

EXTENDING THE UTILITY OF PUBLIC USE MICRODATA

ERIC A. GUTHRIE

This dissertation is approved as a credible and independent investigation by a candidate for the Doctor of Philosophy in Sociology degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

~~Weiwei~~ Zhang, Ph.D.
Dissertation Advisor

Date

Mary Emery, Ph.D.
Head, Department of ~~Socology~~

Date

~~Dean~~, Graduate School

Date

ACKNOWLEDGMENTS

I could not have finished this course of study without the help of many people, and I must acknowledge that I needed and received the help of every person mentioned here. This dissertation would not have been possible without the support and encouragement of my partner, Kristopher, thank you for your unwavering belief in me. My committee, Drs. Zhang, Ulrich-Schad, Yingling, and Gonda provided help and encouragement without which I could not have finished my studies; I am truly grateful. I would be remiss without thanking Drs. Kayongo-Male, Arwood, and Froelich for the innumerable lessons and alterations to my worldview their teaching brought about. Finally, I must thank Dr. Michael McCurry for the years of mentorship and genuine friendship that we shared at SDSU.

CONTENTS

Abstract	vi
Introduction.....	1
Significance.....	5
The Logic of the Method	9
The Original Method.....	9
The Revised Method.....	18
Determining a Weighting Variable.....	28
The Method in Brief.....	33
Testing the Method	37
Usability Testing.....	38
Wide Applicability.....	38
Replicability Testing.....	43
Validity Testing	46
Summary Table of Validity Testing Results.....	54
Discussion.....	55
Conclusions.....	63
Appendix A – Estimates Produced for Usability Testing.....	67
New Jersey Population Between 5 and 17 at or Below 200 Percent of Poverty by School District	67
South Dakota Males Between 26 and 32 Who are Employed by County	79
Appendix B – Estimates Used in Validity Testing.....	81
Appendix C – The Revised Method in Detail.....	82
Reviewing the Geographies	82
Determine the Requirements.....	86
Determine a Weighting Variable	87
Obtain or produce the necessary shapefiles and prepare them for use	91
Combine the Shapefiles	111
Add the weighting variable.....	116
Software Choices (QGIS v. ArcGIS).....	125
Isolating the PUMAs and creating the weights.....	127
Making the Estimates.....	128
Review the Estimates.....	130

Alternate Implementation	132
Brief Steps for an Alternative Implementation	133
Works Cited	135

ABSTRACT

EXTENDING THE UTILITY OF PUBLIC USE MICRODATA

ERIC A. GUTHRIE

2018

Applied demography employs population studies in the effort to answer real world questions and provide insights for the problems that business and civic leaders face on an ongoing basis. To answer these questions the applied demographer sometimes performs primary research, but more often they attempt to leverage and extend the use of publicly available data to answer the questions presented in an efficient and time-constrained manner. The work described here looks at a problem presented by the Michigan Department of Education and the solution presented by the Michigan State Demographer. The problem required estimates for a single-year age group at a non-standard poverty level. These data are not published by the U.S. Census Bureau, but a novel solution was developed to serve an intermediate need until a custom tabulation of Census data could be delivered. With the delivery of a custom tabulation of Census data, there was a unique opportunity to test the results of the interpolation against what would be a gold standard dataset. The results reveal that the process of interpolating estimates devised as a solution could produce estimates that could be useful for a variety of purposes.

Introduction

The work of an applied demographer often involves the production of estimates for very specific groups of people (Morrison and Judson 2011). Many times, data for these groups are not easily obtainable due to their size or geographic dispersion. When presented with a request seeking data on such populations, the person receiving the request has a few options. Generally, these options involve rejecting the request, reframing the question, or developing a novel approach for making the estimates (Merrick 1986; Swanson, Burch, and Tedrow 1996; Son et al. 2012). Conducting primary research into the topic is always possible, but given the cost prohibitive nature of such research for small population subgroups, it is rarely considered an option in the world of an applied demographer, especially one sponsored or employed by a governmental agency (Swanson, Burch, and Tedrow 1996).

Rejecting a request and reframing a research question are two sides of the same coin. In both instances, the applied demographer is indicating the research question as presented is unanswerable without costly data collection. Most things are knowable with enough resources, but in many cases, the reason the question is directed to the applied demographer is due to a resource limitation, which makes primary research an impractical solution (Murdock and Ellis 1991). When rejecting a request, several considerations come into play such as the purpose of the request, the nature of the requester, the current state of research into the topic, and the time required to produce an estimate reasonably expected to have a degree of validity (Swanson, Burch, and Tedrow 1996). The reality is that applied demographers are sought to answer specific questions in a time bound and resource constrained manner (Swanson 2008). Given that, they tend

to give priority to the goals and concerns of the group or body that employs them. This means if they are employed by a non-profit, they will seek to engage the priorities of that group, and if they are employed by a governmental agency, they will likewise seek to engage the priorities of that group (Murdock and Ellis 1991). Being employed by a governmental agency has both positive and negative consequences. On the positive side, government employees are (at least theoretically) employed by all the citizens under the jurisdiction of the governmental body. This allows the applied demographer wider latitude in determining whether a particular request falls within the purview of his or her office. The negative side of this is political considerations can and do come into play when determining who or what can be a subject of investigation (Horton 1999). This means questions requiring significant time and subjectivity may need to be rejected if they carry significant political baggage. This is a relatively rare occurrence, and most often, the researcher tries to help the customer reframe their question into something that can be answered using publicly available data.

When helping a data user reframe a question, the usual alterations come in the form of adjusting the target population, geography, or timeframe (U.S. Census Bureau 2009). Adjusting the target population is sometimes hard for the data user to accept, especially if they have a programmatic reason for the data request. As an example, it is not the same question to ask the population size of a segment of African-American children versus the size of the same segment of “minority” children, but that is a transformation that may make a question answerable. A geographic adjustment, in nearly all cases, involves increasing the size of the area of concern. Take for example the case where a data user is looking for the characteristics of a segment of the population of a

city. Those data may not be available at the city level, but data may be available at the county level. This is very common when talking about the structure of the population as age and sex detail is not produced by the U.S. Census Bureau's Population Estimates Program below the county level on an annual basis. Similar to adjusting the geographic area of concern, adjusting the timeframe of a data request usually involves increasing the timeframe relevant for a request. This is also common when investigating relatively small geographies or low-population geographies. Increasing the timeframe allows flexibility to use pooled estimates from the U.S. Census Bureau's American Community Survey (ACS) 5-year Estimates program. Being able to use the 5-year estimates for a request will allow for a greater number of geographies to be included in a request because of limits on availability between the 1-year and 5-year Estimates. The 1-year Estimates only include data for geographies with a population of greater than 65,000 people. The 5-year Estimates include estimates for all geographies, regardless of the population of the area. As an example, the level of educational attainment for a geography may not be available for 2015, but a pooled, period estimate for 2011-2015 would be available and would get the data user something with which to work. Irrespective of how the applied demographer suggests adjusting the research question, the goal is always (excluding when the data user's question has a systemic flaw) to get the data user information that is as close to the original request as possible. This means the applied demographer is trying to maximize the target population within the geographic and timeframe constraints while minimizing the additional populations, geographies, and time necessary to make an estimate. This means that the applied demographer is attempting to zoom in on the target

population on three aspects, time, space (or geography), and with precision regarding the individuals of interest.

When the data user's question cannot be adjusted, and the user or question is important, the applied demographer must develop a novel solution (Son et al 2012). Such a question was recently presented to me by the Michigan Department of Education when they were seeking help in reworking a legislative funding formula for a pre-K reading program. This program is statewide and consumes hundreds of millions of dollars in educational funding, so the user and the question were of sufficient importance. The group needed to have estimates for the number of four-year-olds who live in families with incomes less than 250 percent of poverty by intermediate school district (ISD). This is not a statistic published by the U.S. Census Bureau in their summary data and there was not sufficient geographic granularity in the publicly accessible microdata to replicate that geography. Given the problems with the request, my first suggestion to the group was to ask the U.S. Census Bureau for a custom tabulation from their American Community Survey (ACS). The group was receptive to the suggestion of purchasing the special tabulation but wanted data prior to the time the Census Bureau could deliver the data. Additionally, the group wanted several iterations of the data to decide on how the final request to the Census Bureau would be submitted. This meant I had to come up with a method for making the estimates. I developed a process using the summary data to make weights for the public use microdata sample (PUMS) data provided by Census to supplement the summary data. This allowed me to transform the PUMS geography into the ISD geography. My results were not perfect, and the group understood they would not be, but they provided early estimates that allowed the group to continue working

while it determined exactly what data to request from Census and to investigate aspects of the formula while we waited for the data to be produced.

There are two major questions this project will seek to answer. My first question: can a methodology be designed to create estimates for alternate geographies and subpopulations using only publicly available data that give a researcher a reasonable approximation of what a special Census Bureau tabulation would provide? I am planning to investigate a multi-step weighting process that should align my estimates more closely with the ones produced by Census through custom tabulations. This type of process will never be perfect, but if it can be used to produce estimates that reasonably correlate with those produced by Census's custom tabulation program, it will be a big help to researchers and those in the advocacy community. My second question: can the methodology be reasonably simple, so that someone with a moderate amount of training and knowledge about the data will be able to easily replicate it and produce estimates for use in their projects? My goal is not to supplant the custom tabulation program, but rather to supplement it with a method that people can use to perform tasks such as creating preliminary data to use for project design. Aside from being time consuming, the process of requesting custom tabulations is also expensive, requiring a minimum of a \$3,000 investment. This process should also ease the burden on researchers who might otherwise need to make multiple requests or ones larger than they actually need.

Significance

The field of applied demography exists as a subfield of demography devoted to answering real world questions in an effort to foster better decision making and to bring

data into play for that purpose. It is distinct from its parent field of formal demography in a few ways. Primarily, it is distinct in its objectives and the manner in which it selects topics of inquiry. Murdock and Ellis (1991) describe the goals of applied demography to be concerned with prediction rather than the explicative goals of formal demography and to further orient the field in its concern for the future over the past. This is a function of its situation as a decision-making science intended to serve the needs of its clients which are generally governments and business rather than formal demographers who focus on advancing knowledge and sharing it with a community of scientists and the public.

The role of applied demographers, as investigators in service to the making of sound decisions, often requires the researcher to invest time in developing unique solutions to the problems presented. The solutions often require the applied demographer to learn new techniques and expand beyond their specific competencies to develop datasets and methods best suited to the questions provided (Swanson, Burch, and Tedrow 1996; Swanson 2008). These novel solutions often are accompanied with the weird twist of having the question, the answer, and often the data provided by the person or group providing the research topic. Demographic data is extremely expensive to produce and usually dependent on the government for its production (Horton 1998). Nowhere is this more obvious than when we consider how much of the work performed by demographers are completely dependent on some form of government data and often funding (Siordia and Wunneburger 2013). This happens wherever the applied demographer is employed but the dynamic is not unique to the applied demographer as it is felt by more traditional demographers and may be a result of the conservative nature of the discipline and the racial make-up of many of its practitioners (Horton 1996; 1998).

A bright spot in the work of current applied demographers is the fact that there has never been such a rich, high-quality source of data available to conduct research, at a fairly granular level, than what is available through the U.S. Census Bureau's American Community Survey (ACS) product. The ACS provides tabular estimates for every area in the nation down to the block group level which is a geographic level accounting for approximately 600-3,000 persons, on average, and the smallest geographic unit for which sample data is published (U.S. Department of Commerce 1994). These summary data provide a great deal of detail on every topic from the age structure of the population, levels of educational attainment, poverty status, and much more. On their own these data help researchers provide products to aid leaders plan for and deal with the challenges they face (Murembya and Guthrie 2015; 2016). Even with these rich datasets, the applied demographer needs to find a creative solution to providing data for some projects. These projects can range from providing justification for where to place a medical school (Beckett and Morrison 2009) to projects like, as in my case, estimating the number of four-year-olds at specific levels of poverty. Surveying the literature provides further examples of possible uses for a process to interpolate alternate geographies or subpopulations from the microdata samples including looking at the links between education and earnings (Doyle and Skinner 2016), applying resources to increase educational attainment (Lazenby 2011), triangulating the results of estimates produced from administrative records (Bakker 2012), and studying how having parents working affects the level of educational attainment for children in a household (Schildberg-Hoerisch 2011), to name just a few.

The state of the latest Public Use Microdata Sample (PUMS) from the ACS makes it better suited for interpolative analysis than it has been in the past. The current set of Public Use Microdata Areas (PUMA) are much more geographically compact and lack the significant fragmentation that was characteristic of many PUMAs in their previous iterations (Siordia and Wunnenburger 2013). This project will be primarily using single year PUMS data as the current set of multi-year PUMS data crosses a period where there are two sets of PUMAs in the data. These data cannot be used to effectively create sub-state level estimates. The first period where multi-year estimates will be useful for this type of work will be released in December 2017. That release will allow for longer period analysis and will allow for time series work with data that do not have overlapping periods (Census 2008).

The Logic of the Method

The ultimate goal of the process being discussed is to increase the utility of publicly available data by increasing the types of estimates that can be produced. As mentioned previously, this project began with a request from the Michigan Department of Education (MDE). The Department needed data for a very specific group that was not part of the normally published summary data that is publicly accessible through the U.S. Census Bureau. So, a novel solution needed to be developed.

Over the course of working with MDE and these data, I developed and then refined a method for moving data from the Public Use Microdata Sample (PUMS), published by the U.S. Census Bureau, to alternate geographies. This allows for the estimation of a wide variety of characteristics for standard or project-specific geographies. In this section, I will first lay out the logic for the original method then talk about what I learned and how I refined the method.

The Original Method

The specific data needed for this group to complete its work was four-year-olds at 250 percent of poverty, which is not available in the published summary data. General data, having differing levels of poverty or poverty for wider age groups, was deemed to be insufficient. Given that, I had to come up with a better solution. This solution only had two possible outcomes. First, the group could purchase a custom tabulation of ACS data from the Census Bureau, and second, I could figure out a method for disaggregating the data contained in the Public Use Microdata Sample (PUMS) and reaggregating it to the necessary geographies. Given the costs and time associated with the purchase of a custom tabulation, a minimum of \$3,000, the group asked me to attempt an interpolation

of the values from the PUMS data. Other solutions, such as ranking ISDs by general youth poverty, were rejected by the group as not being sufficient to use for the distribution of funds or formula testing.

For this request, the geographies are the complicating factor as single year ages are available in the PUMS data. Figures one through three depict the relevant geographies. In Figure 1, we see the local school districts that make up the intermediate school districts (ISD) that are shown in Figure 2. In Figure 3, we see the Public Use Microdata Areas (PUMA), which represent the smallest level of geography for which PUMS microdata is available to the public. Summary data are available from the ACS for the geographies in Figure 1, local school districts, and, as mentioned, microdata are only available for the geographies in Figure 3, PUMAs. Figure 2 represents the geographies to which the group needs the data to be aggregated, ISDs; these are not standard geographies used by the Census Bureau but are comprised of whole local school districts. The real task, in its simplest form, was to take data at the level of PUMAs and transform it into the geographies in ISDs. To do that I chose to use a weighting scheme derived from data in geographic units from local school districts.

My first step was to determine an appropriate way to weight the data for the disaggregation. I did not have data at the level of ISDs, so I had to use the local school districts as an intermediate step to get from the PUMAs to the ISDs. Initially, I started with a process that would use the population in the local school districts as a weight to disaggregate the PUMA data. For example, say we have ten people in theoretical PUMA X, which is comprised of local school districts A, B, and C; five people live in A, two

Figure 1 Local School Districts (Michigan 2012)

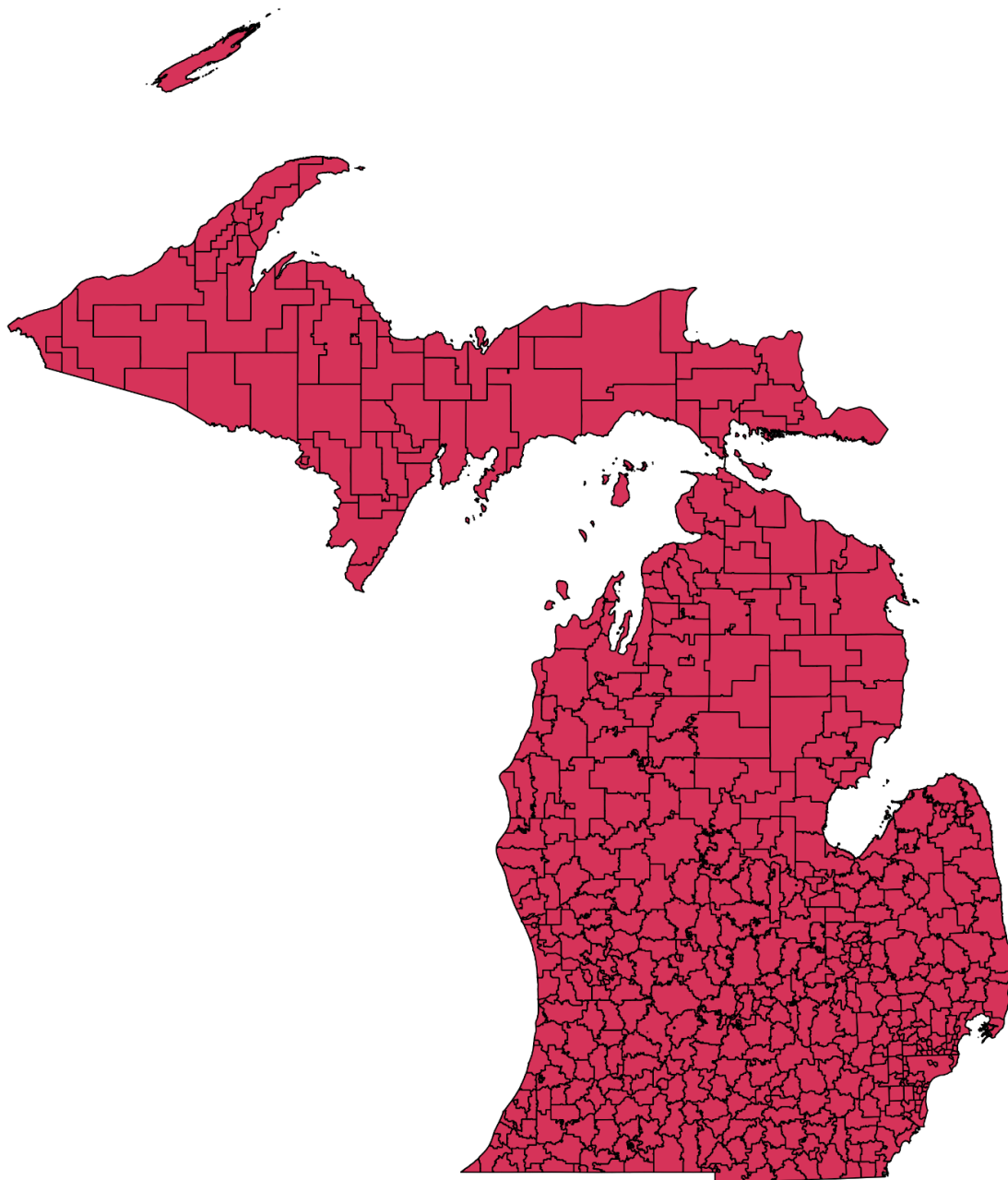


Figure 2 Intermediate School Districts (ISD) (Michigan 2012)

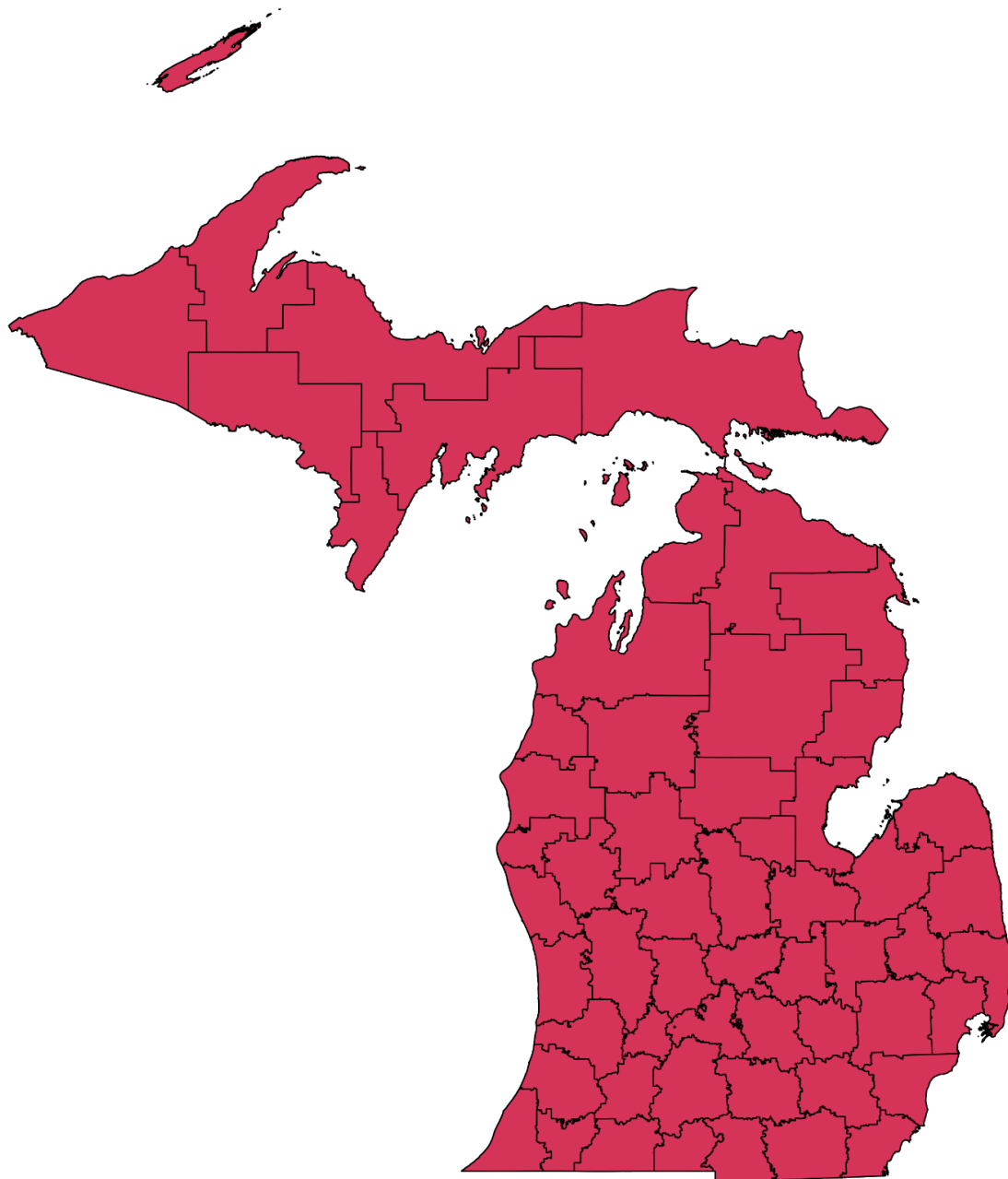
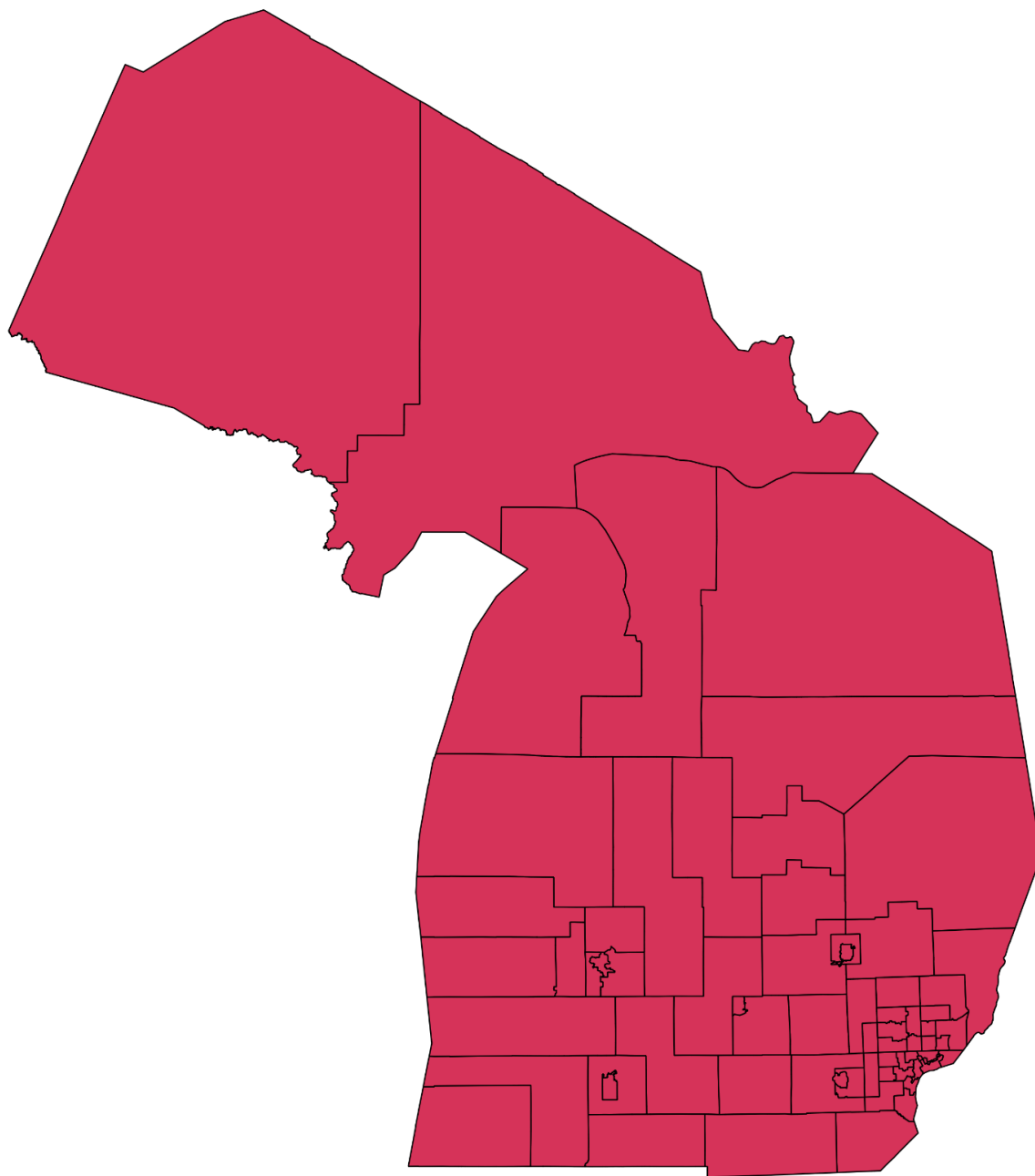


Figure 3 Public Use Microdata Areas (PUMA) (U.S. Census Bureau 2014)

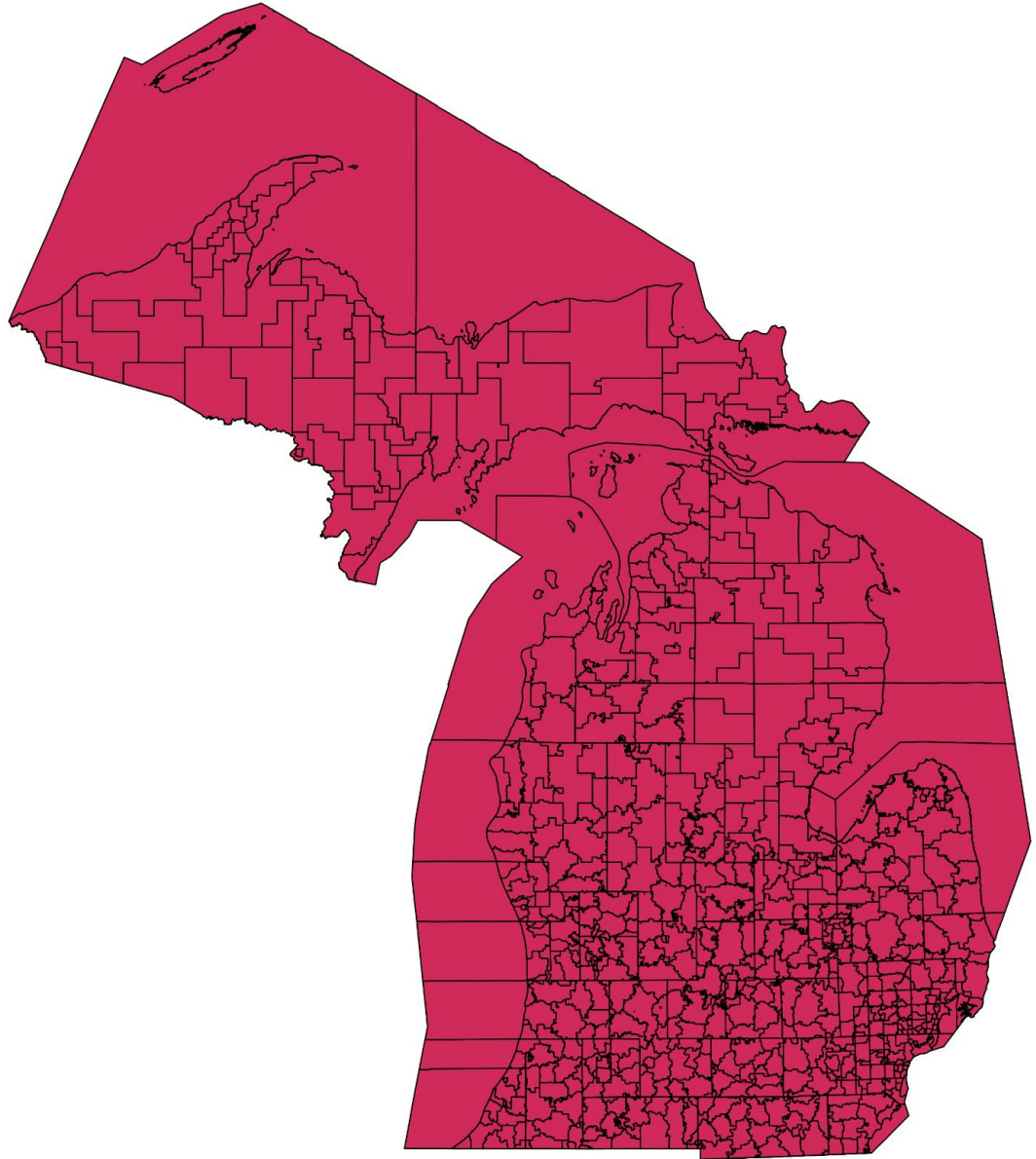


live in B, and three live in C. When I was distributing the persons at 250 percent of poverty, I would send 50 percent to A, 20 percent to B, and 30 percent to C. This is a very simplistic way of describing it, but that is the logic of disaggregation used in the method. This works well for moving data from the local school district to the intermediate level, as the latter is made up completely of whole entities from the former. I abandoned this weighting scheme after the review of the literature proved to me that there are different concentrations of poverty that exist at different levels from simply the concentrations of the population (Owens 2015; Mulherin 2000; Logan et al 2014.)

With that realization, I had to determine a different way to weight the data for my reaggregations in order to provide estimates that would have any validity. I settled on using the number of children between zero and four years of age at the local school district level that were at 100 percent of poverty as the variable to disaggregate the PUMS data from the PUMA level to the local level (U.S. Census Bureau 2015). I used these data because it made logical sense. The populations that I was concerned with, children at four years of age, were wholly contained within this subpopulation in the summary data, and the number of persons at 100 percent of poverty are a significant component within the population at 250 percent of poverty. The use of this group to weight the data for four-year-olds at a specified poverty level seemed logical, but I had no way to verify the correlation and justify the method, however absent a better alternative this is what I used.

In order to move the data from the PUMA level to the local school district level, the project required that the two areas be combined so that the data to be used for the

Figure 4 LSD & PUMA Union



weighting could be distributed to all areas of the local school districts. This was accomplished using Geographical Information System (GIS) software and the required maps for the PUMA level and the local school district level provided by the U.S. Census Bureau and the State of Michigan respectively. The specific map files used in this process are called shapefiles and when combined the state was divided up into every area that was represented by all lines in either file. Figure 4 represents the union of these two levels.

As we can see from the combined geographies in Figure 4, when the two geographies are unioned there is a large number of shapes that have to be accounted for and assigned a certain proportion of the weighting variable. The combination of Michigan's 65 PUMAs with the state's 545 local school districts created 1,104 shapes that had to be accounted for in my distribution scheme. This excludes the shapes that consist solely of water and those that are artifacts created by the combination of maps. The GIS software handles most of these differences very well, but there were 60 shapes that were ignored due to them being artifacts of the union process. These areas were mostly along the border of the state and in total account for a tiny fraction, less than a hundredth of a percent, of the total area in question.

At this point, we have a complete spatial breakdown of the combined PUMA/local school district areas. In order to transform the data from the PUMA geography to the ISD level we need to have a way to pull the PUMA data to certain local school districts based on the weighting variables. The variable that was chosen from the summary data to serve as a weighting variable is the number of children under five whose

family income is equal to or less than 100 percent of poverty. This variable was selected because, as mentioned above, these data comprised part of the target population and there was no other population that contained more of the target population available in the summary data. There is a longer discussion of the concerns with choosing a weighting variable later. In order to distribute the weighting variable over the local school districts, the proportion of the spatial area of the new shapes in the unioned dataset that made up the local school districts in Figure 1 had to be determined. That is a straightforward process that is handled by some GIS software during the process of the union function that created the map in Figure 4. If the particular software being used does not do it automatically, it is not a difficult operation to accomplish in the GIS environment. The process of disaggregating the data for the local school districts consisted of multiplying those proportions by the values for those experiencing poverty. Once those spatial disaggregations have been performed, we have the basis for the weights I used to reaggregate the PUMA data.

To create the PUMA weight, the disaggregated local school district data was reformed into the PUMA geography, which was possible because we have performed the process that created Figure 4. The value of the unioned dataset is that it can be summed to either of the levels that were used in its creation. Then the value for each shape was divided by the sum of those the values that created the larger PUMA. This created a weight that will pull the PUMA data for our population of interest into the shapes from the unioned dataset.

At this point, the majority of the work was completed, and it was simply a matter of multiplying the weights for each geography by the data obtained from the PUMS and summing them according to the ISD to which the original local school district belonged.

This process produced estimates that seemed to represent the distribution of income variation across the state, and it provided MDE with a dataset that they could use to test various scenarios regarding different aspects of the funding formula they were in the process of redesigning. Some of the representatives on the working group had reservations about the estimates for their individual areas, but there was nothing to verify or test the estimates against until the custom tabulation was delivered by the Census Bureau. With the delivery of the custom tabulation, testing of my estimates was possible, and I found that my procedure performed reasonably well. When I compared my estimates to the custom tabulation, I found 87 percent of my estimates were within the confidence intervals provided with estimates. While pleased, I considered possibilities for increasing the accuracy of my interpolated estimates.

The Revised Method

To revise the method, I had to consider the sources of error or deviation causing my estimates to differ from those of the custom tabulation and what I could control. I looked at aspects of how the weighting variable is chosen, the irregularity of the geographies, areal limitation by using land cover data, and at reducing the amount of the weighting variable that was distributed areally. The last of these considerations, limiting the areal distribution of the weighting variable, is where I thought I could have the greatest impact for increasing the accuracy of the estimates. Aside from limiting the area by using land cover data, which has limited benefit when weighed against the drastic

increase in geographic complexity, limiting the areal distribution of the weighting variable seemed to have the greatest benefit.

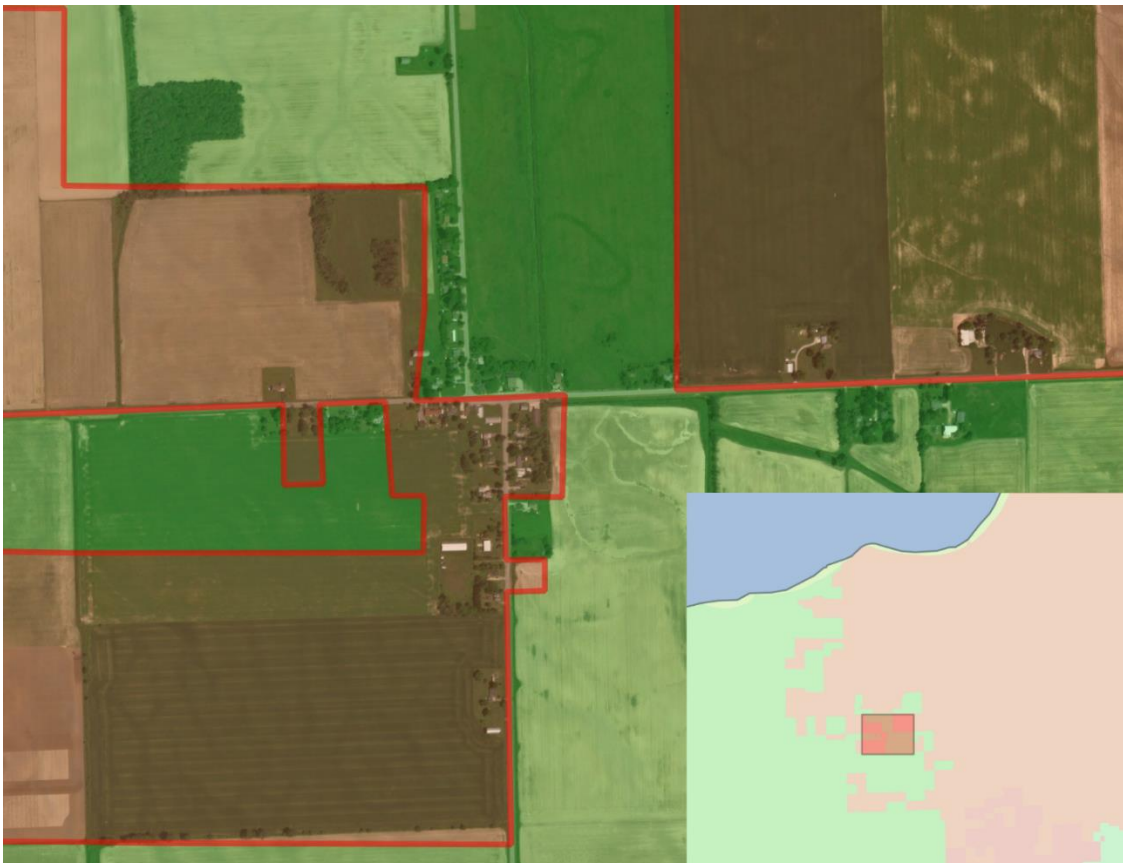
The issues with the ideas for improvement that took them out of consideration were generally impracticality or limited benefit. Issues surrounding the choice of the weighting variable are discussed at length in Appendix C and in the next section, but to summarize, the goal is to find a variable that is maximizing the target population, so if that is done well, there is really not a great deal to do in improving that process that will yield significantly improved estimates. There is always going to be deviation due to the use of a population to weight the PUMS data because there is going to be extra population who do not share the characteristics of the population that is desired in the final tabulation. If there were a perfect weighting variable, i.e. one with perfect correlation to the target population, available in the summary data provided by the Census Bureau, this process would not be needed for specific subpopulations as they could be aggregated or derived directly from those summary data.

The irregularities of the geographic boundaries of the school districts I am working with is another source of error, but one for which there is not an easy or workable solution. The boundaries of the 547 local school districts are set at the local level and are subject to the types of political considerations that may make a particular housing unit part of one district while its next-door neighbors part of another. The types of considerations that can be included in drawing these boundaries can range from which school district a student athlete wants to play in to which school district claims a particular housing unit for purposes of taxation. The types of irregularities that I am pointing to can be seen in the example shown in Figure 5, which shows the boundary

between the North Huron School District and Elkton-Pigeon-Bay Port Schools, which are in Michigan's Thumb region. Michigan's Thumb region thrusts into Lake Huron and forms the eastern boundary of the Saginaw Bay, and "Thumb" is how the area is commonly referenced in the state because of its resemblance to a thumb when viewed on a map.

The larger boundary of the two districts can be seen in the map insert with the North Huron School District shown in pink. The main frame of the map shows an aerial photo of the area, available through Bing maps (2017), with lines and coloring applied to mark the district boundaries. From the inset, it becomes clear that some areas of the North Huron School District are cut out of the neighboring district and are not contiguous with the rest of the district. It can also be observed, from the map's main frame, that

Figure 5 Example Local School District Boundary



some individual households are excised with a high degree of precision that allow the neighbors on either side to be in the neighboring district. This irregularity, I believe, is a source of error in the method, but as mentioned earlier, one for which there is no easy solution.

Removing areas that are not available for habitation was an option I explored, but except for removing the areas from the district geographies under the Great Lakes, it would yield limited benefits. Removing the area that is under the various Great Lakes, reduces the area of the state that will be subject to areal distribution of the weighting variable by around 40 percent. This is a considerable reduction, and it does not greatly increase the complexity of the geographies that are part of the project. The greatest increase in complexity comes from the creation of discontinuous landmasses that form the islands that are part of some school districts. This affects a relatively small number of districts and did not cause problems in the estimation process. Removing inland water (lakes, rivers, streams, etc.) conversely, greatly increases the complexity of the underlying geographies, and provides only slight reductions in the area that is under consideration. Removing inland water reduces the overall area of the state by less than five percent and greatly increases the complexity of the geographies. The overall amount of areal reduction depends on the source and coverage of the shapefiles being used, but all will create significantly more jagged and discontinuous geographies that can cause problems in the estimation process. These problems can be overcome, but the resulting errors in GIS processes and algorithms can be difficult for a user with less GIS experience to overcome. Given that this method is meant for users without extensive GIS

backgrounds, the potential costs in removing inland water greatly outweigh the potential benefits.

Limiting the areal distribution of the weighting variable seemed to be the most viable potential means to reduce deviation between the estimates that are produced with this method and those produced through a custom tabulation of data from the Census Bureau. To that end, a revised method was devised that incorporates a several steps with the intent of distributing the weighting variable to known population centers before areally distributing the remainder to other areas not part of those population centers. The logic of this revision to the methods is straight forward even while the implementation is rather cumbersome and labor intensive. A full explanation and step-by-step walkthrough of the method revision is contained in Appendix C, which are the instructions given to the colleagues assisting with the usability testing discussed later.

The method revision requires that the weighting variable be allocated to population centers first and then to the remaining areas in a school district. This is accomplished by allocating the weighting variable to a new level of geography not discussed previously. This new level of geography is the county subdivision. This level of geography includes incorporated places (in Michigan those are cities and villages), townships, and any area that does not fit into one of those two categories. In Michigan, cities and townships form an exhaustive mosaic of the state's total area, but that is not the case in other, especially western, states that have other organizational configurations. County subdivisions are preferable to what the Census Bureau calls "Places" because of the inclusion of the township administrative unit. For the Census Bureau, a place is an incorporated city, town, or village, and other areas that are not incorporated called Census

Designated Places (CDP). The place level of geography does not form an exhaustive accounting of the state and some of the units overlap with the townships, which are an important level of geography.

The completeness of the county subdivision dataset is one reason to use it as a first level for the allocation of the weighting variable, but there is another important reason as well. Incorporated places do not follow county lines when making boundary decisions. It is often the case that a city will cross one or more county lines. For example, Lansing, Michigan's capital city, exists in three different counties: Ingham, Eaton, and Clinton Counties. This is important because the county subdivision geography divides a city's parts into the respective counties in which they reside. So, for the city of Lansing, pulling data at the county subdivision level will result in three estimates for the city representing the three parts of the city in their respective counties. This is important because when the weighting variable is allocated in the new first stage of the multistage weight process, the population centers to which those data will be allocated will be a subset of the total county subdivision layer that exists wholly within both the local school district level and the PUMA level. As PUMAs tend to respect county boundaries, using the county subdivisions will include more cities compared to what would be included were the place level of geography to be used.

This part of the process requires data for the subset of county subdivisions existing wholly in both the local school district and PUMA, which are easily available from the Census Bureau's website. While the data are readily available, determining the appropriate subset of counties is more challenging. This is done with the aid of GIS techniques and the process is explained in detail in Appendix C. The reason we are

choosing county subdivisions that exist wholly within both the school districts and the PUMAs is to decrease the areal interpolation of the data that will be assigned to these areas. The main purpose of this revision to the original method is to decrease the areal interpolation of the data in the process. If the process just shifted where that areal interpolation was occurring, there would be a potential to increase the error in the estimates, as poverty is often concentrated in unique patterns within population centers.

With the subset of county subdivisions determined, a variety of correspondences and areal differences need to be determined so that data can be accounted for once and not duplicated through the process. To obtain a list of correspondences, a set of straight forward queries in the GIS environment need to be performed, where the end result will be a table that will list for each county subdivision the corresponding local school district and PUMA in which it resides. With that correspondence table in hand the next step is to remove the geographic area of the selected county subdivisions from the area in the local school district layer. When those area have been removed, the two sets, selected county subdivisions and reduced area local school districts can be combined to form a mutually exclusive and exhaustive mosaic of the state's land area. With the combining of the local school districts and the county subdivisions into a single geographic layer in the GIS environment, the data can now be attached, and each geographic unit's special area should be recorded.

With the data attached to the combined school district and county subdivision layer, the data is ready to be combined with PUMA layer. This is accomplished in the GIS environment with a union algorithm. This task takes the two layers—the PUMA and the combine layer created in the previous step—and forms a single layer. The union

algorithm in the GIS software will take every line in each file and combine them into a single layer. The effect of this combination is the fracturing of every shape in each set into a set that represents a mosaic layer of non-overlapping geographies that form a set of shapes that can be summed or combined to form either of the two input layers, the local school districts or the PUMAs. This transferability of data is what makes the whole method work. The method takes data from the local school districts that most closely relates to the target population, and forms weights that then pull the PUMS data to these small, unioned shapes that can then be reformed to the local school districts.

Once the data has been attached to the small areas created from the union process. The work shifts out of the GIS software and into a spreadsheet where data can be manipulated to create the final estimates. Before the data that has been attached to the small areas in the unioned layer can be turned into weights, it first must go through a deduplication process. This is necessary for two reasons. The first is that the county subdivisions overlap with the school districts of which they are a part. So, if the populations that are being allocated to the school districts were not removed from their school districts, this process would not have the full desired effect of reducing the amount of areal interpolation. For example, say we had a school district with a population of 100 students in the weighting variable's population, and forty of those students lived in a county subdivision. When the deduplication process is completed successfully, 40 percent of the weighting variable is attributed to the county subdivision (40 out of 100 students). If we do not deduplicate the data only about 29 percent of the weighting variable's population will be attributed to the county subdivision (40 out of a 140 population).

The second reason that the data needs to be deduplicated is similar to the first, but it would have the opposite effect. Even with selection of county subdivisions that exist wholly within school districts and PUMAs, it is possible for a county subdivision to be separated into multiple, non-contiguous parts during the difference and union processes. Deduplication needs to check for and correct the data to account for these areas. Not accounting for these areas would overemphasize the population centers. For example, using same population numbers from the previous example, 100 persons in the school district and 40 in the county subdivision, and assuming a particular county subdivision was split into two parts, not accounting for the duplication resulting from the splitting of the county subdivision would pull 44 percent of the data into the county subdivision (80 of 180) versus the forty percent which it actually constitutes. This problem would compound for every additional part that a county subdivision is divided.

The process of deduplication occurs in spreadsheet software and is completed algorithmically rather than correcting individual entity data. To correct for the first and more likely type of duplication, the total for each data point for the local school districts will need to be reduced by the amount accounted for by its constituent population centers. This can be accomplished using the correspondence tables created during the process. To correct for the possibility of noncontiguous portions of a county subdivision, the amounts for each county subdivision should be multiplied by the ratio of the area for the county subdivision from the unioned dataset to that of the original county subdivision. If a county subdivision has not been split, the ratio would equal one. If it had been split, the ratio would send some of the weighting variable to each portion of the county subdivision, depending on its relative size. When these two procedures are applied to the

data correctly, the new totals for the weighting variable data should equal or approximately equal (there may be some small amount gained or lost due to the rounding of the areal splits in the second step) the total amount attributable the original school districts. This procedure is explained fully with the accompanying spreadsheet formulas in Appendix C.

With the weighting variable attached to each polygon in the unioned set and following the deduplication process, the data can be turned into weights. At this point the steps return to those that were developed for the original implementation of the method. As mentioned previously, this is a fairly easy process where the current amount for a polygon is divided by the total amount for the PUMA of which it is a part. This is the most technically simple part of the process and is accomplished with a single function, but it is very important as is what allow the transfer of the PUMS data to the unioned polygons.

When the weights have been created, they can be multiplied by the PUMS estimates for the target population. This will create an estimate of the target population for each polygon. As discussed earlier the most important property of the unioned polygon set is that it can be summed to either the PUMA areas or the school district areas. Once the PUMS estimates have been disaggregated to the unioned polygons with the weights, the last step is to sum them to the desired geography, the school districts. With that step complete, the final estimates of the target population by the desired geography are complete.

Determining a Weighting Variable

The purpose of the weighting variable is to determine how much of the total subpopulation from the PUMS to attribute to each polygon created by merging the PUMA geographies with the target or bridging geographic areas. The goal of selecting a weighting variable should be to pick one that most closely correlates with the subpopulation for which you are trying to make an estimate. It may be possible to use a variable from the summary data that represents your entire subpopulation, though that would be unlikely, and would only occur if the researchers were making estimates for a standard population or subpopulation of a user defined geography. The more likely scenario would be having to select from a variety of imperfect matches available in the summary data.

Taking the example of the request from MDE, the need was for a variable that I could use to pull data from the PUMS data to the polygons created from the merging of the PUMA and local school district geographies. MDE needed estimates for the four-year-old population that was at or below 250 percent of poverty. The final set of estimates I produced for the project used children 0 to 4 years of age who are at or below 100 percent of poverty, what I call early childhood poverty. My first attempt at this process used total population as a weight, but the results from that were unsatisfactory considering the knowledge and literature that point to poverty being distributed differently from the general population distribution (Iceland and Hernandez 2017). The value of the exercise with the total population was more to prove the process worked and could be used to transform the data from the PUMA level to another geographic level.

To arrive at that as a weighting variable, I generated several sets of estimates, each using a different variable from the summary data to serve as the weighting variable.

The first set of alternate estimates used the general population in poverty as a weighting variable. This produced a set of data with a different distribution than the first dataset which used the general population. The distribution also made more sense when compared to the set produced with the general population as a weight. My overall impression with this set was the process of refining the weighting variable was having a positive impact on the final product, so I decided to produce at least three additional sets to see how they compared to known distributions. The variables I used in this process were youth poverty (0 to 17), extreme or “deep” poverty (population below 50% of poverty, regardless of age), and population below 200 percent of poverty.

The variables all worked as variables for the process, but some worked better than others and others looked very similar. The final preschool poverty variable seemed to make the most sense when I presented the results to the group for them to consider. This is also the distribution that matched most closely with past funding distributions and counts that were available to the program administrators. While it performed the best, meaning it produced estimates that most closely matched the expectations of the group and distributions produced for previous years funding reports, other variables worked well too. The overall youth poverty worked well and could have been a final variable, though it was not, in the end, tested against the purchased data. Similarly, the extreme poverty weight seemed to work well, but it seemed to favor the dense urban areas more and seemed to short-change some of the more rural districts. Lastly, the 200 percent of poverty weight tended to look more like the first attempt that just used population. Part

of the explanation for these observations relate to how poverty is distributed. While rural areas definitely have pockets of poverty and individuals that are experiencing extreme poverty, areas of concentrated poverty tend to exist in more urban areas. There are also higher concentrations of children in urban areas as the population is tending to migrate to more urban areas. Similarly, as we move up the scale of income to poverty ratios, we are including more and more of the population, so a map of persons who are at or below 200 percent of poverty would look more like the general population than would a map of persons at or below 100 percent of poverty,

What each of the attempted weighting variables have in common is their selection of the population by various levels of poverty. The variables that produced the best results were variables that eliminated as much of the population that was not part of my target population. For example, if we think of the youth population in terms of the age groups represented, we have persons between and inclusive of the ages zero and seventeen, while the target population for estimates was specifically four-year-olds. This means that in terms of the age groups represented, the four-year-olds would be about five and a half percent. That would, of course, vary depending on the age structure of the particular geography, but four-year-olds are only one age group out of a possible 18 in that range. Similar issues made the other attempted weighting variables underperform when compared to the final choice of the early childhood poverty variable.

In an ideal world, there would be an estimate of the population that you are trying to estimate so that you can compare and find the best weighting variable. However, this process is meant to assist with making estimates for populations that do not have independent estimates with which to make the determination. The judgement of the

researcher and the insights gleaned from a review of the data and subject area literature will need to guide the choice. There are however a few guiding principles that should aid in this choice.

1. The weighting variable should maximize the target population. This should help to prevent non-target portions of the weighting variable from exerting undue influence over the final estimates. For example, if your target population is four-year-old children, use early childhood poverty instead of general children in poverty because four-year-olds make up a greater proportion of the early childhood ages than the general childhood ages.
2. Maximize the target characteristics in the weighting variable to give the estimates as much geographic specificity as possible. The assumption is social characteristics are autocorrelative in nature (Males and Brown 2014; Frank 2003), and they will cluster, so maximizing the characteristics in the weight will help them to reflect the actual social conditions. This would make the 100 percent poverty data work better than the extreme poverty data, which was 50 percent or less of poverty, as the final estimates were for 250 percent of poverty. Care needs to be exercised when implementing this principle as maximizing a social characteristic may affect the proportional size of the population. For example, the summary data do not provide age specificity for poverty data at levels other than 100 percent, so using data at 200 percent of poverty would mean that the weighting variable would be using the entire population for which poverty status was determined. This dramatically reduces the proportion of the target population in the weighting variable.

3. Reduce, as much as possible, known cohort effects that will distort the final estimates. For example, the use of general poverty or youth poverty were not as good of choices as was early childhood poverty because of the known negative correlation between poverty status and age. As age increases, the probability of someone experiencing poverty decreases (DeNavas-Walt and Proctor 2014), so by using general or youth poverty, the weighting variable is simultaneously decreasing the proportion of the target population and including more population that have a different (lower) probability of experiencing poverty.

The Method in Brief

The core function of this method is to move data from the Public Use Microdata Area (PUMA) to the desired geography. This should be able to be done with any subpopulation that has sufficient observations in the dataset. PUMAs are statistical areas designated by officials in the states' State Data Centers (SDC) according to the guidelines provided by the Census Bureau. The guidelines require that every PUMA area contain a minimum of 100,000 people, and for PUMA areas created for data releases since 2012, the PUMA must be constructed from contiguous census tracts with emphasis on keeping counties whole when possible ([U.S. Census Bureau 2011](#)). These guidelines provide areas that do a good job of preserving respondent confidentiality as they require the PUMAs to maintain a large minimum population. This property is also a drawback as their size often makes them large and unwieldy, especially in rural areas where they likely group several counties together in a single PUMA. Due to the guidelines for PUMA creation using census tract and population thresholds as requirements, the PUMAs usually do not conform to any recognizable areas with the exception of counties and cities that have populations that exceed 100,000. This can create problems in putting them to use when trying to answer demographic questions below the state level, and makes it impossible to isolate all but highly populated areas. This is a problem when trying to make estimates for small areas or for areas with low populations. The revised method will not solve all the problems experienced by individuals trying to work in low or geographically diverse areas, but it will provide assistance and another tool in the demographic data user's toolbox.

The basic steps needed to produce an interpolated set of estimates for alternate geographies and/or subpopulations are as follows:

1. Review the geographic requirements to determine if it is possible to use a census geography to create or approximate the geography required for the projects. The requirements of the request from the MDE called for the data to be presented in Intermediate School Districts (ISD). The ISD is not a geographic unit for which the Census Bureau published data, but all ISDs are aggregations of various numbers of Local Education Agencies (LEA), which are geographies tabulated by the Census Bureau. This is a very important step as a census geography is necessary to build the weights that will be used to distribute the values from the PUMS data. The estimates produced for New Jersey (discussed later) will be at the school district level, and the South Dakota (discussed later) estimates will be for the counties. New Jersey and South Dakota will use county subdivisions and census tracts, respectively, as intermediate geographies in the first stage of the weighting process. A valuable lesson learned through revising this method is how much effort is saved by using as many shapefiles from the same source as possible. For the estimates, all shapefiles were obtained from the Census Bureau's shapefile collection (U.S. Census Bureau 2016)
2. Determine if the geographic requirements for your project require an intermediate geography that from which you will aggregate to your final geography. As mentioned above, all estimates produced for testing will use the county subdivision layer as an intermediate level for the initial state of weighting.

3. Determine an appropriate weighting variable to move data from the PUMS level to the geographic level you have selected from the previous steps.
4. Obtain or produce shapefiles for all the geographic levels needed, and prepare them for use.
5. Add the weighting variable values to the geographic area you will use to move the PUMS data to the project specific geography.
6. Combine the PUMA shapefile with the shapefile that will have the weighting variable data added. This should provide an exhaustive accounting of the study area with polygons that can be added to either the PUMA areas or the weighting variable areas.
7. Once the data are attached and the shapefiles are combined, the deduplication process will be performed.
8. Weights are created that will be applied to the PUMS data. The total of these weights should exactly total the number of PUMAs in the state.
9. With the final weighting variable in place, the project is ready to make the interpolations of the needed or intermediate geography. The nature of the unioned dataset allows for the weights to move the data from the PUMAs to the required geography.
10. Review the estimates for face validity and perform any other validity checks that are possible based on the local level data that are available or compare them to gold standard data that may be available. It is unlikely that any real gold standard data will be available as such data would make this process unnecessary. This project does have those gold standard data as the group moved to purchase data

from the Census Bureau to fulfill the programmatic requirements of the project. I will use those data to test the results of the method.

Testing the Method

This method is primarily intended to be used in areas and for subjects about which there are no other data sources available or in some instances to verify other forms of data that may be collected for which there may not be appropriate comparators. Given its nature as a weighted disaggregation and interpolation of other estimates, this method will never suffice as a final arbiter of funding or resource decisions nor will it independently prove any position or theory. It will provide estimates for geographies and subpopulations for which other sources are mute. The method's limited applicability does not mean that it can be employed without testing and validation on an independent level. The testing regime that I will employ for the method will look at the validity of the estimates that were produced and the relative ease with which the method can be employed.

To look at the relative ease of implementation, I have asked colleagues to take the detailed method section (Appendix C) and try to recreate the estimates and provide feedback regarding the method. Additionally, I will produce estimates for some other states to demonstrate the portability of the method. These will include county level estimates for the state of South Dakota, and school district level estimates for the state of New Jersey. Michigan is a state that has a wide variety of area types and the states of South Dakota and New Jersey are further away from that middle, each varying toward a different extreme. South Dakota is a state with a large amount of open, rural area, which presents challenges that will be different from those in Michigan. New Jersey goes in the opposite direction and is a very urban state. The very high population densities will test the ability of the method to produce estimates for areas with very small geographies and

very high populations. Additionally, New Jersey is a physically smaller state than either Michigan or South Dakota. New Jersey has less than one tenth the area of Michigan. The differences in size, density, and the respective rural/urban splits for the three states make them useful tests for the method.

To test the validity of the method, I will perform some statistical tests to compare the estimates I produced with estimates that were purchased through the American Community Survey Office's custom tabulation program. I have access to a few of these custom tabulations that I have access through my work with the State of Michigan.

Usability Testing

The usability testing will be discussed in two sections. The first is the general applicability of the method. In this phase I will be looking at expanding the method to other locations, I will look at producing estimates for South Dakota counties and New Jersey school districts. I see nothing in the method procedures that will prevent the method from being applied across different areas, but testing is necessary to verify.

The second phase of usability testing will involve asking colleagues to replicate my Michigan ISD estimates using the detailed explanation of the method provided in Appendix C. The whole point of this work is to produce a method that other researchers can employ to gain better understanding. Putting a tool in their kit that they cannot use will serve no purpose.

Wide Applicability

The method can be used to create estimates and distributions for a variety of subpopulations and geographies. To look at how implementation can be extended to

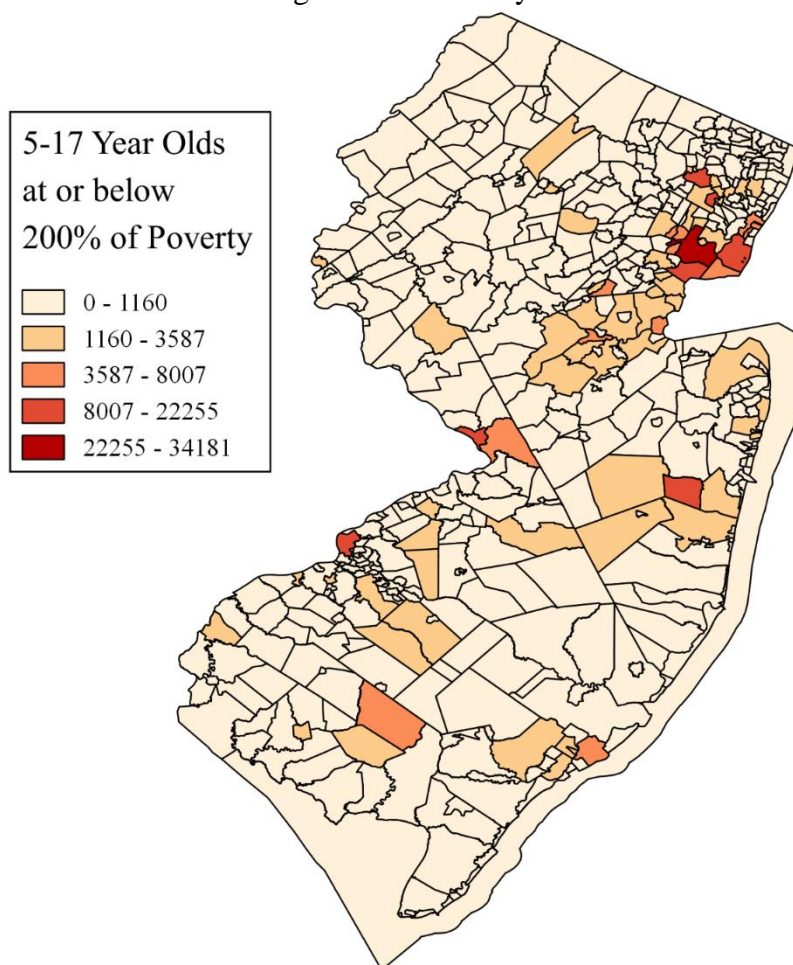
locations beyond Michigan and the populations discussed at length in the previous sections, I have created estimates for different populations for the states of New Jersey and South Dakota. For New Jersey, I made estimates for school the school age (5-17) population who live in families whose income is at or below 200 percent of the federal poverty level by school district. For South Dakota, I made estimates for the county level number of males between the ages of 26 and 32 who are employed. The complete table of estimates for each geography and subpopulation are available for review in Appendix A.

The estimates for the New Jersey were not any more difficult to complete in terms of the calculations or GIS procedures, but there are peculiarities relevant to the state of New Jersey that made the production of the estimates more challenging than they might have otherwise been for a researcher with more familiarity with the state. New Jersey has three sets of school district level geographies which had to be reconciled in order to create an exhaustive and mutually exclusive set of polygons for the state. The three types of school district geography provided by the Census Bureau in its data products and shapefiles are elementary, secondary and unified. From an investigation of the shapefiles for these three geographies it appears that secondary school districts are aggregations of the elementary district level polygons. The unified school district level is a unit that is entirely exclusive of the secondary or elementary levels, and when combined with either the elementary or secondary level geographies, creates an exhaustive and mutually exclusive set of polygons for the state. With that knowledge, I decided to combine the elementary districts with the unified districts and make the estimation according to those delineations. My thoughts were that the smaller elementary districts would better serve

to isolate pockets of the variable of interest and they could be aggregated to the secondary districts if the interest or need arose.

As mentioned above, the estimates produced for New Jersey can be seen in Appendix A. Once completed, I began reviewing the estimates to see if there were any obvious problems. I noticed one thing immediately—there seemed to be a number of districts where there were zero of the population of interest estimated. I followed my method backward through the data to see if there was an error, but not finding one, I turned to an alternate data source to glean some insight. I looked at the data published by the State of New Jersey’s Department of Education, which included a count of students enrolled in the free or reduced priced programs provided in area schools (New Jersey 2017). Those data agreed with the estimates I produced, which lent credence to the estimates. Without a data source to use for testing, I do not have a way to statistically test my estimates as was possible for the Michigan estimates, but the procedures worked well, and I believe that the procedure might even work better because of the larger number of PUMAs covering a smaller geographic area. The map below is a representation of the estimates shown in Appendix A.

Figure 6 New Jersey Estimates

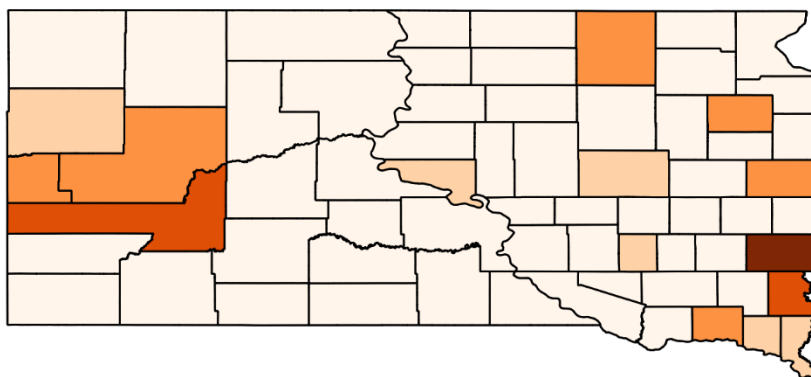
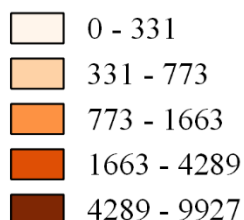


The procedure seemed to work well in an urban state, New Jersey, and a state with a wide mixture of urban and rural, but there was a need to round out the testing with a more rural state to see if the method was really as widely applicable as I believed. I chose to produce estimates for the State of South Dakota's counties. In this test dataset I made estimates for males between 26 and 32 who are employed for each county in the state. The results for this test can be seen in Appendix A.

Considering I was producing estimates for South Dakota counties, I used the alternate implementation of the method described in Appendix C, as it is a shorter procedure and easier to implement. To complete the estimates, I used census tracts as a bridging geography. There were not any problems in the procedure, but as would be expected in a rural setting, the census tracts that I was using as the weighting geography were often quite large. This did not seem to have an effect on the quality of the estimates, but was something that was striking having just created estimates for Michigan and New Jersey whose geographies were much more populous and therefore smaller in size. A graphic representation of the estimates produced can be seen in the map below.

Figure 7 South Dakota Estimates

Employed Males
Between 26 and
32 years



The method is clearly capable of producing estimates for a wide range of geographies and topics. The results of the testing for wide applicability seem to be positive, but without gold standard data to test these estimates against there is question as to the validity of the method. This will be discussed more in the Validity Testing section to follow.

Replicability Testing

This phase of testing involved giving the method to colleagues to have them follow the instructions (found in Appendix C) and produce estimates. For this I enlisted the aid of two individuals with whom I work. Both of these individuals have knowledge of and experience with GIS software and extensive experience working with data. I asked these colleagues to replicate my estimates for Michigan ISDs with the same parameters I used in the production of my estimates.

One of my colleagues was able to produce estimates based on my written instructions, and both provided valuable feedback that made the instructions more clear and readable while they were in the process of producing the estimates. The colleague who was unable to complete the process indicated that it was related to computer issues—he had to have lengthy computer repairs initiated while he was attempting to make the estimates. Additionally, after they completed the process, I conducted an informal interview to gain additional feedback. Those interviews were guided by the following questions:

1. Were you able to make the estimates with the instructions provided?
2. Where there any major problems with the method as described?
3. Do you have any suggestions to improve the method?
4. Can you see a use for this type of method in relation to the work that you do, and if yes, what would that be?
5. How would you rate the ease of using the method to create estimates?
6. With this experience and my written instructions, if you wanted to create your own estimates for some other geography or subpopulation, do you think you could accomplish the task?

When my colleagues attempted to complete the process, both had to get clarification on some of the issues that caused confusion. Those requests for clarification provided opportunity to improve the instructions and make the process clearer. The instructions in

Appendix C are the instructions that were given and revised through the testing process. The goal of those instructions was two-fold. The first goal was to provide instructions on how to complete the steps necessary to create an estimate set, and the second was to provide enough explanation so the user could repeat the process outside of a training environment. If the process was going to be useful to someone other than me, they had to understand why things were being done so they could use the process, and not just follow my steps. One of my reviewers acknowledged this when he indicated that the length of the instructions may have been an impediment, but “but sacrificing clarity for conciseness is dangerous.”

Both of the reviewers indicated that the process is not something that they would implement in their current position, but they could see the utility in the future. This response is not a surprise as the jobs held by these colleagues are not positions where they are required to consult on data availability with other agencies in the state government. One of my reviewers is an Economic Analyst for the State of Michigan and the other is an Employment Projections Specialist. Neither of these positions in state government require the production of data outside of what can be aggregated from those data that are publicly available. Both reviewers indicated that this method would be something that they would remember and possibly use or use some variation of in the future.

The reviewers also indicated that the method was not particularly easy to implement. One user described, “using the face pain scale I’d say about a 4.” He was referencing the Wong-Baker Faces Pain Rating Scale in a joking manner to indicate the process was moderately painful to complete, but the pain was not insurmountable. Both reviewers

indicated that learning a complicated method was not easy to do from written instructions, but that it was clearly possible, and were it necessary to their work, the pain of learning the method would have been worth it. On that point, both reviewers also indicated that they could implement the method outside of the example they were given if the need arose.

Given that both of my colleagues indicated that they understood the method and were confident that they could implement the method if they were called to in the future, I consider this part of the testing to be successful. Both colleagues I called on to assist me were people with some GIS experience, but neither were GIS professionals. Similarly, both had passing familiarity with data from the Census Bureau, but neither are called on to use it on a daily basis, as I am required to do in my position as Michigan's State Demographer.

Validity Testing

The opportunities for testing the validity of results from an improvised method are sparse as the production of appropriate comparators is expensive and not typically possible. The heart of the original and revised methods described in this work is figuring out a distribution scheme for public use microdata, which are drawn from the same sample as the weighting variable, to obtain a rough proxy estimate for the subpopulation or geography under consideration. The best comparator would be an actual estimate produced by the American Community Survey Office (ACSO), which is the organizational unit at the U.S. Census Bureau responsible for the collection, production and publication of all the data products for the ACS program.

The ASCO will commission estimates by request for external parties, but these requests are handled on a fee for service basis, and the minimum charge to produce a custom tabulation is \$3,000. The reason to pointing out the expense of obtaining a custom tabulation is not to criticize the expense or amount charged. The production requires the work of several staff member from the ASCO and the review of boards within the Bureau charged with ensuring respondent confidentiality. Given the work required and the general expense of conducting the ACS, the \$3,000 charge is extremely fair and probably less than one would expect. Even with the acknowledgement that the fee is reasonable, it is still a barrier for some as is the time required to produce a custom tabulation, approximately six to eight weeks at a minimum. It is definitely a constraint in the context of obtaining data to test the method described in this work.

I have been able to obtain three custom tabulations I will be using to test the original and revised methods. These datasets were purchased to assist in the

administration of various programs in the state and those datasets have been made available to me to assist with the testing of this method. The custom tabulations represent two years' worth of data purchased to assist with the administration of the Michigan Great Starts Reading Program (GSRP) for children four years of age who are living in households at or below 250 percent of the federal poverty level and the third is for individuals 60 years and over who are living in households at or below 150 percent of the federal poverty level. The GSRP data is organized by Intermedia School District (ISD) and the elder data is organized by Michigan counties. The datasets are for different years and the dataset for the counties is drawn from the 2006-2010 5-year estimates. These datasets give me a diverse set of estimates to test against and cover different years, so that the tests can be seen to be independent of one another. The school district data are drawn from the 2014 and 2015 estimates.

One issue arose when I was requesting the 2015 data from the U.S. Census Bureau. Between my request for the 2014 data and the request for the 2015 data, there were some changes on the disclosure review board at the Census Bureau. That change meant that while I was able to get a complete set of 1-year data for 2014 which represented every intermediate school district in the state, I was not able to get the same data for 2015. For that year, the disclosure review board decided that the same constraints that were in place for the 1-year summary estimates also needed to be applied to the custom tabulations. The effect of that decision limited the 2015 data to only those ISDs whose total population exceeded 65,000 residents, which reduced the school districts available for testing in the 2015 data from the 57 available in the 2014 data to 32.

The statistical tests I will be performing will consist of testing two estimated distributions, one from the original method and one from the revised method, against the data obtained from the custom tabulations purchased from the U.S. Census Bureau. The methodological description in the preceding pages describes both methods I developed to fulfill the request from the Michigan Department of Education.

When I am performing the validity testing I tested my original method and my revised method against the custom data. I have gone into detail about both the original method and the revised in this work. The original method areally distributes all of the weighting variables to the unioned geographies and those areally distributed data are the building blocks for the weights used to distribute the PUMA data. The revised method first allocates portions of the weighting variable to population centers and then areally distributes the residuals to the larger areas. The procedure for creating the weights is the same for both methods. I am comparing both methods here because I do not know if the revisions I have made will make the method better. Conceptually, they should make the estimates better, but I am only guessing until I test both versions of the method to see which performs better, or if there is any difference at all.

I first performed a paired t-test to determine if the distributions were different from the data purchased from the Census Bureau. My thinking here was a situation where the null hypothesis can be rejected for a t-test would demonstrate evidence that the distributions were different, which would end my need for testing considering the desired result was equivalent. If I was unable to reject the null hypothesis for the t-test, I would continue to perform a signed-rank test for stochastic equivalence (Dinno 2017). I planned this two-step testing procedure because the failure to reject the null on the t-test

means I cannot say that they are different, but that does not mean that I can say they are the same but rejecting the null on the signed-rank test for stochastic equivalence did provide evidence of equivalences. A reader might ask why I did not just perform a signed-rank test for stochastic equivalence in the first place as that provided the result in which I was interested? The reason for this two-step process is the signed-rank test for stochastic equivalence is a nonparametric test and does not have a high sensitivity and the null hypothesis is sometimes rejected when there is a difference. The signed-rank test for stochastic equivalence, due to its low sensitivity, also frequently returns an intermediate result where the researcher can neither say the distribution is equivalent nor different. In that circumstance the signed-rank test for stochastic equivalence is simply underpowered and unable to make that determination. The conclusion from that result is usually that there is a trivial difference that is preventing the conclusion of equivalence, but not sufficient to allow for a finding of difference (Dinno 2017).

Before I could begin to test my results, I had to contend with a problem that I had begun to consider when I was producing my estimates. There are slight differences between the results that are produced from the public use microdata and the custom tabulations that are produced by the U.S. Census Bureau. These differences are to be expected as they are technically being drawn from different datasets. The custom tabulation is being drawn from the full, restricted-use dataset available to Census Bureau employees, and the dataset I am producing is being derived from the PUMS data. The PUMS data is a sample of the restricted-use data, so the estimates should be similar, but they will not be the same. For example, when looking at the year 2015, estimates produced from the PUMS data put the total number of four-year-olds at or below 250

percent of poverty at 60,548, while the total number from the custom tabulation was reported by the Census Bureau as 61,435. This is a small difference and the estimates are about the same population in the same time period, but it was a difference for which there needed to be a correction made. To bring these estimates in line with one another for testing purposes, I raked the interpolated estimates so that the total would agree with the estimates produced through the custom tabulation program. This raking procedure would adjust every estimate in my series by the same proportion, so the distribution would remain constant. This was a necessary step to ensure that differences or equivalencies can be attributed to the distributions and the procedures rather than to the differences between the datasets from which they were drawn.

I started with testing to see if there was a significant difference between the ACS estimates and my original method which was solely based on areally interpolating the estimates, without the step of pulling data to the county subdivision level for 2014. There was no significant difference detected between the original method ($M=1169.14$, $SD=303.87$) and the ACS estimates ($M=1169.21$, $SD=323.3$); $t(56)=-0.0013$, $p=.9989$. That result was expected as I was generally happy with my original estimates, and thought they generally represented the population well. During the process of performing this test, I discovered that my data had some outliers and that the distribution of the differences may not be normally distributed. The values of the differences did not deviate sufficiently enough from the normal distribution to be able to be detected on a Shapiro-Wilk test for normality, but they were not sufficiently normal to pass a graphical investigation. This was a problem because this violated two assumptions of the t-testing procedure therefore I needed to reevaluate my testing plans.

To accommodate the new reality of the data I was working with, I had to embrace non-parametric testing procedures for the entirety of the process. This was less desirable, but a necessary step to produce valid results. The signed-rank test for stochastic equivalence is a statistical testing procedure and package for the STATA statistical software that is actually a combination of three different tests. The first is a Wilcoxon signed-rank test that is used to test for differences and is a non-parametric comparator to the t-test. This test portion of the procedure will determine if there is sufficient evidence to determine if a set of estimates is sufficiently different from the test data provided by a custom tabulation. The second portion of the testing procedure involves performing two one-tailed tests that determine if there is sufficient evidence to conclude the two distributions being compared are equivalent. This procedure is explained and validated by Stefan Wellek in excruciating detail (2003). The tests for equivalence checks if one distribution dominates the other (consistently get higher ranked positive or negative ranks) and judges that dominance against a predetermined amount ϵ which is expressed in units of the z distribution (Dinno 2017). The value of ϵ used in the tests was 1.645 which corresponds to the 90% margin of error the Census Bureau uses when publishing summary estimates and for the custom tabulation program.

The specific hypotheses used for testing the distributions for differences are as follows:

$$H_0: \sum(x_1 - x_2) = 0$$

$$H_a: \sum(x_1 - x_2) \neq 0$$

Where x are the paired estimates from the interpolation methods and the gold standard data obtained from the custom tabulations of the ACS data, respectively.

When testing the estimates for equivalence, as mentioned previously, the procedure involves two separate tests. For the distributions to be equivalent, the null hypotheses for both need to be rejected. The specific hypotheses are as follows:

$$H_{01}: \varepsilon - z \leq 0$$

$$H_{a1}: \varepsilon - z > 0$$

-and-

$$H_{02}: z + \varepsilon \leq 0$$

$$H_{a2}: z + \varepsilon > 0$$

The table below shows the results for the tests of the data produced for 2015, 2014, and 2010. The results of the testing are encouraging if anticlimactic. Every dataset seemed to have trivial differences that allowed them to pass the test for difference with very large p-values suggesting they were far from being able to be considered different, but they also were not able to reject both the null hypotheses in the second stage of testing involving the two one-tailed tests, which is required to conclude that there is evidence that the distributions are equivalent. In each case the second stage of the tests had one of the two tailed tests that was able to reject the H_0 for one tail but not the other. The author of the test package for the Stata software refers to this result as an intermediate result. An interpretation concluding the distributions have trivial differences seems appropriate considering there has been an *a priori* determination of no difference

made in advance of the two one-tailed tests (Dinno 2017). Complete tables with the results from the estimation procedures are in the Appendix B, which would allow for independent confirmation of these results.

Summary Table of Validity Testing Results

Signed-rank Test for Stochastic Equivalence							
		2015		2014		2010	
		Original Method (areal)	Revised Method (two-stage)	Original Method (areal)	Revised Method (two-stage)	Original Method (areal)	Revised Method (two-stage)
Wilcoxon Signed-Rank Test (Difference)							
$\alpha = 0.05$	Total Pairs	32	32	57	57	83	83
	Total Ranks	528	528	1653	1653	3486	3786
	Number Positive	17	15	27	28	38	38
	Sum Positive Ranks	280.5	277	800.5	803.5	1653.5	1648.5
	Number Negative	15	16	30	29	45	45
	Sum Negative Ranks	247.5	250	852.5	849.5	1832.5	1837.5
	Zero Ranks	0	1	0	0	0	0
	z	0.309	0.252	-0.207	-0.183	-0.406	-0.429
	p-value	0.7577	0.8007	0.8363	0.855	0.6845	0.6679
Two One-Tailed Tests for Equivalence							
$\varepsilon = 1.645$ (expressed in units of z distribution)	z_1	1.336	1.393	1.852	1.828	2.051	2.074
$\alpha = 0.05$	p-value	0.0907	0.0819	0.032*	0.0338*	0.0201*	0.019*
	z_2	1.954	1.897	1.438	1.462	1.239	1.216
	p-value	0.0254*	0.0289*	0.0752	0.0718	0.1077	0.112

*Significant at $p < 0.05$

Discussion

The goal of this project was to determine if data interpolated from public use microdata files could approximate the distributions that are produced when a data user purchases data from the U.S. Census Bureau through the custom tabulation program. To accomplish this task there are abundant examples of data that can be useful to produce, but precious few opportunities to test the results against what I am considering gold standard data produced from the restricted use microdata files available to the U.S. Census Bureau through the ACS program.

In the course of my duties for the State of Michigan, I was asked to interpolate estimates from the PUMS data for use in redesigning a funding formula for an early childhood reading program, but I had significant reservations about using those estimates for the distribution of funds, so I recommended that the group purchase data from the U.S. Census Bureau as those would be the best data available to the group to answer the questions posed. The group responded positively to this suggestion and to date have purchased two sets of data from Census to accomplish the group's goals. These are data for the years of 2014 and 2015 and consist of estimates of four-year-olds who live in families whose income is less than or equal to 250 percent of the federal poverty level. With the permission of the group, I now had access to two datasets against which I could test my estimates procedures and determine if I could produce estimates that were equivalent to the custom tabulations.

I gained access to a third dataset through connections in the Michigan Office of Aging Services. That group periodically purchases data from the custom tabulation program for persons 60 years and over who are living in families at or below 150 percent

of the federal poverty level. These data were provided to me and permission was given to use them in this project. Those estimates were from the 2010, 5-year estimates and corresponded to the counties in the state. In addition to adding another set for which I can attempt make equivalent estimates, this set also provides me with the opportunity to see how the method compares to estimates produced from the ACS 5-year estimates.

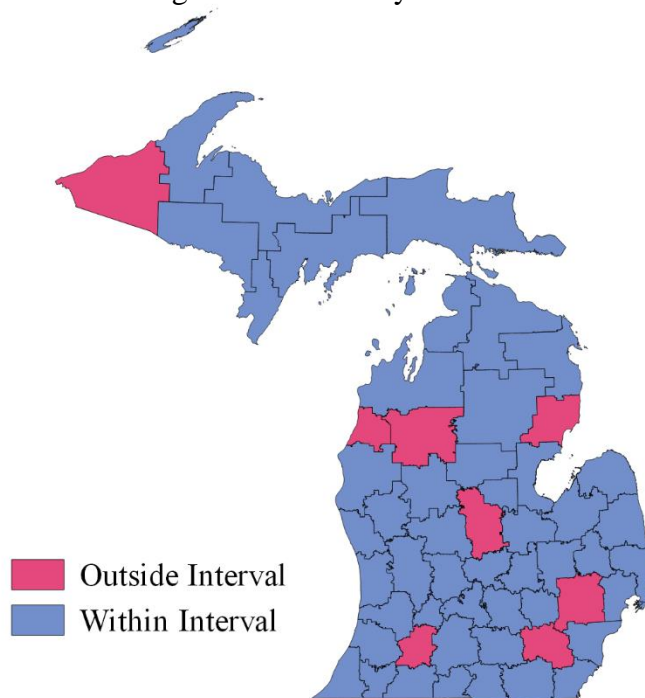
This project began with a request to interpolate data for the Michigan Department of Education (MDE). That request was time constrained and very important in terms of who was making the request and the data need. These factors meant it was a request that I could neither refuse nor ignore. The request was very specific (four-year-olds at or below 250% of poverty) and was not open to change to make it something capable of being answered with publicly available data. I told the group that I was not able to produce estimates of sufficient quality on which to base distribution of the hundreds of millions of dollars that MDE was charged with distributing through the Great Starts Reading Program (GSRP), and that the best option was to purchase estimates from the U.S. Census Bureau's custom tabulation program. The group accepted this recommendation but needed data to start redesigning the funding formula in advance of receiving the estimates. The custom tabulation program requires a minimum of two months to produce estimates, and often takes considerably longer. Given the confluence of these needs and constraints, I needed to devise a novel solution to allow MDE to continue their work while we waited for the final estimates that would be used to actually distribute the funds.

To fulfill the request from MDE, I devised a method for moving PUMS data to the ISD level through the intermediary geography of the Local Education Areas (LEA).

This was necessary because the ISD was not a standard Census geography but was made up of aggregations of the LEA level geography. The process I used was to areally weight the PUMS data by attaching a weighting variable to the LEA geography and then combining the LEA and PUMA geographies using GIS software. This allowed me to produce estimates from the PUMS data at the necessary geography for the required subpopulation. I was pleased with the technical aspects of my method, but I had no way to test the output from the method. The final delivery of the estimates from the custom tabulation request allowed for this testing. I was broadly pleased with the estimates and found that my estimates fell within the confidence interval for the census produced estimates 86 percent of the time. This was a good result, but I wanted to think of a way to make the estimates better.

In reviewing the estimates for 2014, I found that many of the areas that did not fall within the confidence intervals for my method were in the more rural parts of the

Figure 8 New Jersey Estimates



state as can be seen in the map below. The exception to this are the Oakland and Washtenaw districts in the southeastern part of the state and to a lesser extent the Kalamazoo district in the southwestern part of the state.

To improve the estimates, I devised a modification to the estimates procedure that distributes the weighting variables to the intermediate geography in two stages. The first stage of this distribution is to a smaller geography that exists within both the intermediate and the PUMS geographies. My thinking was that a better targeting of the population centers would produce better estimates as they would be less reliant on areal disaggregation. This is the method that was detailed earlier in this paper.

In the methodological explanation provided in the previous sections, I used the county subdivision geography as a first stage disaggregation geography because it fit well within the other geographies I was using, and it was of sufficient size as to actually be able to pull significant portions of the weighting variables. I wanted to use the largest geography I could in this stage because I was concerned about the multiplicative effect of distributing error across the first state geographies. I thought county subdivisions would work well because they were generally larger than something like census tracts or block groups, and they were able to differentiate portions of cities that crossed county lines. This was important because the PUMA boundaries generally respect county boundaries.

The multi-stage process I devised to produce the estimates made the process much more complicated, which is evidenced by the extensive data manipulation required shown in the Excel work detailed in the methodological description in Appendix C. Unfortunately, the added methodological complication did not actually improve the estimates. As shown in the testing section, trivial differences remain in both

implantations of the method, and the same geographies fall outside of the confidence intervals produced from the custom tabulations. The raked estimates actually performed a little worse as an additional school district fell outside of the confidence intervals for the 2014 estimates set.

The two estimate sets (both the original and revised methods) perform reasonably well and produce estimates with only trivial differences when compared to the purchased ACS data. Also, the clear majority of the time they both produce estimates that will fall within the confidence intervals produced by a custom tabulation. However, the increased complexity of the revised method does not seem to produce better estimates and may actually make the estimates less accurate.

There are myriad reasons why the estimates produced in this methodological exercise might not perfectly replicate the distribution produced through a custom tabulation of ACS data. First, we must remember that this project was never meant to produce a general method that would be capable of perfectly estimating ACS distributions. This method starts with the acknowledgement that what will be produced are weighted interpolations of distributions that would not otherwise be available without costly investments in terms of both money and time spent. If a researcher has the time and money to purchase custom tabulations from the Census Bureau to meet the data needs of a project, that would always be preferable to implementing this method to gather data. When that is not possible, this method provides an alternative to the “no data” situation, but the causes of error need to be in the forefront of a researcher’s mind.

The first source of error in implementing this method is simply user error. This process is complicated and there are many chances for the user to make a mistake that

will create error. During the course of this research I have probably replicated (through the course of refinement and description) the process at least 50 times. Through the course of those iterations, I have become quite proficient with the process, but I still find that I make mistakes that bring in error. I am able to catch those errors and correct them, but that is because I have done it so many times, and I know what to expect and how to spot errors. For example, the “unduplication” that is done in excel should not produce negative numbers, but often will if the formula is not implemented perfectly. This happened to me and both of my testers. Through the feedback of my testers, I have refined the instructions for that part of the process, so hopefully that will not be a sticking point for those trying to implement the process, but that is just one place where user error can derail the method. As mentioned in the usability testing section, both users were able to complete a set of estimates, but both users also needed guidance at points to help them at points where they became stuck. Those sticking points have been addressed in the methodological description above, but there are still ample opportunities for a user to go wrong and make a mistake that will result in inaccurate estimates.

Another source of error comes from low correlation between the population used for the weighting variable and the target population for which the estimates are being made. This method uses populations that are contained in the summary data available through the ACS tabulations to make estimates that are not contained in the summary data. This is inherently a problematic part of the method as you are using a population other than the one being studied to make estimates of the target population. As mentioned previously, efforts need to be made to increase the correlation as much as possible, but that correlation is something that is unknowable as implementation of this

method presupposes that there are not estimates of the target population available. To put it another way, if the data required to make an estimation of the correlation between the weighting variable data and the target population were available, the method would be unnecessary as the data the method is trying to estimate would be available.

Error with the school district geography may also come from the very irregular nature of the constituent polygons. School district boundaries are determined at the local level and therefore are amongst the most irregular geographies that exist in census geography, at least in the state of Michigan. In some areas the school districts are discontinuous and include some households while excluding those that are next door. Some of this irregularity reflects the desires of locals for the particular school district in which they wish to live, the need of school districts to maintain a tax base, and the sometimes-opaque machinations of local political debates. Regardless of the source of the irregular boundaries, they sometime shift populations in a manner for which an algorithmic interpolation method cannot account.

The irregular nature of the school district boundaries is something in particular that contributed to the lower performance of the revised, two-stage method. I came to this conclusion after reviewing the proportion of estimates that fall within the confidence intervals formed from the data provided with the custom tabulations. For the school district geographies, there was no improvement when implementing the revised, two-stage method, however when the revised method was applied to the county level geography there was considerable improvement. When looking at the school district level geography, the results between the two methods were nearly identical. The only difference was one school district was not within the confidence interval with the revised

method for the 2014 data that was with the original, areal method. When looking at the county level estimates, the revised method improved the estimates where 76 percent of the estimates fell within the confidence intervals created from the custom tabulations versus 71 percent for the original method.

The irregularity of the school districts is different from the irregularity that is caused by removing the areas of the state that would be under various portions of the Great Lakes. This also creates irregularities in the polygons, but it does not seem to be detrimental to the estimates. This, conceptually at least, is an element that would improve the estimates, and there was nothing in the review of the estimates that indicated it had a deleterious effect on the estimates.

The work of an applied demographer can be very frustrating at times because there is often a request made that does not appreciate the nature of the data being requested nor the lack of any accepted method for the production of what is being requested. Applied demographers are often left to their own devices and are forced to “make due” with whatever data they can get and make as best an estimate as time and resources allow. The exciting part about this project was the promise of gold standard data at the end against which my estimates could be tested. This project allowed me, as a researcher and a social scientist, to put all my skills into a project and then to test how well I did at the end. That is a rare occurrence in my work, and I am extremely grateful to have had the opportunity.

Conclusions

As with most projects in the areas of applied demography, this project started with a request for data that was outside of what was freely and readily available. The request came from an important area in state government and involved making decisions that affected the distribution of enormous amounts of money. The data need was very specific and could not be updated or modified to make it more answerable by publicly available data. In short, this project is definitely the type of project that required the applied demographer to devise a novel solution to fulfill the data need.

Given the importance of the program and the large amount of money being distributed, I was reluctant to produce estimates using an untested procedure. Luckily, given the size of the program, there were funds available to purchase estimates that would satisfy the member of the committee that were requesting data. In the course of making that request, some new issues arose that brought me back to the need to make estimates for the group. The committee needed data to begin working through their formula redesign in advance of the point where the custom tabulations from census would arrive, and the committee wanted to see the effects of different grouping and poverty scenarios. With the need present, I agreed to produce estimates with the understanding that they would not be used for the final funds distribution. During the course of producing these estimates, I also gained permission to use the custom tabulations to test how well my estimates approximated the custom tabulation.

With the availability of these datasets, I now had the ability to test my estimates against those produced by the U.S. Census Bureau. The results from these tests are seen in the table in the previous section and show that my interpolated estimates, while not

identical, are close enough to be useful for a variety of purposes. The intention of this project was never to create a process that could take the place of the custom tabulation programs, a project with that goal would not be likely to succeed. Rather, the goal was to develop and test a method to create estimates that could be used for a variety of purposes that would allow for greater availability for data and projects that would otherwise not be possible. I cannot create an exhaustive list of uses for these types of estimates, but some uses include triangulation of test or survey results, survey frame development, commuting analysis, and many others. These estimates, by themselves, are not precise enough for things like resource allocation.

The results of the testing of my original and revised methods are mixed and not necessarily consistent with my hopes, but this is often the case with work in the social sciences. I am pleased that my procedures—both the original and revised methods—are able to produce estimates that are similar to those produced by the custom tabulations programs. They are not perfectly equivalent, but they are close, and the differences can be said to be trivial. I would have liked to be able to produce estimates that were closer to the mark, but an algorithmic disaggregation of data using weighting variables that are not exactly what I want in the final data might not be able to come closer without additional data and techniques that were unavailable to me, or that would be unavailable without significant increases in cost.

I am disappointed that my revision to the method did not improve the estimates and may have made them worse. Given these results and the significant increases in complexity, it would be better for any implementation of this method to use the original, general areal disaggregation method rather than the two-stage weighting method. Wider

testing of the method may yet vindicate this procedure, but at present time it does not appear to improve the method. I believe that because there was not improvement with the decrease of areal disaggregation, the major source of error in the estimates is the distance of the data used for the weighting variable from the true values of the target population. This is an area that I plan to pursue with future research.

The method is widely applicable in that it can be used to produce estimates in a wide variety of places. I produced estimates for South Dakota and New Jersey to determine if there were peculiarities that made the method work in Michigan but not other places. I did not find any indications that the method would not work in other areas. The method worked well in both rural and urban settings.

The method is also available to a variety of researchers with basic GIS skills and a knowledge of Census data. I was pleased that my testers were able understand the method well enough to reproduce my school district estimates for Michigan. The testers that I enlisted to work with me and are economists by training, but that should not be held against them. They were able to reproduce the estimates from my instructions, but they went the extra step and provided feedback that allowed me to make the instructions more clear and concise. This was a welcome benefit that improved the quality of this project.

In the end, this project has described and validated a procedure that can be used to create estimates for a variety of projects, but that may not be precise enough for many uses. The data that was available to me for testing was not sufficient to be able to say what parts of the procedure may be flawed, but, as mentioned above, I suspect more error is introduced from the selection of the weighting variable than from the areal distribution of the data. That seemed to be confirmed by the lack of improvement in the estimates

through the two-stage method which was meant to reduce the areal distribution of the weighting variable.

APPENDIX

Appendix A – Estimates Produced for Usability Testing

New Jersey Population Between 5 and 17 at or Below 200 Percent of Poverty by School District

GEOID	School District	Estimate
3400004	The Chathams School District	108
3400008	Great Meadows Regional School District	125
3400009	Somerset Hills Regional School District	19
3400660	Absecon City School District	42
3400690	Alexandria Township School District	31
3400720	Allamuchy Township School District	57
3400750	Allendale Borough School District	176
3400769	South Hunterdon Regional School District	14
3400780	Allenhurst Borough School District	0
3400810	Alloway Township School District	43
3400840	Alpha Borough School District	142
3400870	Alpine Borough School District	117
3400900	Andover Regional School District	4
3400930	Asbury Park City School District	2,705
3400960	Atlantic City School District	4,592
3401020	Atlantic Highlands Borough School District	199
3401050	Audubon Borough School District	200
3401110	Avalon Borough School District	25
3401140	Avon-by-the-Sea Borough School District	15
3401170	Barrington Borough School District	0
3401200	Bass River Township School District	172
3401230	Bay Head Borough School District	37
3401260	Bayonne City School District	6,200
3401290	Beach Haven Borough School District	11
3401320	Bedminster Township School District	54
3401350	Belleville Town School District	2,135
3401380	Bellmawr Borough School District	1,013
3401410	Belmar Borough School District	160
3401440	Belvidere Town School District	91
3401500	Bergenfield Borough School District	844
3401530	Berkeley Heights Township School District	190
3401560	Berkeley Township School District	923
3401590	Berlin Borough School District	103
3401620	Berlin Township School District	192
3401650	Bernards Township School District	138
3401710	Bethlehem Township School District	146

3401740	Beverly City School District	145
3401800	Blairstown Township School District	32
3401830	Bloomfield Township School District	2,442
3401860	Bloomington Borough School District	76
3401890	Bloomsbury Borough School District	26
3401920	Bogota Borough School District	366
3401950	Boonton Town School District	85
3401980	Boonton Township School District	0
3402030	Bordentown Regional School District	257
3402100	Bound Brook Borough School District	616
3402130	Bradley Beach Borough School District	95
3402160	Branchburg Township School District	127
3402220	Brick Township School District	1,672
3402250	Bridgeton City School District	3,392
3402280	Bridgewater-Raritan Regional School District	759
3402310	Brielle Borough School District	12
3402340	Brigantine City School District	275
3402370	Brooklawn Borough School District	190
3402400	Buena Regional School District	1,010
3402430	Burlington City School District	726
3402460	Burlington Township School District	723
3402520	Butler Borough School District	47
3402550	Byram Township School District	21
3402580	Caldwell-West Caldwell School District	252
3402610	Califon Borough School District	10
3402640	Camden City School District	13,577
3402700	Cape May City School District	169
3402760	Cape May Point Borough School District	2
3402790	Carlstadt Borough School District	55
3402820	Carteret Borough School District	2,456
3402850	Cedar Grove Township School District	155
3403000	Cherry Hill Township School District	892
3403030	Chesilhurst Borough School District	56
3403060	Chester Township School District	33
3403090	Chesterfield Township School District	21
3403120	Cinnaminson Township School District	522
3403150	Clark Township School District	304
3403180	Clayton Borough School District	433
3403240	Clementon Borough School District	444
3403270	Cliffside Park Borough School District	1,090
3403300	Clifton City School District	3,175
3403330	Clinton Town-Glen Gardner School District	91
3403360	Clinton Township School District	139
3403390	Closter Borough School District	104
3403420	Collingswood Borough School District	402

3403450	Colts Neck Township School District	425
3403480	Commercial Township School District	588
3403510	Corbin City School District	24
3403540	Cranbury Township School District	0
3403570	Cranford Township School District	64
3403600	Cresskill Borough School District	174
3403630	Deal Borough School District	28
3403660	Deerfield Township School District	86
3403690	Delanco Township School District	0
3403720	Delaware Township School District	0
3403780	Delran Township School District	235
3403810	Demarest Borough School District	0
3403840	Dennis Township School District	566
3403870	Denville Township School District	266
3403900	Deptford Township School District	988
3403930	Dover Town School District	1,795
3403960	Downe Township School District	22
3403990	Dumont Borough School District	474
3404020	Dunellen Borough School District	295
3404050	Eagleswood Township School District	4
3404080	East Amwell Township School District	64
3404110	East Brunswick Township School District	1,555
3404140	East Greenwich Township School District	329
3404170	East Hanover Township School District	449
3404200	East Newark Borough School District	603
3404230	East Orange City School District	6,196
3404290	East Rutherford Borough School District	432
3404320	East Windsor Regional School District	452
3404350	Eastampton Township School District	151
3404410	Eatontown Borough School District	339
3404440	Edgewater Borough School District	777
3404470	Edgewater Park Township School District	216
3404500	Edison Township School District	1,965
3404530	Egg Harbor City School District	299
3404560	Egg Harbor Township School District	2,318
3404590	Elizabeth City School District	14,451
3404620	Elk Township School District	34
3404650	Elmer Borough School District	113
3404660	Elmwood Park Borough School District	1,211
3404680	Elsinboro Township School District	38
3404710	Emerson Borough School District	150
3404740	Englewood City School District	1,093
3404770	Englewood Cliffs Borough School District	35
3404830	Essex Fells Borough School District	32
3404860	Estell Manor City School District	41

3404890	Evesham Township School District	1,483
3404920	Ewing Township School District	803
3404950	Fair Haven Borough School District	27
3404980	Fair Lawn Borough School District	468
3405010	Fairfield Township School District	0
3405040	Fairfield Township School District	322
3405070	Fairview Borough School District	743
3405130	Farmingdale Borough School District	159
3405190	Flemington-Raritan Regional School District	1,201
3405220	Florence Township School District	279
3405250	Florham Park Borough School District	119
3405280	Folsom Borough School District	41
3405310	Fort Lee Borough School District	771
3405340	Frankford Township School District	42
3405370	Franklin Lakes Borough School District	121
3405400	Franklin Borough School District	60
3405430	Franklin Township School District	593
3405460	Franklin Township School District	43
3405490	Franklin Township School District	1,353
3405520	Franklin Township School District	6
3405550	Fredon Township School District	170
3405580	Freehold Borough School District	764
3405640	Freehold Township School District	515
3405670	Frelinghuysen Township School District	13
3405700	Frenchtown Borough School District	36
3405730	Galloway Township School District	1,083
3405760	Garfield City School District	2,968
3405790	Garwood Borough School District	19
3405850	Gibbsboro Borough School District	9
3405880	Glassboro Borough School District	1,038
3405940	Glen Ridge Borough School District	0
3405970	Glen Rock Borough School District	78
3406000	Gloucester City School District	385
3406030	Gloucester Township School District	2,738
3406090	Green Township School District	12
3406120	Green Brook Township School District	173
3406150	Greenwich Township School District	24
3406180	Greenwich Township School District	248
3406210	Greenwich Township School District	128
3406240	Guttenberg Town School District	1,040
3406270	Hackensack City School District	1,833
3406300	Hackettstown Town School District	423
3406330	Haddon Heights Borough School District	38
3406360	Haddon Township School District	207
3406390	Haddonfield Borough School District	110

3406420	Hainesport Township School District	371
3406450	Haledon Borough School District	1,171
3406480	Hamburg Borough School District	42
3406510	Hamilton Township School District	850
3406540	Hamilton Township School District	4,073
3406570	Hammonton Town School District	746
3406600	Hampton Borough School District	24
3406630	Hampton Township School District	39
3406690	Hanover Township School District	174
3406720	Harding Township School District	80
3406780	Hardyston Township School District	0
3406810	Harmony Township School District	55
3406840	Harrington Park Borough School District	146
3406870	Harrison Town School District	1,664
3406900	Harrison Township School District	79
3406930	Hasbrouck Heights Borough School District	319
3406960	Haworth Borough School District	32
3406990	Hawthorne Borough School District	633
3407080	Hi-Nella Borough School District	24
3407110	High Bridge Borough School District	26
3407170	Highland Park Borough School District	443
3407200	Highlands Borough School District	357
3407230	Hillsborough Township School District	330
3407260	Hillsdale Borough School District	435
3407290	Hillside Township School District	991
3407320	Ho-Ho-Kus Borough School District	0
3407350	Hoboken City School District	999
3407380	Holland Township School District	107
3407410	Holmdel Township School District	369
3407440	Hopatcong Borough School District	601
3407470	Hope Township School District	53
3407500	Hopewell Township School District	105
3407530	Hopewell Valley Regional School District	62
3407560	Howell Township School District	1,050
3407650	Interlaken Borough School District	0
3407680	Irvington Township School District	7,222
3407710	Island Heights Borough School District	33
3407740	Jackson Township School District	1,909
3407770	Jamesburg Borough School District	509
3407800	Jefferson Township School District	2,452
3407830	Jersey City School District	17,696
3407860	Keansburg Borough School District	443
3407890	Kearny Town School District	3,587
3407920	Kenilworth Borough School District	0
3407950	Keyport Borough School District	449

3408010	Kingwood Township School District	38
3408040	Kinnelon Borough School District	198
3408070	Knowlton Township School District	164
3408100	Lacey Township School District	1,160
3408130	Lafayette Township School District	40
3408160	Lakehurst Borough School District	252
3408220	Lakewood Township School District	22,255
3408280	Laurel Springs Borough School District	68
3408310	Lavallette Borough School District	57
3408340	Lawnside Borough School District	316
3408370	Lawrence Township School District	190
3408400	Lawrence Township School District	493
3408430	Lebanon Borough School District	29
3408460	Lebanon Township School District	41
3408520	Leonia Borough School District	407
3408580	Lincoln Park Borough School District	143
3408610	Linden City School District	2,464
3408640	Lindenwold Borough School District	1,147
3408670	Linwood City School District	95
3408671	Longport Borough School District	0
3408700	Little Egg Harbor Township School District	674
3408730	Little Falls Township School District	495
3408760	Little Ferry Borough School District	641
3408790	Little Silver Borough School District	160
3408820	Livingston Township School District	303
3408850	Lodi Borough School District	2,287
3408880	Logan Township School District	304
3408910	Long Beach Island School District	57
3408940	Long Branch City School District	3,173
3409000	Lopatcong Township School District	327
3409030	Lower Alloways Creek Township School District	103
3409120	Pennsville Township School District	685
3409150	Lower Township School District	1,065
3409180	Lumberton Township School District	409
3409210	Lyndhurst Township School District	577
3409240	Madison Borough School District	400
3409270	Old Bridge Township School District	887
3409300	Magnolia Borough School District	273
3409330	Mahwah Township School District	77
3409390	Manalapan-Englishtown Regional School District	403
3409420	Manasquan Borough School District	97
3409450	Manchester Township School District	1,543
3409480	Mannington Township School District	93
3409510	Mansfield Township School District	0
3409540	Mansfield Township School District	62

3409600	Mantua Township School District	377
3409630	Manville Borough School District	690
3409660	Maple Shade Township School District	857
3409690	Margate City School District	42
3409720	Marlboro Township School District	233
3409750	Matawan-Aberdeen Regional School District	845
3409780	Maurice River Township School District	32
3409810	Maywood Borough School District	152
3409840	Medford Lakes Borough School District	0
3409870	Medford Township School District	158
3409900	Mendham Borough School District	51
3409930	Mendham Township School District	451
3409960	Merchantville Borough School District	0
3409990	Metuchen Borough School District	58
3410020	Middle Township School District	463
3410050	Middlesex Borough School District	141
3410110	Middletown Township School District	1,340
3410140	Midland Park Borough School District	314
3410170	Milford Borough School District	0
3410200	Millburn Township School District	609
3410230	Millstone Township School District	202
3410290	Milltown Borough School District	233
3410320	Millville City School District	3,097
3410350	Mine Hill Township School District	71
3410380	Monmouth Beach Borough School District	63
3410470	Monroe Township School District	1,932
3410500	Monroe Township School District	379
3410530	Montague Township School District	130
3410560	Montclair Town School District	891
3410590	Montgomery Township School District	588
3410620	Montvale Borough School District	623
3410650	Montville Township School District	199
3410680	Moonachie Borough School District	129
3410710	Moorestown Township School District	336
3410770	Morris Plains Borough School District	44
3410810	Morris Township School District	1,877
3410860	Mount Arlington Borough School District	0
3410890	Mount Ephraim Borough School District	189
3410920	Mount Holly Township School District	612
3410950	Mount Laurel Township School District	1,381
3410980	Mount Olive Township School District	712
3411010	Mountain Lakes Borough School District	93
3411040	Mountainside Borough School District	174
3411070	Mullica Township School District	141
3411100	National Park Borough School District	110

3411130	Neptune City School District	309
3411160	Neptune Township School District	840
3411190	Netcong Borough School District	291
3411220	New Brunswick City School District	5,144
3411250	New Hanover Township School District	261
3411280	New Milford Borough School District	678
3411310	New Providence Borough School District	63
3411340	Newark City School District	34,181
3411370	Newfield Elementary School District	84
3411400	Newton Town School District	392
3411430	North Arlington Borough School District	764
3411460	North Bergen Township School District	4,873
3411490	North Brunswick Township School District	1,580
3411520	North Caldwell Borough School District	105
3411550	North Haledon Borough School District	277
3411580	North Hanover Township School District	347
3411640	North Plainfield Borough School District	1,472
3411670	North Wildwood City School District	171
3411790	Northfield City School District	807
3411820	Northvale Borough School District	97
3411850	Norwood Borough School District	141
3411880	Nutley Town School District	696
3411910	Oakland Borough School District	63
3411940	Oaklyn Borough School District	54
3411970	Ocean City School District	251
3412030	Ocean Gate Borough School District	118
3412060	Ocean Township School District	1,354
3412090	Ocean Township School District	67
3412120	Oceanport Borough School District	494
3412150	Ogdensburg Borough School District	70
3412180	Old Tappan Borough School District	223
3412210	Oldmans Township School District	38
3412240	Oradell Borough School District	20
3412270	Orange City Township School District	4,078
3412300	Oxford Township School District	0
3412360	Palisades Park Borough School District	130
3412390	Palmyra Borough School District	260
3412420	Paramus Borough School District	476
3412450	Park Ridge Borough School District	45
3412480	Parsippany-Troy Hills Township School District	676
3412540	Passaic City School District	11,636
3412660	Long Hill Township School District	174
3412690	Paterson City School District	21,445
3412720	Paulsboro Borough School District	1,970
3412810	Pemberton Township School District	3,285

3412840	Penns Grove-Carneys Point Regional School District	1,233
3412870	Pennsauken Township School District	1,016
3412900	Pequannock Township School District	0
3412930	Perth Amboy City School District	8,007
3412960	Phillipsburg Town School District	1,714
3412990	Pine Hill Borough School District	414
3413020	Pine Valley Borough School District	0
3413050	Piscataway Township School District	2,483
3413080	Pitman Borough School District	203
3413110	Pittsgrove Township School District	210
3413140	Plainfield City School District	6,339
3413200	Pleasantville City School District	3,302
3413230	Plumsted Township School District	262
3413260	Pohatcong Township School District	165
3413290	Point Pleasant Borough School District	432
3413320	Point Pleasant Beach Borough School District	165
3413350	Pompton Lakes Borough School District	578
3413380	Port Republic City School District	7
3413410	Princeton Public Schools	269
3413470	Prospect Park Borough School District	853
3413500	Quinton Township School District	82
3413530	Rahway City School District	1,632
3413590	Ramsey Borough School District	159
3413650	Randolph Township School District	294
3413680	Hazlet Township School District	398
3413710	Readington Township School District	542
3413740	Red Bank Borough School District	766
3413770	Ridgefield Borough School District	175
3413800	Ridgefield Park Township School District	470
3413830	Ridgewood Village School District	579
3413860	Ringwood Borough School District	97
3413890	River Edge Borough School District	131
3413950	River Vale Township School District	144
3413980	Riverdale Borough School District	177
3414010	Riverside Township School District	370
3414040	Riverton Borough School District	32
3414070	Rochelle Park Township School District	69
3414100	Rockaway Borough School District	89
3414130	Rockaway Township School District	604
3414160	Rockleigh Borough School District	10
3414220	Roosevelt Borough School District	11
3414250	Roseland Borough School District	0
3414280	Roselle Borough School District	2,574
3414310	Roselle Park Borough School District	366
3414340	Roxbury Township School District	576

3414370	Rumson Borough School District	160
3414430	Runnemede Borough School District	420
3414460	Rutherford Borough School District	419
3414490	Saddle Brook Township School District	275
3414520	Saddle River Borough School District	0
3414550	Salem City School District	744
3414610	Sandyston-Walpack Township School District	37
3414640	Sayreville Borough School District	1,915
3414670	Scotch Plains-Fanwood Regional School District	159
3414730	Sea Girt Borough School District	20
3414760	Sea Isle City School District	0
3414790	Seaside Heights Borough School District	325
3414820	Seaside Park Borough School District	9
3414850	Secaucus Town School District	560
3414880	Shamong Township School District	179
3414970	Shrewsbury Borough School District	47
3415000	Somerdale Borough School District	81
3415030	Somers Point City School District	617
3415090	Somerville Borough School District	341
3415120	South Amboy City School District	159
3415150	Lake Como Borough School District	97
3415180	South Bound Brook Borough School District	204
3415210	South Brunswick Township School District	1,538
3415240	South Hackensack Township School District	239
3415270	South Harrison Township School District	0
3415330	South Orange-Maplewood School District	1,092
3415360	South Plainfield Borough School District	617
3415390	South River Borough School District	1,044
3415420	Southampton Township School District	117
3415510	Sparta Township School District	222
3415540	Spotswood Borough School District	200
3415570	Spring Lake Borough School District	53
3415600	Spring Lake Heights Borough School District	0
3415630	Springfield Township School District	195
3415660	Springfield Township School District	454
3415690	Stafford Township School District	546
3415720	Stanhope Borough School District	184
3415750	Stillwater Township School District	149
3415810	Stone Harbor Borough School District	4
3415840	Stow Creek Township School District	9
3415870	Stratford Borough School District	422
3415900	Summit City School District	365
3415960	Sussex-Wantage Regional School District	317
3415990	Swedesboro-Woolwich School District	619
3416020	Tabernacle Township School District	93

3416080	Teaneck Township School District	1,184
3416110	Tenaflly Borough School District	486
3416170	Tewksbury Township School District	0
3416200	Tinton Falls Borough School District	467
3416230	Toms River Regional School District	2,737
3416260	Totowa Borough School District	174
3416290	Trenton City School District	10,786
3416320	Tuckerton Borough School District	48
3416350	Union Beach Borough School District	345
3416380	Union City School District	6,127
3416440	Union Township School District	45
3416470	Barnegat Township School District	894
3416500	Union Township School District	2,032
3416530	Upper Deerfield Township School District	249
3416560	Upper Freehold Regional School District	29
3416590	Upper Pittsgrove Township School District	86
3416620	Upper Saddle River Borough School District	116
3416650	Upper Township School District	93
3416680	Ventnor City School District	330
3416710	Vernon Township School District	392
3416740	Verona Borough School District	286
3416800	Vineland City School District	5,853
3416830	Voorhees Township School District	168
3416860	Waldwick Borough School District	486
3416890	Wall Township School District	738
3416920	Wallington Borough School District	1,422
3416950	Wanaque Borough School District	85
3416980	Warren Township School District	67
3417010	Washington Borough School District	651
3417040	Washington Township School District	45
3417070	Washington Township School District	752
3417100	Robbinsville Township School District	0
3417130	Washington Township School District	197
3417160	Washington Township School District	30
3417190	Watchung Borough School District	74
3417250	Waterford Township School District	241
3417280	Wayne Township School District	796
3417310	Weehawken Township School District	189
3417340	Wenonah Borough School District	33
3417400	West Cape May Borough School District	84
3417430	West Deptford Township School District	465
3417490	West Long Branch Borough School District	55
3417520	West Milford Township School District	365
3417580	West New York Town School District	4,942
3417610	West Orange Town School District	1,514

3417640	Woodland Park Borough School District	112
3417670	West Wildwood Borough School District	0
3417700	West Windsor-Plainsboro Regional School District	637
3417730	Westampton Township School District	48
3417760	Westfield Town School District	600
3417790	Westville Borough School District	161
3417820	Westwood Regional School District	288
3417850	Weymouth Township School District	39
3417880	Wharton Borough School District	270
3417910	White Township School District	78
3417940	Wildwood City School District	748
3417970	Wildwood Crest Borough School District	33
3418000	Willingboro Township School District	2,463
3418030	Winfield Township School District	167
3418060	Winslow Township School District	2,012
3418090	Woodbine Borough School District	146
3418120	Woodbridge Township School District	3,468
3418150	Woodbury City School District	1,521
3418180	Woodbury Heights Borough School District	9
3418210	Woodcliff Lake Borough School District	0
3418240	Woodland Township School District	28
3418270	Woodlynne Borough School District	685
3418300	Wood-Ridge Borough School District	306
3418330	Woodstown-Pilesgrove Regional School District	419
3418360	Wyckoff Township School District	205
3434001	Joint Base McGuire-Dix-Lakehurst	93
3499997	School District Not Defined	0
Grand Total		451,429

South Dakota Males Between 26 and 32 Who are Employed by County

GEOID	County	Estimate
46003	Aurora	89
46005	Beadle	595
46007	Bennett	87
46009	Bon Homme	130
46011	Brookings	1,268
46013	Brown	1,663
46015	Brule	139
46017	Buffalo	59
46019	Butte	390
46021	Campbell	42
46023	Charles Mix	253
46025	Clark	111
46027	Clay	542
46029	Codington	1,235
46031	Corson	95
46033	Custer	243
46035	Davison	694
46037	Day	174
46039	Deuel	144
46041	Dewey	159
46043	Douglas	73
46045	Edmunds	85
46047	Fall River	184
46049	Faulk	69
46051	Grant	217
46053	Gregory	124
46055	Haakon	71
46057	Hamlin	208
46059	Hand	92
46061	Hanson	94
46063	Harding	48
46065	Hughes	773
46067	Hutchinson	188
46069	Hyde	43
46071	Jackson	114
46073	Jerauld	55
46075	Jones	24
46077	Kingsbury	151
46079	Lake	331
46081	Lawrence	1,003
46083	Lincoln	3,113
46085	Lyman	133

46087	McCook	221
46089	McPherson	42
46091	Marshall	187
46093	Meade	1,006
46095	Mellette	54
46097	Miner	63
46099	Minnehaha	9,927
46101	Moody	186
46102	Oglala Lakota	269
46103	Pennington	4,289
46105	Perkins	100
46107	Potter	71
46109	Roberts	262
46111	Sanborn	71
46115	Spink	240
46117	Stanley	120
46119	Sully	51
46121	Todd	253
46123	Tripp	150
46125	Turner	315
46127	Union	592
46129	Walworth	168
46135	Yankton	947
46137	Ziebach	59
Grand Total		34,949

Appendix B – Estimates Used in Validity Testing

ISD	2014					2015				
	ACS Custom Tabulation	Original Interpolation Method (Areal)		Revised Method (Two-Stage)		ACS Custom Tabulation	Original Interpolation Method (Areal)		Revised Method (Two-Stage)	
		Initial	Raked	Initial	Ranked		Initial	Raked	Initial	Ranked
3	355	437	446	418	426	865	1,117	1,127	1,134	1,144
4	175	159	162	165	168	(X)	203	(X)	206	(X)
8	335	114	116	113	115	(X)	162	(X)	161	(X)
9	1,040	942	961	955	973	145	297	299	303	306
11	1,330	1,193	1,217	1,074	1,095	935	1,287	1,298	1,249	1,261
12	395	332	338	332	338	(X)	197	(X)	197	(X)
13	995	970	989	971	990	1,255	1,258	1,270	1,253	1,264
14	365	328	334	366	373	(X)	140	(X)	153	(X)
15	405	368	376	380	387	(X)	348	(X)	357	(X)
16	440	319	325	325	331	(X)	483	(X)	486	(X)
17	595	639	652	639	652	(X)	449	(X)	449	(X)
18	400	358	365	340	347	(X)	298	(X)	290	(X)
19	280	103	105	103	105	(X)	151	(X)	148	(X)
21	330	450	459	452	461	(X)	293	(X)	295	(X)
22	115	133	136	131	133	(X)	75	(X)	73	(X)
23	285	177	181	160	164	420	220	222	221	223
25	3,530	3,285	3,350	3,276	3,341	2,870	3,119	3,147	3,107	3,137
27	325	131	133	131	133	(X)	55	(X)	55	(X)
28	660	840	857	848	864	1,020	1,013	1,023	1,020	1,030
29	705	409	417	378	385	905	400	404	399	403
30	330	286	292	280	286	(X)	300	(X)	297	(X)
31	210	190	194	190	194	(X)	116	(X)	116	(X)
32	185	247	252	239	244	(X)	217	(X)	212	(X)
33	2,565	3,203	3,267	3,229	3,293	1,610	1,721	1,736	1,718	1,734
34	485	428	436	410	418	635	253	255	237	239
35	110	225	230	214	218	(X)	170	(X)	161	(X)
38	410	524	534	530	540	810	1,171	1,182	1,186	1,197
39	2,210	3,165	3,227	3,193	3,257	1,750	1,933	1,951	1,944	1,962
41	3,840	4,545	4,635	4,583	4,674	4,155	3,264	3,293	3,304	3,335
44	610	433	442	461	470	240	441	444	466	471
46	335	342	349	339	346	525	395	399	392	396
47	495	536	547	536	546	470	325	328	325	329
50	4,430	4,699	4,792	4,660	4,753	4,865	4,604	4,645	4,569	4,612
51	45	156	159	157	160	(X)	165	(X)	167	(X)
52	300	262	267	262	267	255	158	159	158	159
53	445	509	519	534	545	(X)	340	(X)	357	(X)
54	770	595	607	535	545	415	375	378	346	349
55	115	208	212	208	212	(X)	121	(X)	121	(X)
56	465	473	483	493	503	330	122	123	122	123
58	975	1,009	1,029	1,011	1,031	170	122	123	122	123
59	635	495	504	471	480	535	374	378	342	345
61	1,495	1,280	1,305	1,274	1,299	930	984	993	977	987
62	685	625	637	657	670	(X)	342	(X)	360	(X)
63	4,140	5,087	5,188	5,090	5,191	5,510	5,729	5,780	5,736	5,790
64	340	363	370	381	388	(X)	213	(X)	223	(X)
70	1,825	1,232	1,256	1,222	1,247	1,710	1,710	1,725	1,689	1,705
72	340	275	280	259	264	(X)	293	(X)	272	(X)
73	1,365	1,497	1,526	1,486	1,516	1,110	1,304	1,315	1,294	1,306
74	1,285	1,294	1,319	1,339	1,365	1,195	1,073	1,083	1,112	1,123
75	550	631	644	635	648	610	373	376	375	378
76	320	355	362	347	354	(X)	340	(X)	335	(X)
78	495	269	274	261	266	465	388	391	386	390
79	470	363	371	355	362	(X)	430	(X)	423	(X)
80	625	906	924	991	1,011	290	539	544	563	568
81	1,670	970	989	956	975	1,720	1,709	1,725	1,704	1,720
82	17,660	15,445	15,751	15,453	15,759	15,850	16,306	16,453	16,307	16,460
83	355	541	552	552	563	(X)	565	(X)	572	(X)

Appendix C – The Revised Method in Detail

Reviewing the Geographies

The first step in making an estimate for an alternate geography and/or subpopulation is to determine the geographic constraints of the project. Projects vary widely in terms of the geographic specificity required, so having a complete understanding of these constraints will not only save time when the estimates are being made, but it will also make sure time is not wasted in making estimates that are not part of the study area and that sufficient geographic coverage exists to make the required estimates.

When we take the project that I was presented with that started this work, the project for the MDE, the end result needed were estimates for all Intermediate School Districts (ISD) in the state. Upon first inspection of the required geographies, I became concerned as I noticed that what MDE was referring to as ISDs were not part of the geographic levels that were tabulated by the Census Bureau. This required further investigation, as I would have to see how close I could get to the required geography based on the levels provided in the summary data. There are two basic approaches that I used when trying to find geographic units that would serve as a mechanism to make estimates. One approach would be to find the smallest area that would be able to aggregate to the required geography, and the other would be to find the largest geography that would aggregate to the required level. I generally look for the largest area because all of the data I will be working with are estimate, which have associated error. Using the largest areas possible will mean that I am using fewer areas, and therefore introducing less error.

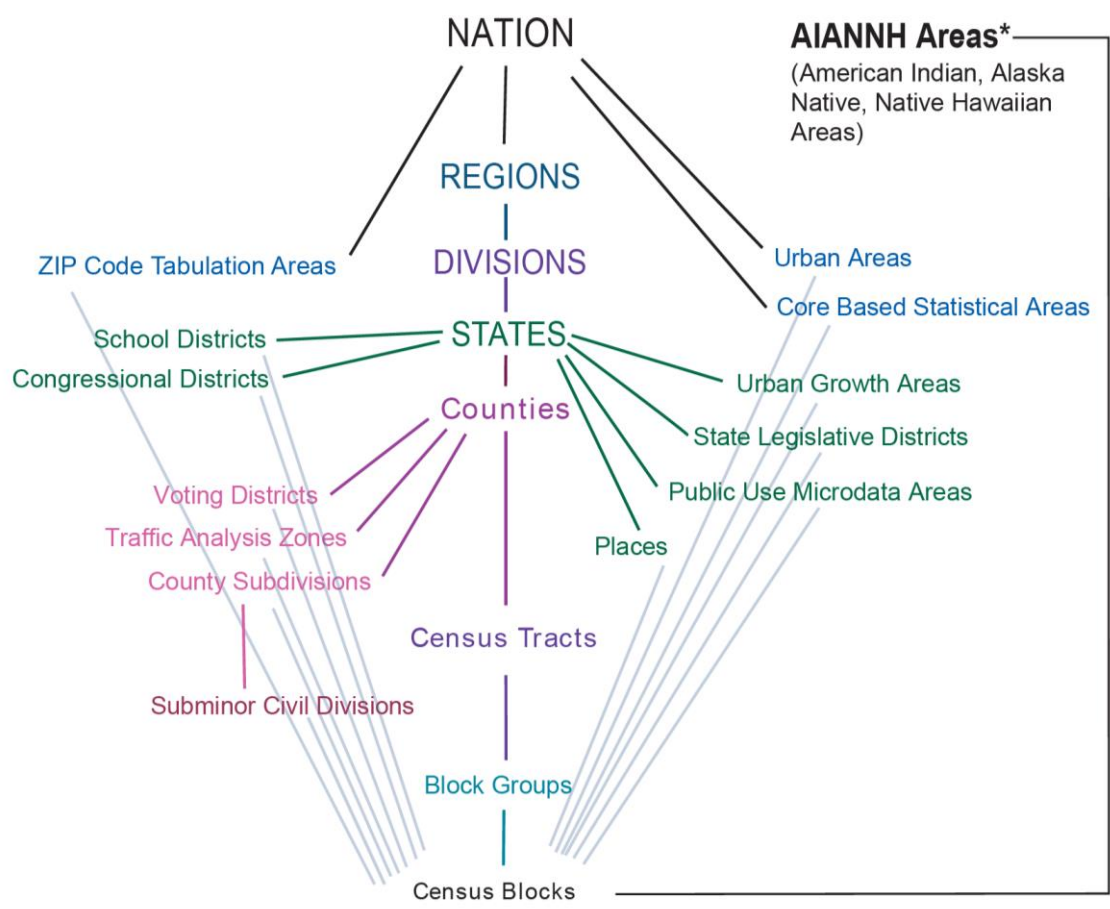
If it made sense to look for the smallest area that would be able to aggregate the study area, I would look at the block group level and begin to work up from there. Block groups are the smallest geographic level available for survey data from the Census Bureau and are made up of various numbers of census blocks which represent the smallest geographic level for which data from the decennial counts are available. A census block is what most people think of when they think of the block in which they live. It is generally, especially in urban areas, a piece of land that is surrounded by a road. As areas become more sparsely populated, these boundaries are sometimes other natural or manmade lines, such as rivers, lakes, railroad tracks or field lines. Blocks in urban areas tend to be very small and grow as population decreases and consequently the number of manmade boundaries decreases. Census blocks do not have a set or average population threshold, but block groups are generally groups of blocks that are within the same census tract whose populations total between 600 and 3,000 persons. These are not strict limits for block groups but are averages that allow for data privacy standards to be maintained. Disaggregating the data or a weighting variable would allow me to aggregate the data to any geographic level that is above it in the Census Bureau geography hierarchy below (Figure C-1). Working from the block group level would create two problems, one specific to the request from MDE and one more general that would affect any procedure or application of the process I am describing.

The first and the problem specific to the request from MDE involves the unique nature of school districts in census geography. Census geographies on the central line in Figure 1 all nest within the geography that is directly above it. So, census blocks nest within block groups, and block groups nest within census tracts etc., until you reach the

largest geographic unit, the nation. School districts do not exist on that central line and are instead on a diagonal and aggregations to the school district level are only possible from the census block level. This means that I would not be able to cleanly aggregate to the school district level from the block group level. This is enough to make the block group and all geographies above it inappropriate for use in the project, but there is another problem that could result from aggregating from a small level, if it were possible.

The second problem, which would affect any such aggregation, is the nature of census survey data and all survey data for that matter. Data collected from a survey and weighted to represent the total population in question always have associated margins of

Figure C-1 Standard Hierarchy of Census Geographic Entities



Source: U.S. Census Bureau

error, as noted above. When aggregating many geographies at from a small level all with associated margins of error, some quite significant, the resultant estimates can vary significantly from the true values due to the multiplicative effect of combining numerous estimates with wide confidence intervals. Given these issues and the issue discussed above, choosing a small geography from which to aggregate was not possible.

One caveat to the discussion above regarding census geography is the placement of the PUMA geography on a diagonal linking to the Census Block. The instructions (U.S. Census, 2011) that were given to produce the PUMA geographies in 2011, which were done by the State Data Center (SDC) agencies, were to build contiguous areas that contained at least 100,000 people out of the Census Tract geography. That means that the diagonal that links the PUMA geography to its source geography should really lead to the Census Tract rather than the Census Block. I know from experience that there is perfect correspondence between tracts and PUMAs in the states of Michigan and South Dakota. I have worked extensively with both geographies in Michigan, and I personally drafted the PUMA areas in South Dakota. It is entirely possible that there may be small variations in certain parts of the nation that made the link between the PUMA and Census Block necessary for purely technical reasons, but in practice most researchers who are investigating topics as a geographic level at or above the census tract could build a perfect correspondence table and make weighted estimates with no need for areal interpolation. This is an important note, because that places the county level geography on the list of geographies where this would be possible.

I will discuss the process for producing county level estimates from the PUMA data briefly after the discussion of the School District level geography. This discussion is

ultimately a simplification of the method as it can be done almost entirely in a spreadsheet, requires very little to no GIS work, and no areal interpolation.

Determine the Requirements

When considering the geographic entities that were available from the Census Bureau, I decided to see how the geographies listed as school districts related to the ISD that was being requested by MDE. To get an exhaustive accounting of the land area in the state, two different school geographic units were necessary from the census geography, which were the “elementary” and “unified” school districts. The combination of these two geographic units make up the total listing of the state’s Local Education Agencies (LEA), which also form the building blocks of the state’s ISDs. That meant if I could get data to the LEA level, I could sum it to the ISD level. This was the link that I needed to bridge my requested geography to the census geography. The problem with the LEA geography as a bridging geography is it crosses the boundaries of the Public Use Microdata Areas (PUMA) which means I would need to disaggregate some of the data areally across those spaces that existed in two different PUMAs. This created an opportunity for significant error to enter in to the process, especially in the rural areas of the state where large swaths of the LEAs are sparsely populated.

Selecting geographies is a vital step that will be the starting point of any project of this nature. If the target geography is not a census geography you will need to select an appropriate bridging geography. When making that selection there will be choices to be made and any of these choices will be an opportunity for the introduction of error into the process. A goal of reviewing the geographies is to select a geographic unit as a target or bridging geography where the smallest amount of error will be introduced. This is

especially problematic in a process like this where the only reason to engage in this process is because of a lack of data. The researcher has little or no data to verify the end estimates. Thus, the need for careful and meticulous thought in the early parts of the process.

Determine a Weighting Variable

The purpose of the weighting variable is to determine how much of the total subpopulation from the PUMS to attribute to each polygon created by merging the PUMA geographies with the target or bridging geographic areas. The goal of selecting a weighting variable should be to pick one that most closely correlates with the subpopulation for which you are trying to make an estimate. It may be possible to use a variable from the summary data that represents your entire subpopulation, though that would be unlikely, and would only occur if the researchers were making estimates for a standard population or subpopulation of a user defined geography. The more likely scenario would be having to select from a variety of imperfect matches available in the summary data.

Taking the example of the request from MDE, the need was for a variable that I could use to pull data from the PUMS data to the polygons created from the merging of the PUMA and LEA geographies. The final set of estimates I produced for the MDE project used children 0 to 4 years of age who are at or below 100 percent of poverty, what I call early childhood poverty. My first attempt at this process used total population as a weight, but the results from that were unsatisfactory considering the knowledge and literature that point to poverty being distributed in a manner different from the general population distribution (Ranjith and Rupasingha 2012). That value of the exercise with

the total population was more to prove the process worked and could be used to transform the data from the PUMA level to another geographic level. To arrive at that as a weighting variable, I generated several sets of estimates, each using a different variable from the summary data to serve as the weighting variable.

The first set of alternate estimates used the general population as a weighting variable. This produced a set of data with a different distribution than the first data set. The distribution also made more sense in general when looked at in comparison with the set produced with the general population as a weight. My overall impression with this set was the process of refining the weighting variable was having a positive impact on the final product, so I decided to produce at least three additional sets to see how they compared to known distributions. The variables I used in this process were youth poverty (0 to 17), extreme poverty (population below 50% of poverty, regardless of age), and population below 200 percent of poverty.

The variables all worked as variables for the process, but some worked better than others and others looked very similar. The final preschool poverty variable seemed to make the most sense when I presented the results to the group for them to consider. This is also the distribution that matched most closely with past funding distributions and counts that were available to the program administrators. While it performed the best other variables worked well too. The overall youth poverty worked well and could have been a final variable, though it was not, in the end compared to the purchased data to see how close it comes. Similarly, the extreme poverty weight seemed to work well, but it seemed to favor the dense urban areas more and seemed to short-change some of the

more rural districts. Lastly, the 200 percent of poverty weight tended to look more like the first attempt that just used population.

What each of the attempted weighting variables have in common is their selection of the population by various levels of poverty. The variables that seem to produce the best results were variables that eliminated as much of the population that was not part of my target population. For example, if we think of the youth population in terms of the age groups represented, we have persons between and inclusive of the ages zero and seventeen, while the target population for estimates was specifically four-year-olds. This means that in terms of the age groups represented, the four-year-olds would be about five and a half percent. That would, of course, vary depending on the age structure of the particular geography, but four-year-olds are only one age group out of a possible 18 in that range. Similar issues made the other attempted weighting variables underperform when compared to the final choice of the early childhood poverty variable.

In an ideal world, there would be an estimate of the population that you are trying to estimate so that you can compare and find the best weighting variable. However, given that this process is meant to assist with making estimates for populations that do not have independent estimates with which to make the determination. The judgement of the researcher and the insights gleaned from a review of the data and subject area literature will need to guide the choice. There are however a few guiding principles that should aid in this choice.

1. The weighting variable should maximize the target population. This should help to prevent non-target portions of the weighting variable from exerting undue influence over the final estimates. For example, use early childhood poverty

instead of general children in poverty because four-year-olds make up a greater proportion of the early childhood ages than the general childhood ages.

2. Maximize the target characteristics in the weighting variable to give the estimates as much geographic specificity as possible. The assumption is social characteristics are autocorrelative in nature (Poudyal et al. 2016), and they will cluster, so maximizing the characteristics in the weight will help them to reflect the actual social conditions. This would make the 100 percent poverty data work better than the extreme poverty data, which was 50 percent or less of poverty, as the final estimates were for 250 percent of poverty. Care needs to be exercised when implementing this principle as maximizing a social characteristic may affect the proportional size of the population. For example, the summary data do not provide age specificity for poverty data at levels other than 100 percent, so using data at 200 percent of poverty would mean that the weighting variable would be using the entire population for which poverty status was determined. This dramatically reduces the proportion of the target population in the weighting variable.
3. Reduce, as much as possible, known cohort effects that will distort the final estimates. For example, the use of general poverty or youth poverty were not as good of choices as was early childhood poverty because of the known negative correlation between poverty status and age. As age increases, the probability of someone experiencing poverty decreases, so by using general or youth poverty, the weighting variable is simultaneously decreasing the proportion of the target

population and including more population that have a different (lower) probability of experiencing poverty.

Obtain or produce the necessary shapefiles and prepare them for use

This method is meant for social researchers with some GIS experience rather than true GIS professionals, so the actual production of shapefiles is not going to be discussed. The estimates produced with this method will require modifications to shapefiles, but the base shapefiles are available through public sources.

Shapefiles for all census geographies are available through the Geography Division at the U.S. Census Bureau. A simple internet search will locate the website, or the researcher can search for them on www.census.gov. These will likely be the primary source of shapefiles for projects that are using census data, however other state or local level shapefiles could prove useful for various purposes.

This step in the process is centered more on the preparation of the shapefiles for use in the project. The method should be replicable with any GIS software package, but the steps would be unique to the software package being used. This section will detail the process in the Quantum GIS (QGIS) software package as it is open source and available to anyone with an internet connection and a Windows, Apple, or Linux based computer. There are features in the ArcGIS software package produced by the Environmental Systems Research Institute (ESRI) that make the process easier, so I will mention how the process differs in ArcGIS at the end of the section.

As noted above, the process is more complicated in the QGIS environment, so I will tackle that portion of the description first. The part of the process that is more

complicated in QGIS deals with the spatial disaggregation of the weighting variable in the first part of the process. The QGIS program does not currently have a function that will automatically distribute the weighting variable to geographies that are created from combining shapefiles, so that part of the process must be completed manually. This is accomplished by recording the area of the polygons at two specific points in the process. The areas that must be recorded are the areas at the point of combining the county subdivisions with the school districts and the final combined areas for the final union shapefile. I will point out where to record the areas, and how to save them to a CSV file that will be used later in the process when the work shifts to Excel (or other spreadsheet program) for the completion of the estimates. These proportional areas will be used in the next steps to distribute the weighting variable's residuals to the school district parts after they have been allocated to the county subdivisions.

A thorough exploration of the previous two steps, exploring the requirements of the projects and determining a weighting variable should have put the researcher in a position where the necessary geographies for the project are known. At this point the shapefiles for those geographies need to be obtained. As mentioned above, those shapefiles should generally be available from the U.S. Census Bureau or similar data provider. Once those shapefiles are obtained, the process of transforming them into the working shapefiles for the project can begin. I will detail the process using the example that has been discussed up to this point, making estimates for four-year-old children at or below 250 percent of poverty by intermediate school district.

To begin the process, I obtained the shapefiles for the Public Use Microdata Areas (PUMA), elementary school districts, unified school districts, and county subdivisions

from the Census Bureau's website. An additional shapefile that is necessary contains the basic shapes for the great lakes. I will discuss the utility of this last shape later, but in brief, it will be used to remove the areas from the other shapefiles that are part of the great lakes and therefore not areas where people generally live.

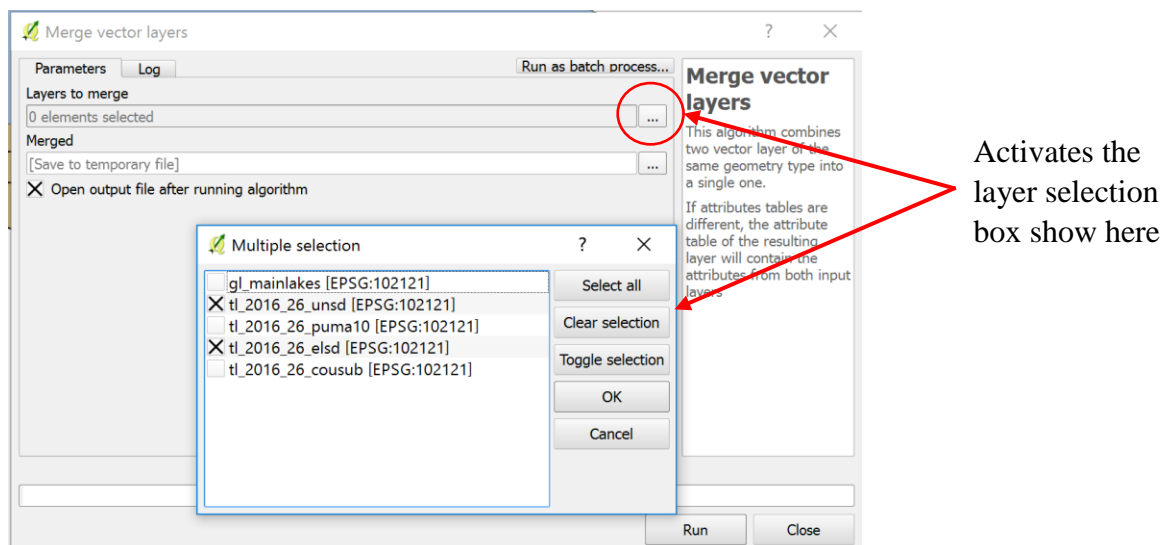
The first step likely to be necessary, especially if shapes were obtained from different sources, is to align the projections for all the shapefiles. The term projection and coordinate reference system (CRS) are often used interchangeably despite their specific and distinct and different meanings. I will use these terms interchangeably as their precise meaning is less important to this project. This is a vital step as the transformations that will be necessary later will not work or be accurate if the shapefiles being used are of different projections. This is a relatively simple process where the shapefiles are opened in the GIS software of choice and saved to a different location as the desired projections. In this project, all the shapefiles were saved as "NAD_1983_Michigan_Georef_Feet_US", because it best represents the whole of Michigan in the measurement system I am most comfortable with, feet.

At this point the researcher should have the necessary shapes to complete the project and s/he can begin the process of transforming the files as necessary to complete the project. This process needs to be very methodical and planned prior to the beginning. If it is not, the researcher is likely to spend time repeating the process several times to account for details that could have been foreseen. It is extremely helpful to sketch out the transformations in advance of performing the operations in the GIS software.

For this example, the steps for shapefile transformation I have outlined are:

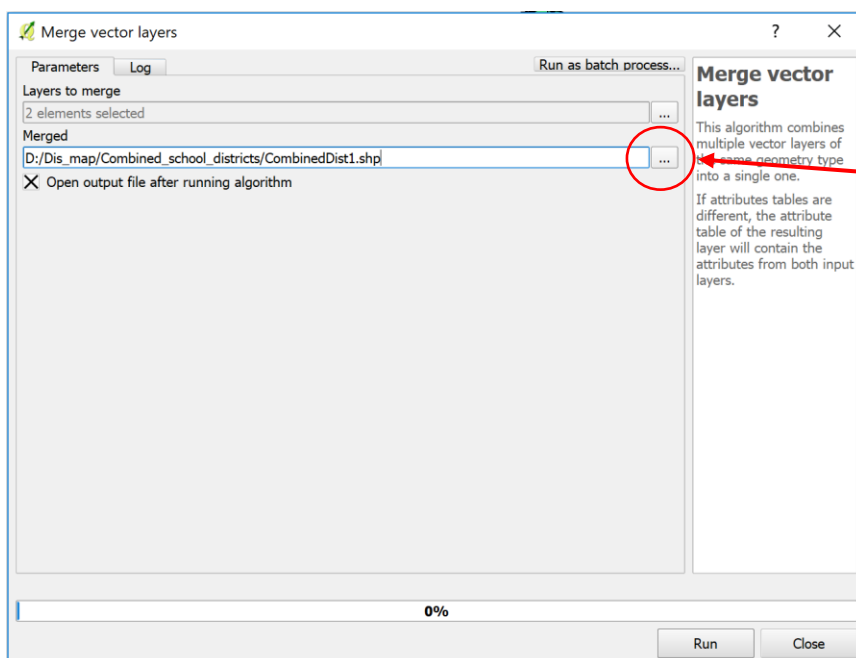
1. Combine the two types of school districts (Elementary and Unified school districts) into a single shapefile. This will result in an exhaustive, mutually-exclusive accounting of the geographic area of the state. To accomplish this in QGIS, follow the menu options Vector > Data Management Tools > Merge vector layers. That menu/command progression will bring up a menu where you can select the appropriate shapefiles and indicate a location where to save the new shapefile as seen in the two images below:

Figure C-2 Vector Layer Merge



This process will add a shapefile to your project that is the combination of the selected shapefiles. In the case of this example this combines the two types of school district areas which creates an exhaustive mosaic of the state. A peculiarity of the QGIS program adds the shape to the map with the title of “Merged” even if it was named differently in the creation process. It is advised that the user right click on the shape and rename it consistent with the name that it was saved as to avoid confusion later in the process.

Figure C-3 Vector Layer Merge Saving



Activates menu to name and determine location of new shapefile

2. Remove the portion of all shapefiles that are clearly not part of the habitable portions of the geographies. For this project that will be removing portions of all shapes that are covered by great lakes water bodies. Depending on the project, there may be other areas that could be removed for better estimates, for example, roads, inland lakes, etc. For this project removing the great lakes areas will improve the estimates without increasing the overall complexity of the project to an unmanageable level.

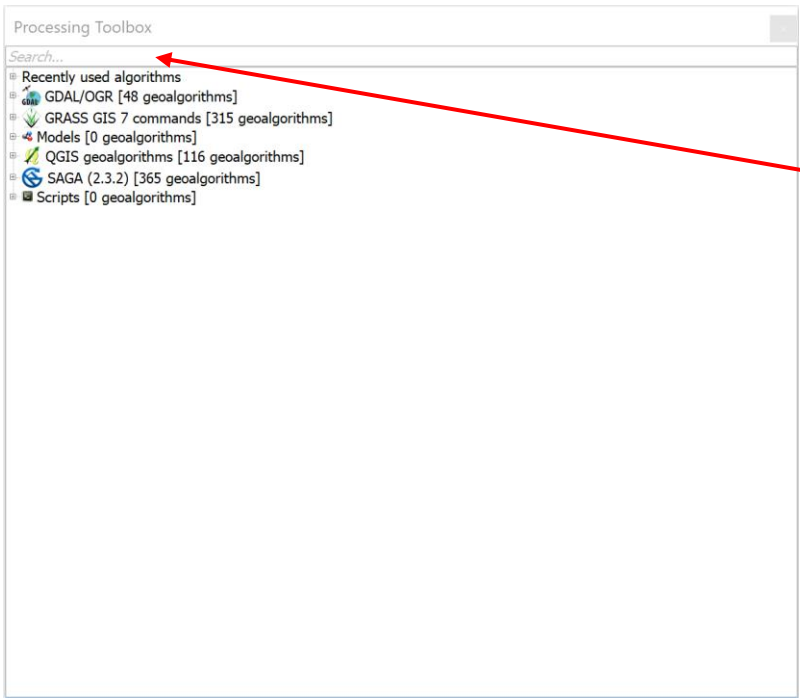
The requirements of every project should be evaluated independently to determine what, if any, areas should be removed to better account for the target geography.

In the case of the State of Michigan, removing the area under the great lakes reduces the study area by about 40 percent. In contrast, removing the inland waters, (rivers and lakes) would reduce the target area by just over one percent.

For the 40 percent decrease in area there are some modest increases in polygon complexity that account for the island areas of the state and the jagged nature of coastlines. For the minimal increase that would be provided by removing the inland water areas, there would be a larger increase in the complexity resulting from the transformation of nearly all polygons into multipart, discontinuous geographies. Given the large increase in complexity and the minimal improvement of the target areas, the decision was made to not remove the inland water from the target areas. This sort of tradeoff would need to be evaluated for every application of the method.

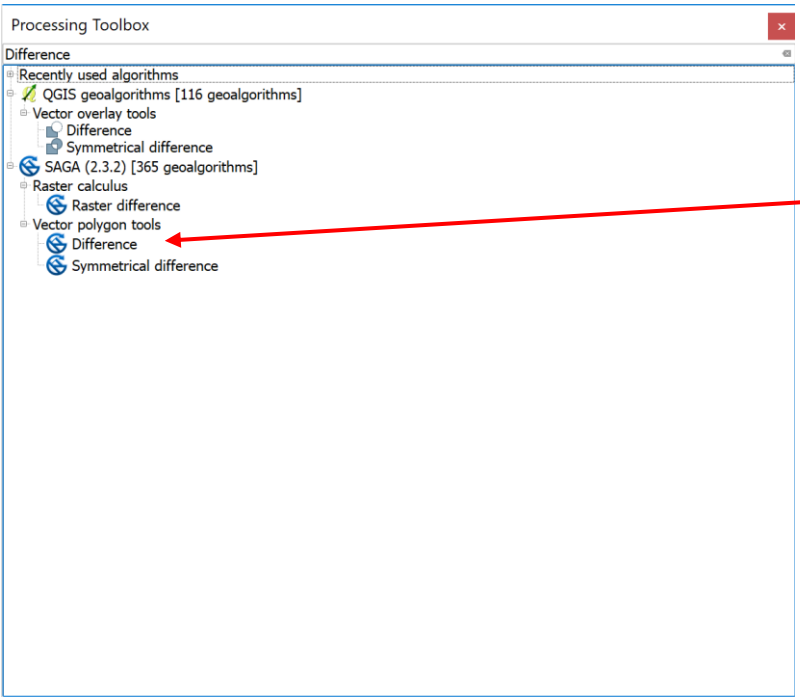
For each layer, county subdivisions, school districts, and PUMAs, the process of removing the water areas is accomplished by cutting them out with the difference function. QGIS actually provides multiple difference algorithms based on its nature as an open-source software and its incorporation of other, open-source, GIS programs. Through multiple iterations of this process the functionality that I have found to work best for this project is the functionality provided by SAGA algorithm set. To access these functions, the user will navigate to the Processing Toolbox found by navigating the menus Processing > Toolbox. This is also available with the keyboard shortcut Ctrl+Alt+T. This will open the Processing Toolbox where the user can search for the difference algorithm provided by the SAGA software:

Figure C-4 Search for Difference Algorithm



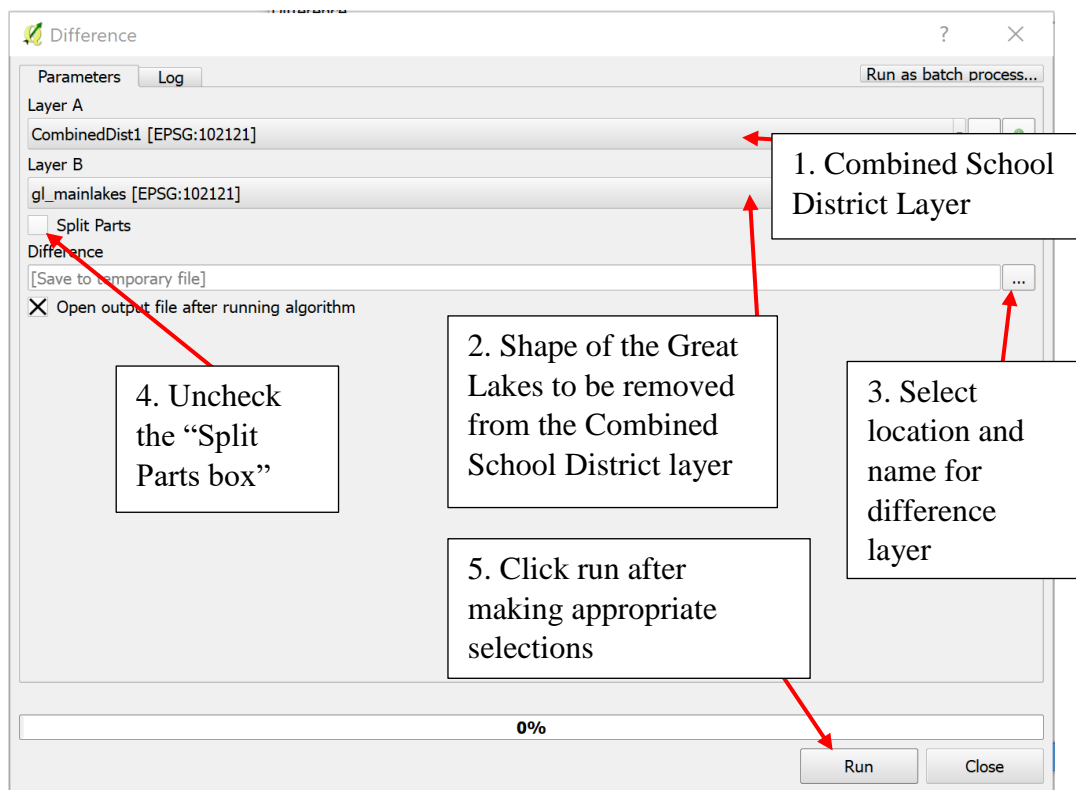
Use search function to find "Difference" algorithm

Figure C-5 Select for Difference Algorithm



Selecting the Vector Difference tool will bring up the difference tool

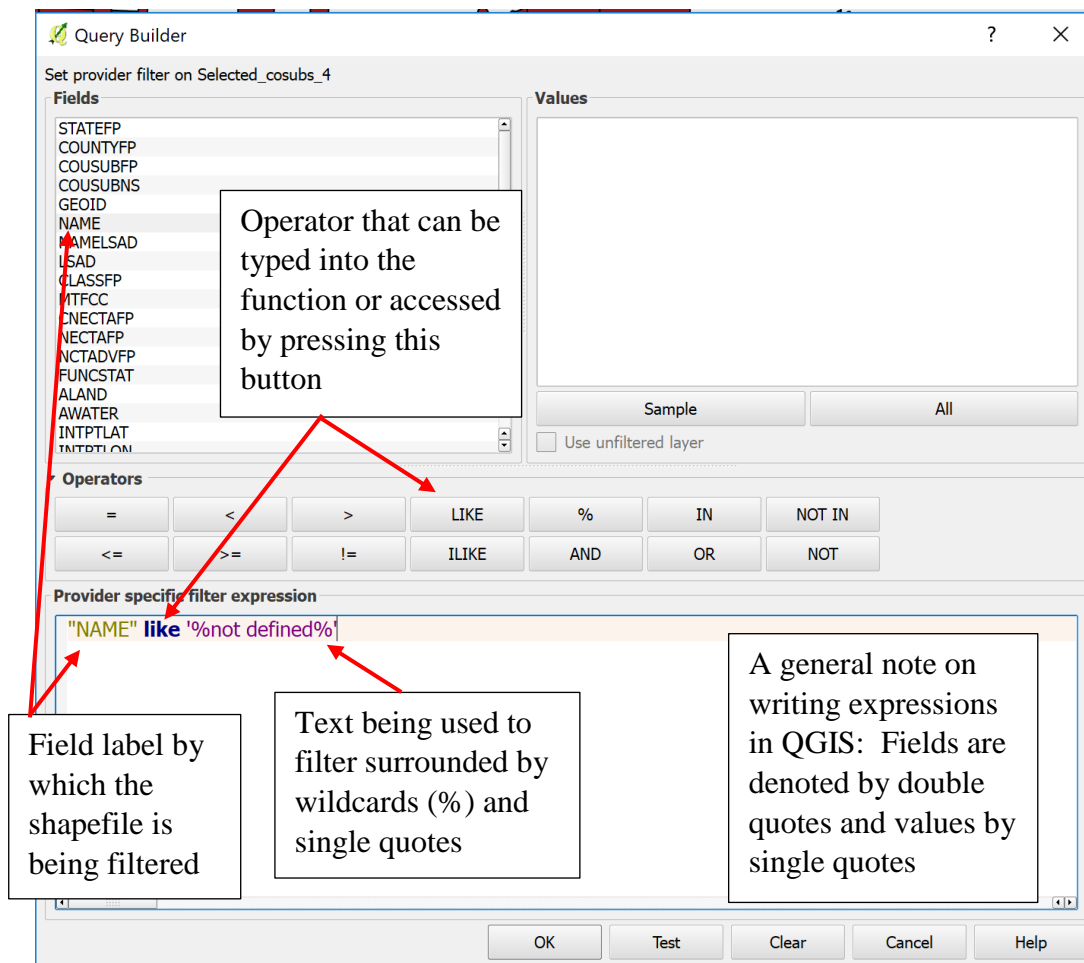
Figure C-6 Difference Settings



After the difference algorithm is applied to the shapefiles, the user will need to inspect the remaining shapes to try to find any sliver polygons that exist as artifacts of the difference process rather than areas appropriate for disaggregating the target population. The sliver polygons created in this operation are the result of portions of the school district that are in areas that are sometimes covered by water. We know this because the visual inspection of the remaining polygons reveals that the slivers are all listed with the name "School District Not Defined." That designator means that the geographic area is not part of a school district but was included in the shapefile to create an exhaustive geographic accounting of the state. The only area that would fit that descriptor would be an area that was under water.

Given the similarity of the sliver geographies noted above, it is a fairly easy operation to highlight and remove them from the school district shapefile that was created in this step. The filter command found in the Layer menu, with a navigation path of Layer > Filter (or Ctrl+F), will provide the mechanism for removal. The menu can be operated as follows:


Figure C-7 Look for Undefined Areas



The results of the filter expression above are the shapes in the image below

Figure C-8 Look for Sliver Areas



It is clear from a visual inspection that these shapes are all along coastlines of the state that result from differing marks of where the school districts begin and end. Because we will be distributing portions of the data based on areal coverage, it is important to remove as much of the uninhabited land area as possible. In this case it is a simple matter of selecting the area with the rectangle selection button () . Once depressed, the user can trace a rectangle over the intended geographies and they will be selected. Once selected, they can be removed by


selecting the layer edit button (). This will open the layer for editing and expose all polygon nodes with red “x” marks shown here:

Figure C-9 Select Sliver Areas



Once the geographies are selected in edit mode, it is a simple matter to delete the unwanted shapes. With all the geographies selected, press “Delete” and they will be removed from the active frame. At this point the user will again press the pencil icon which will remove the layer from edit mode. The user will be asked if the changes to the shapefile should be saved. The user should click save. At this point the user will be presented with a completely blank active window, which

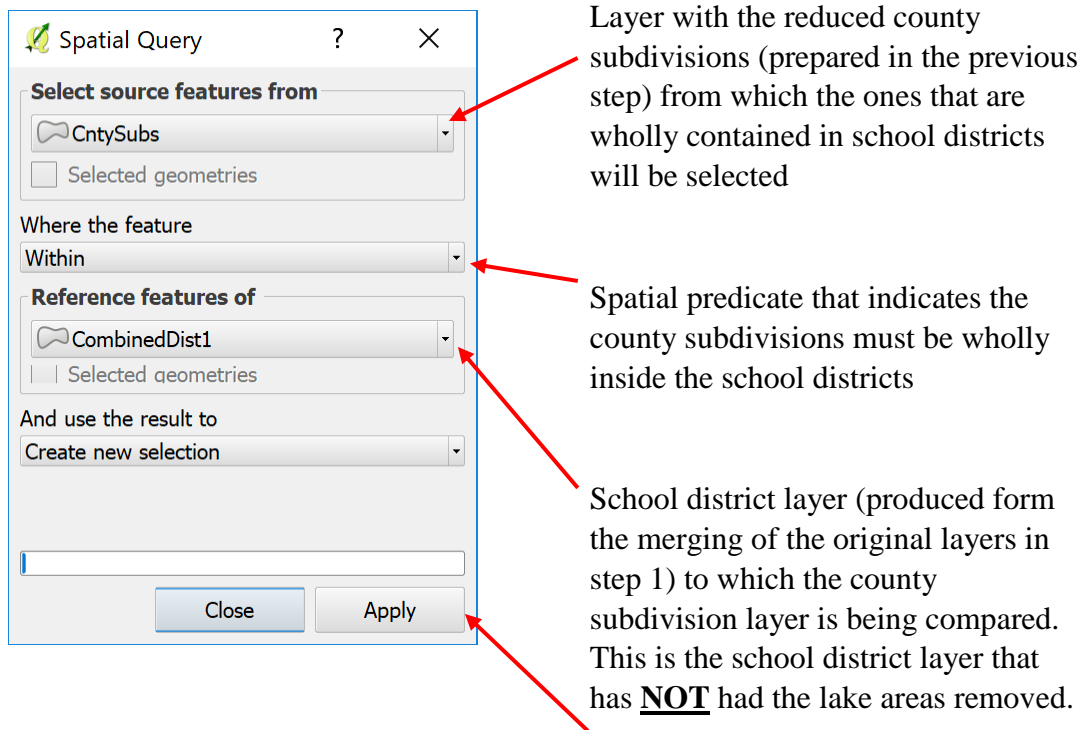
would be disconcerting at first, but that is because the shape is still being filtered to look for the shapes that have just been deleted. To get the desired geographies back the user should go back into the Filter window, either through the menu or by pressing Ctrl+F, and clicking the clear button. Once the filter has been cleared the desired geographies will reappear in the active window.

The description for this step has centered on the school district shapefile, but as mentioned at the beginning of the step, the process needs to be iterated for all current shapefiles in use, this will include the county subdivision and PUMA shapefiles.

Once the differences have been taken for all the relevant layers, the shapefile for the great lakes can be removed from the project. Right-clicking on the layer in the layer selection window will bring up a context menu where “Remove” can be selected. Removing this layer will reduce the number of active layer which will help reduce error by decreasing the chance of selecting incorrect layers in the processes to follow. It is good idea to remove layers when they are no longer going to be used for a procedure later in the process.

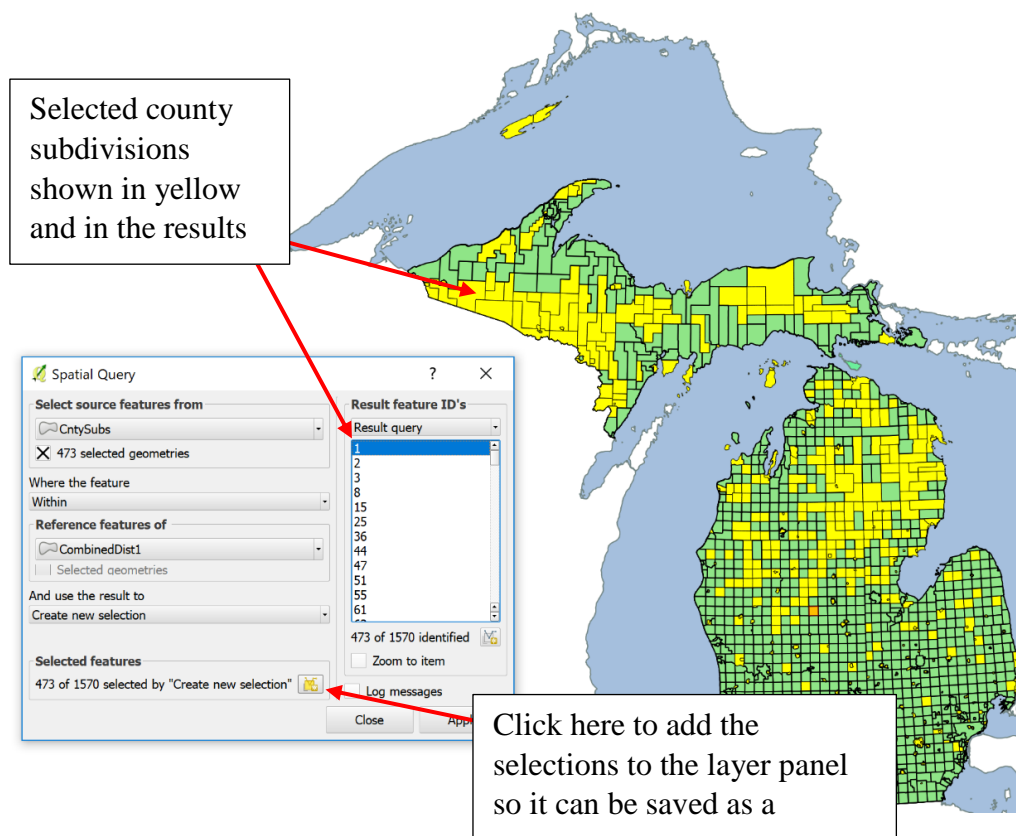
3. Identify county subdivisions that are wholly within school districts and save those geographies as a subset in a spate shapefile. To accomplish this in QGIS follow the menu options Vector > Spatial Query. That progression will bring up a dialog box to select the required shapefiles and the spatial predicate to be applied. Here are how the selections for this process are made:

Figure C-10 Spatial Query



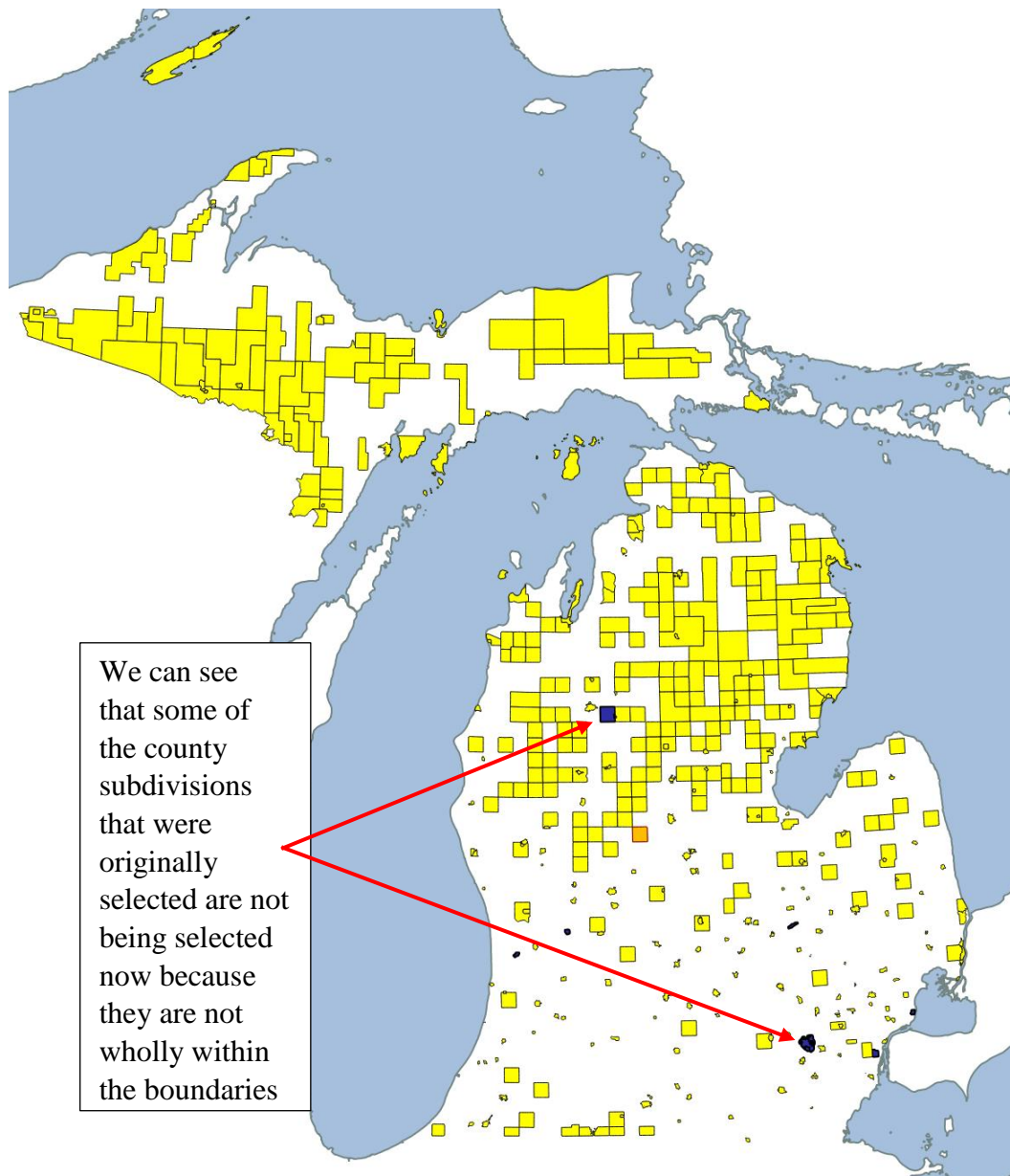
The result of this process is a selected subset of the county subdivision layer which can be added as a new layer and then that layer can be saved as a shapefile to be used later in the project. That part of the process can be seen here:

Figure C-11 Spatial Query Results



4. Identify county subdivisions from previous step that also exist wholly within PUMA boundaries and save that subset as a shapefile. This process is performed in the same way as in step 2, substituting the newly created selected county subdivisions layer and the PUMA layer for the original county subdivision and combined school district layer respectively. The results of this process can be seen here and should be saved as a shapefile named to indicate the iterative nature of this process.

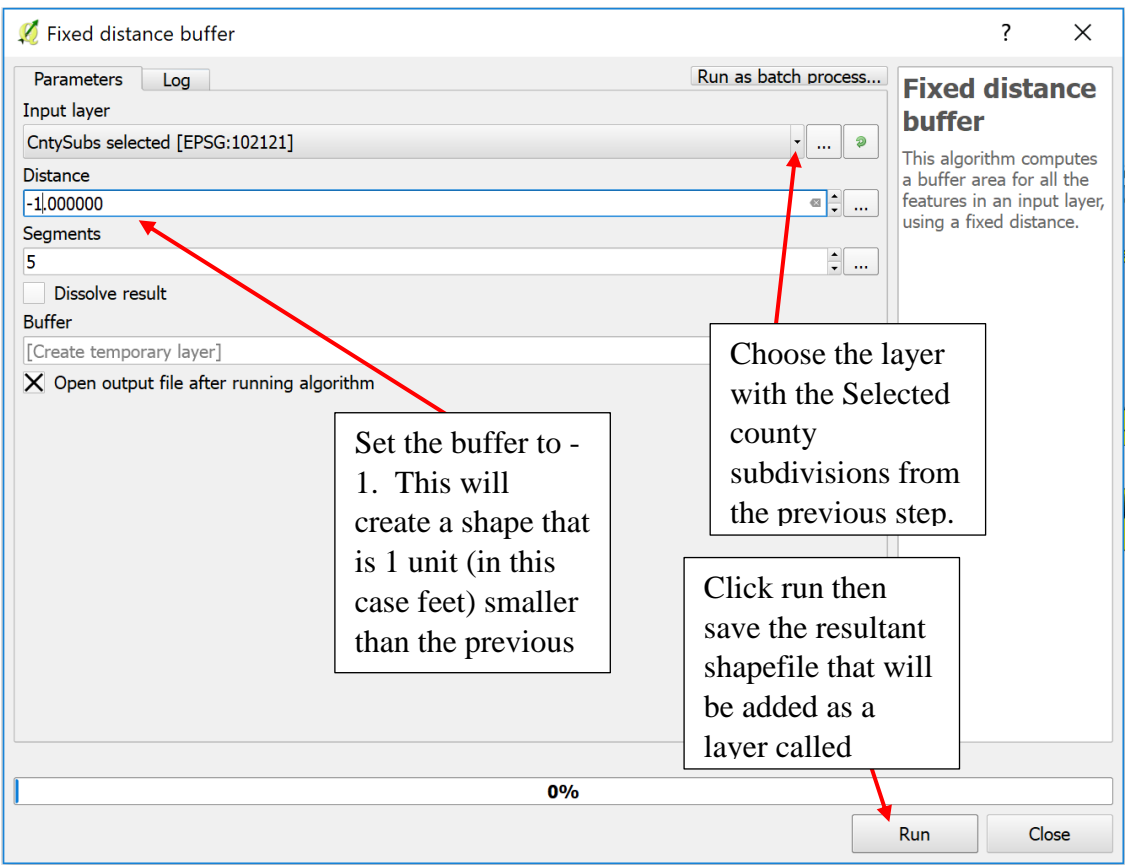
Figure C-12 Combined Query Results



5. Attach district labels to the county subdivisions identified in previous step. This process is accomplished by using the “Join attributes by location” function in the vector menu. Before that process is run, there is another algorithm that should be processed to help the correlations between county subdivisions and other geographies process. That algorithm does not consistently join locations if they

share a boundary. Many of the county subs identified by the spatial query share a boundary with one or both other shapes, so reducing their size by a tiny amount helps this process run. To reduce the size of the selected county subdivisions, the fixed width buffer function is very useful. To employ this algorithm, use the menu progression Vector > Geoprocessing Tools > Fixed Distance Buffer. This menu progression will activate the buffer window below:

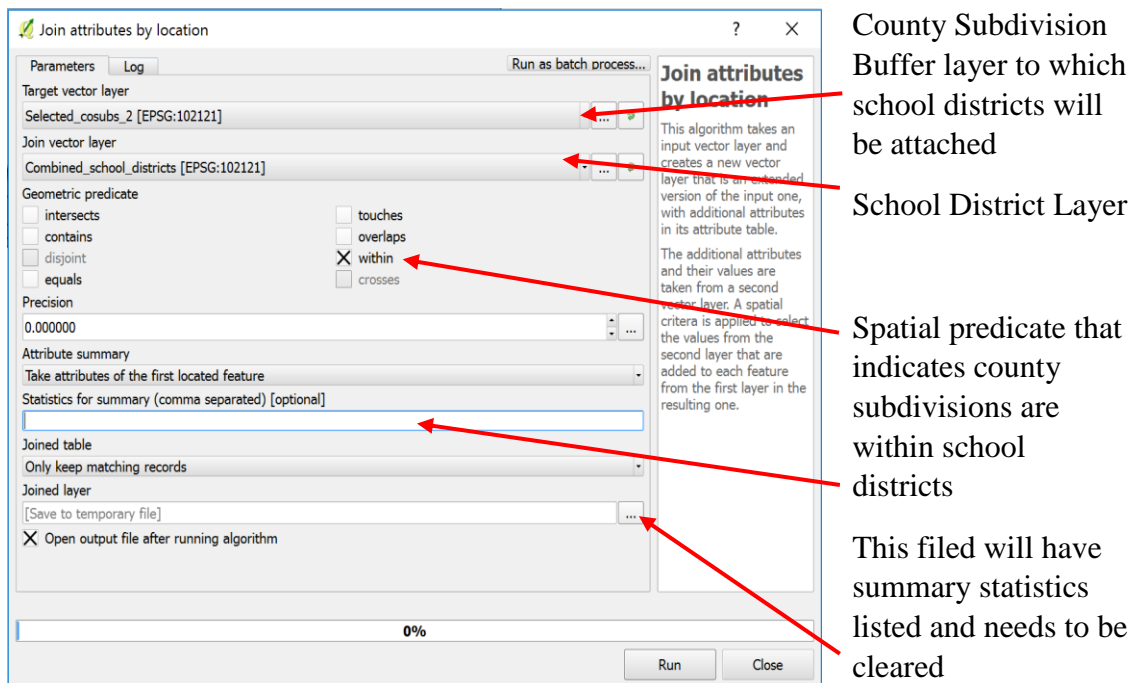
Figure C-13 Buffer Settings



The buffered layer will be nearly identical to the previous layer and you will only be able to see a difference through visual inspection if zoom in to a boundary. This buffer layer should be used to create the correspondence tables in the remainder of this step.

The menu progression for the Join attributes by location is Vector > Data Management Tools > Join attributes by location. Below is a graphic

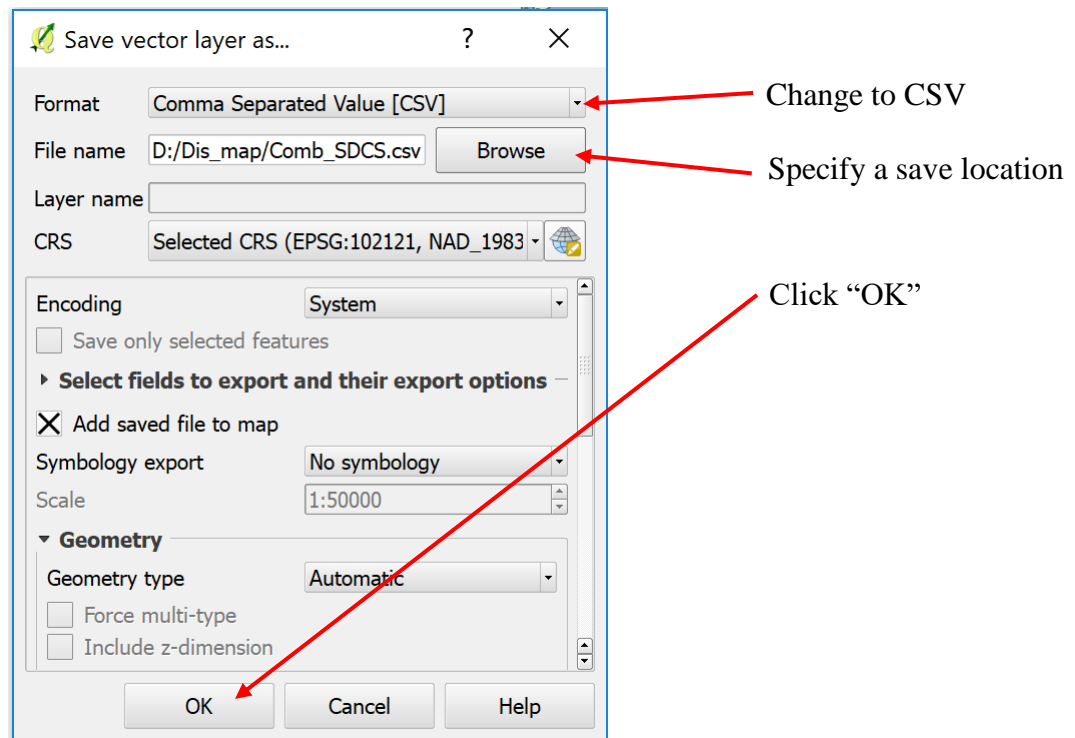
Figure C-14 Join Settings



representation of how to interact with that menu:

Once the process is complete the layer will be added to the active window in QGIS. This layer is needed to identify the correspondence between school districts and county subdivisions. This layer should be saved as a CSV file to be used later. Once the layer is saved as a CSV, it can be removed from the active project. The buffer layer can also be removed from the project at this point. The process of saving a layer to a CSV file is accomplished by right-clicking on the shapefile in the layer panel and selecting Save As. This will bring up a save window as seen below:

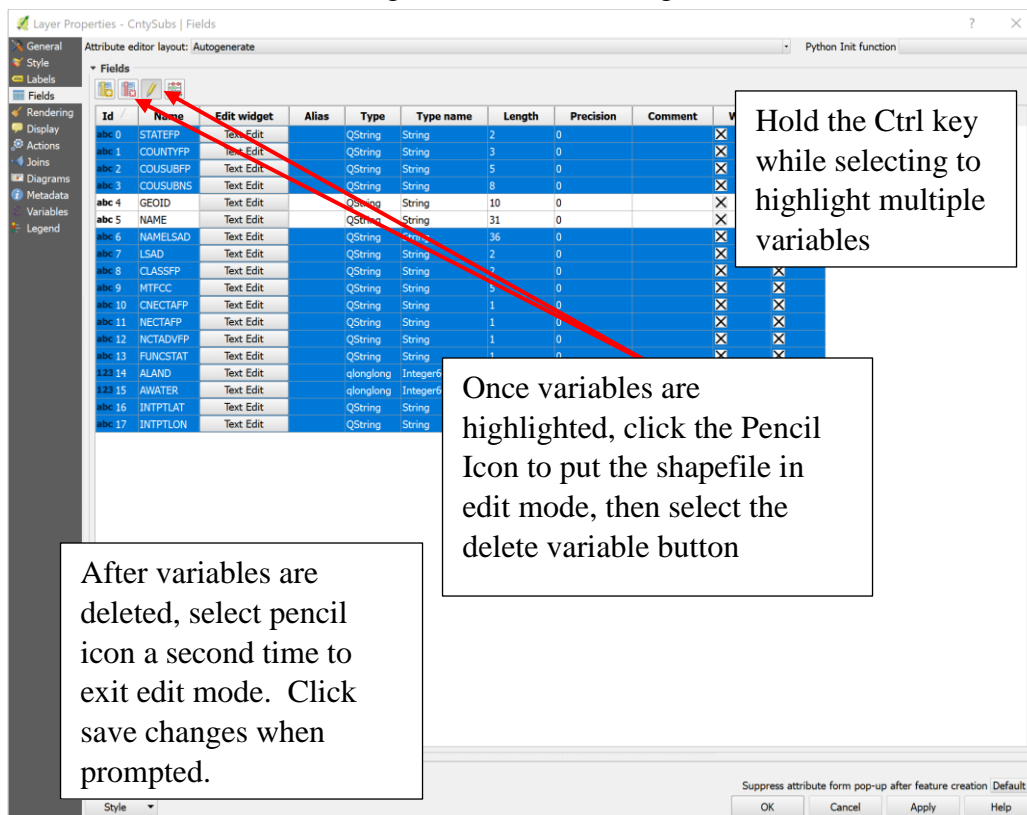
Figure C-15 Save as CSV Settings




6. Cut the county subdivisions geographies out of the school district shapefile, and save the new school district file with the county subdivisions removed using the procedures detailed in step two.
7. Combine the last iterations of the shapefiles for school districts and county subdivisions, which should form an exhaustive accounting of the geographic area of land area of state using the procedures detailed in step one. At this point, there may be problems merging the shapefiles because of the transformations that have been performed. The most common error received is one describing a mismatch in data types between variables. The easiest way to solve this situation is to delete all the fields in the shapefiles except for the GEOID and NAME fields. This makes the spreadsheet work later easier as well, so it is recommended even if

there are no issues that would cause merge errors. Deleting variables from the shapefiles is an easy procedure. See the graphic example below.

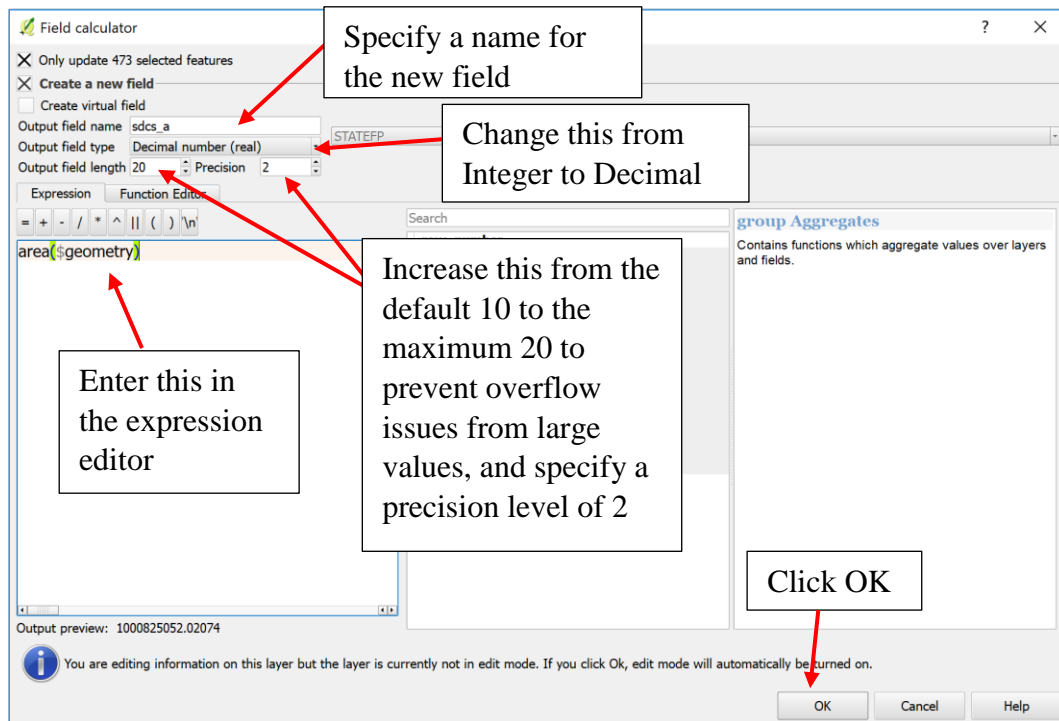
Figure C-16 Join Settings



8. Area for geographies in the combined County Subdivisions/School Districts layer need to be recorded. This will be the shape areas that will be used in later steps to areally disaggregate the weighting variable. To obtain polygon areas, select the shapefile for which areas are to be determined and then open the field calculator whose button is styled as an abacus (). The field calculator can be opened in a few different ways. The first way is to double click on the layer for which you want to make calculations, then click on the “Fields” tab, then you will see the field calculator button on the windows ribbon. The layer properties menu can

also be opened by right-clicking on the layer and then clicking “Properties” from the context menu. Finally, the field calculator menu can be opened by clicking on the field calculator button from the top ribbon when the appropriate layer is

Figure C-17 Area Settings



selected. When the calculator is opened the fields should be specified as follows:

This process will calculate the area for each polygon in the shapefile selected.


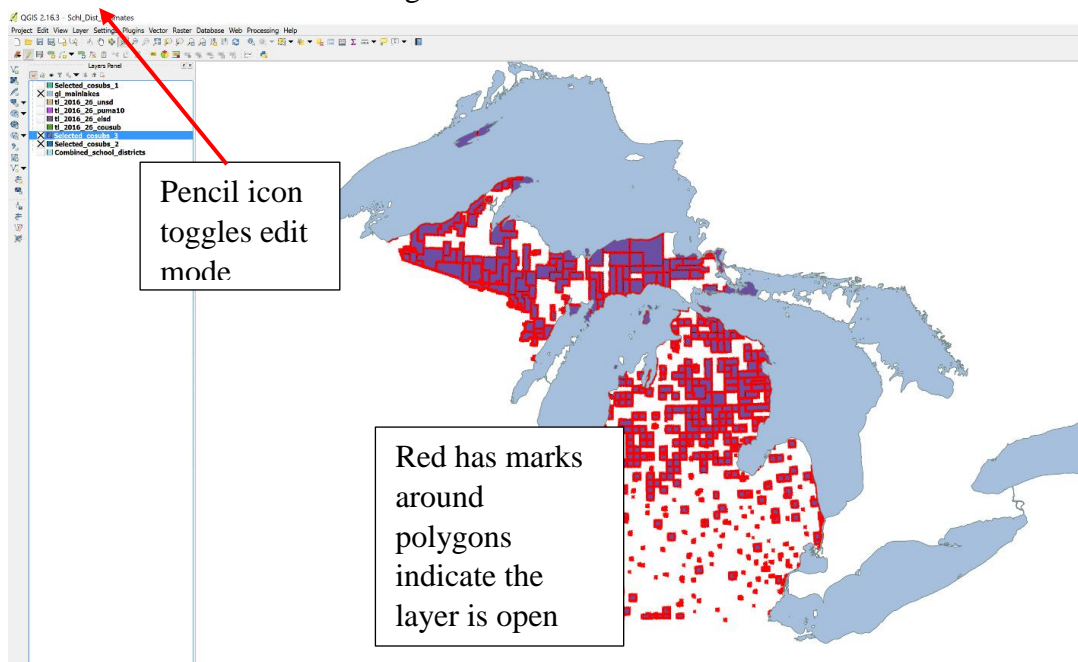
When the process runs it will put the layer in editing mode. This needs to be turned off by clicking on the pencil icon (). This will be displayed as indicated below. When the pencil icon is selected to turn off edit mode, the program will ask if the user wants to save changes. Click “Save” to keep the areas calculated.

Figure C-18 Save Areas



9. Once the areas have been recorded, the layer must be saved as a CSV file to preserve these areas. Save this file to the same directory that the correspondence table was saved in step five.

This will be the end of the process of preparing the shapefiles for use. At this point, the user will have two end shapefiles, one that is an exhaustive accounting of the state made up of a combination of school districts and county subdivisions and a second that contains the PUMA geographies, both have had the areas covered by the great lakes removed. Both shapes can now be processed with the “Union” algorithm which will combine them into a single shapefile, as described in the next step.

Combine the Shapefiles

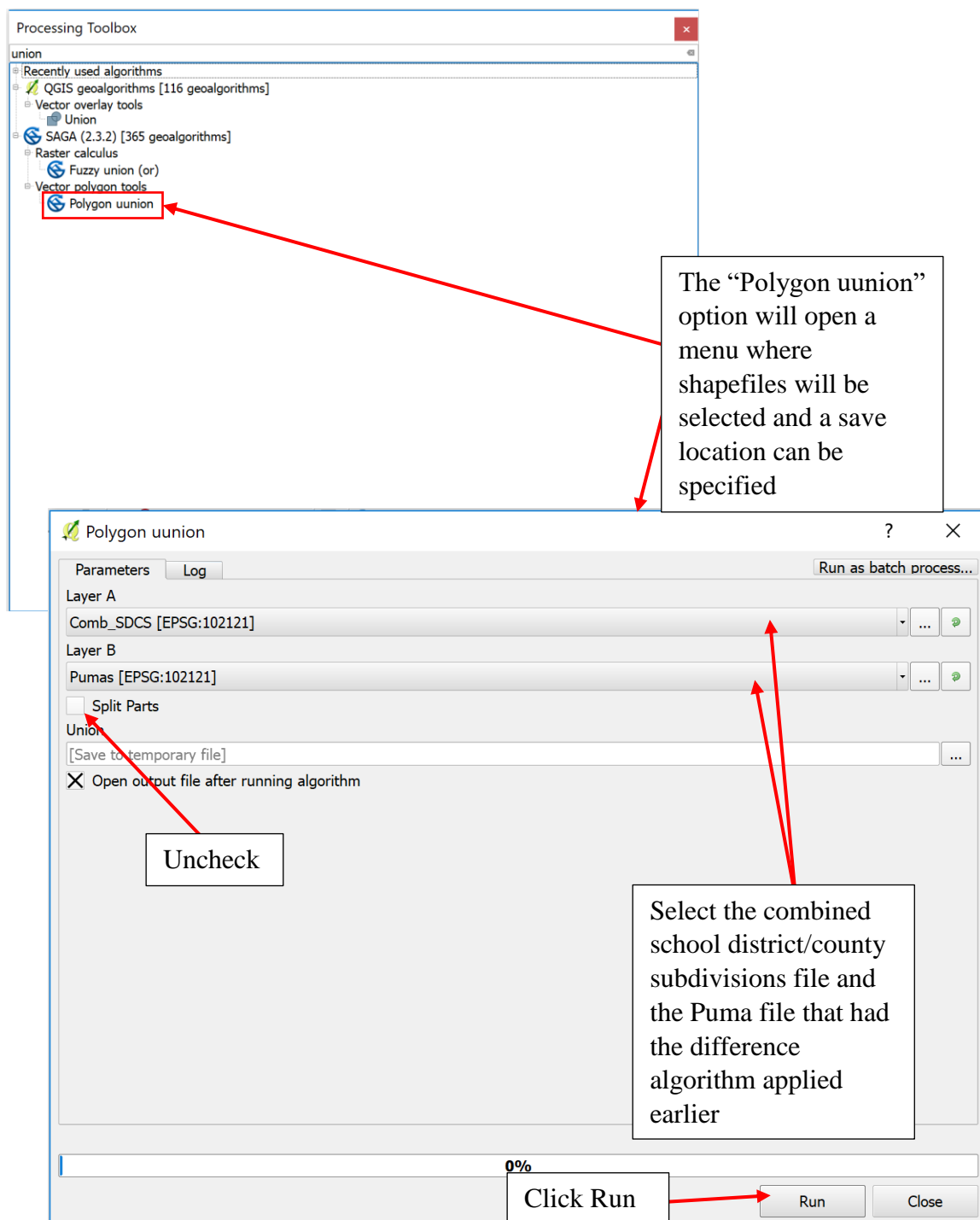
Combining the shapefiles that have been produced to this point, the combined school district/county subdivision shapefile and the PUMA shapefile, is necessary to

bring all the shapes together and allow the user to make the interpolated estimates.

Combining the shapefiles in QGIS is not a difficult process, but needs to be completed with a version of QGIS that is 2.18 or higher. Earlier versions of QGIS have a bug that may prevent the user from completing the process. An additional step that was not detailed above should be performed prior to the union algorithm. That is to delete all the extraneous variables from the PUMA shapefile. This will make for a cleaner end file and prevent errors from stopping the union process.

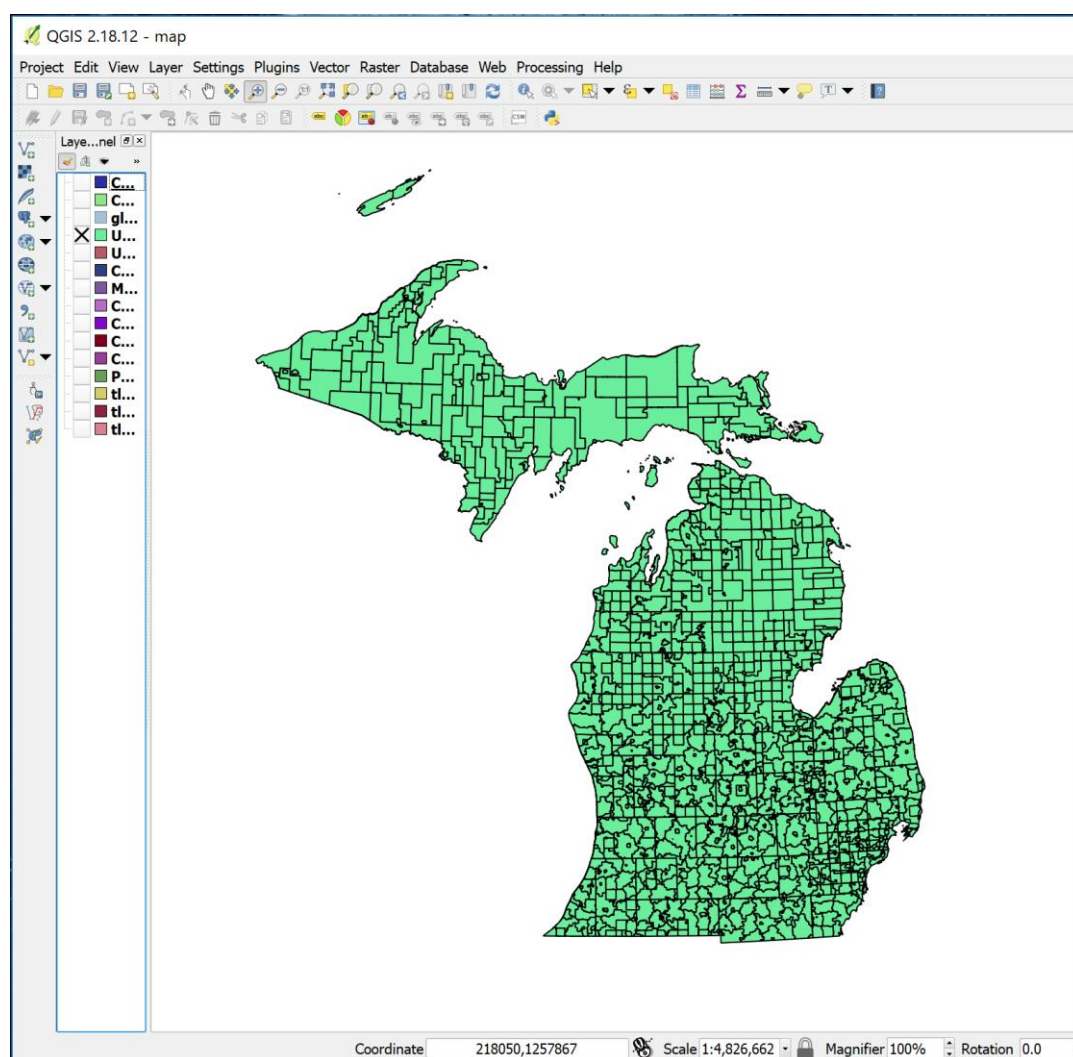
As with the difference algorithm used in the previous section, QGIS offers several different ways to perform the union process. The one that produced the best results for this procedure was, again, provided by the SAGA geo-algorithms. That functionality can be accessed from the Processing Toolbox used previously when performing the Difference operations. Instead of searching for difference, search for “Union” and select the “Polygon union” algorithm. The selection can be seen in the image below. As a reminder, the processing toolbox can be accessed through the menu progression Processing > Toolbox or by using the keyboard shortcut Ctrl+Alt+T.

Figure C-19 Union Settings



The union process, in QGIS, will add the new shapefile to the active window, which is a combination of both shapefiles, as seen in the image below. This is the point in the process where the researcher will record the area of all polygons in the new shapefile and export the shapefile to a CSV file. This new file should be used in Excel where that can be manipulated, and where the final steps of the estimation process will take place.

Figure C-20 Union Polygons



An important thing to note, when the new shapefile is generated and presented to the user, it may not look like all the shapes have been combined. As can be seen it may appear that there has been a partial combination and that some areas have not been combined. This relates to the way that QGIS combines the underlying polygons, assigns reference ids, and renders the polygons. The union algorithm creates a new layer that has all the original polygons and all the intersecting or overlapping polygons. As a consequence of this process, the program does not always render an image that looks like and exhaustive accounting of the intersections of the polygons form the underlying shapes. When the shapefile is exported from the QGIS program and opened in Excel, it will be sorted so that only the relevant shapes are being considered. This process will be discussed further in the following sections.

The last step prior to exporting the unioned shapefile to a CSV file will be to record the area for all the geographies in the final shapefile. This will give us the numerators for the proportions that we will use later to areally distribute portions of the data from the PUMS data. I will discuss the steps to follow for the remainder of the estimation process in Microsoft Excel.

Add the weighting variable

The task now is to get data from the U.S. Census Bureau's American Community Survey to attach to the county subdivision and school district level geographies that were saved in the CSV file in the previous step. To accomplish this the researcher will pull data for all county subdivisions, elementary school districts, and unified school districts from the Census Bureau's data dissemination tool, American FactFinder, available at <http://factfinder.census.gov>. These data should be downloaded at the same time in a single data pull so all data will be in the same CSV file. We can use data for all the geographies even though we are only using a subset of the county subdivisions because we will be using a function in Excel that will search for geographic codes from the data file that match those in the CSV file we exported previously. It is much easier and faster to pull all the geographies than to try to only select the ones we will be using to make the estimates. As I have mentioned in earlier sections, this method assumes a certain familiarity with census and ACS data, so I am not going to detail the steps in pulling the data, but the user will be delivered a CSV file via download from the Census Bureau that will have all the necessary data, and where the second column will be a geographic code, a Federal Information Processing Standards (FIPS) code, which matches the *GEOID* in the CSV file we exported from the work with the shapefiles.

The data we will need to attach to the data from the unioned shapefile data will be from the ACS table B17001. This is the basic poverty table that breaks down the population by age, gender, and poverty status. The first thing that will be necessary will be to total the males and females that are less than five years of age and below the poverty level. This is a simple operation in Excel and involves summing two cells for all

geographies that were downloaded previously. Once that minor task is complete, the user will have two open spreadsheets, one with the unioned data and one with the poverty data. Once the poverty data is summed for all geographies, the work will be in the file with the unioned data.

When the user opens the CSV obtained from QGIS, as mentioned previously, the file will contain data for all the shapes that existed independently and that were created when the geographies overlapped. The CSV file will only have the fields that were not deleted from the previous shapefiles before they were unioned. For example, the union file that was produced here resulted in total of 5 variable fields represented by columns in the CSV file. If variables had the same name between files, the data will be merged into a single column in the resultant file. For example, all the FIPS codes are in a single column labeled *GEOID* in the saved CSV file. In addition to the *GEOID*, the CSV file had values for the *NAME*, which corresponds to the name of either a school district or a county subdivision; *GEOID10*, which corresponded to the *GEIOD* for the PUMAs; *NAMELSAD10*, which represent the long name for the PUMA regions; and *fnl_area*, which was the name I used to represent the final areas recorded for geographies after the union algorithm was applied to the PUMA and combined school district/county subdivision shapefiles.

The file will need to be sorted to get rid of all the entries for the original shapes that are not part of the exhaustive, mutually exclusive mosaic of polygons that are in the unioned file. The entries that the researcher will want to keep will have entries in all the fields as those are the final polygons. The researcher can accomplish this with a series of filters, but an easier solution is to scroll through the data and find the cut point where data

stops filling all the columns. The SAGA union process places the overlapping geographies at the beginning of the file and the individual accounting of the underlying geographies at the end. With that structure, it is easy to scroll through and find that point. The user would then highlight all the rows below that point and delete those rows. The example below shows that point in the spreadsheet with the highlighted rows already marked for deletion.

Figure C-21 Cleaning Data

	A	B	C	D	E	F	G	H	I	J	K
1	GEOID	NAME	GEOID10	NAMELSAD10	fnl_area						
1468	2636390	Whittemore-Prescott Area Schools	2601300	Iosco, Gladwin, Roscommon, Ogemaw & Arenac Counties PUMA	2613770762						
1469	2636420	Williamston Community Schools	2601801	Ingham County (South & East) PUMA	1579143072						
1470	2636480	Wolverine Community Schools	2600400	Northwest Lower Peninsula (East) PUMA	0						
1471	2636480	Wolverine Community Schools	2600300	Northeast Lower Peninsula PUMA	4565927636						
1472	2636485	Woodhaven-Brownstown School District	2603205	Wayne County (Southwest) PUMA	74564767.73						
1473	2636485	Woodhaven-Brownstown School District	2603206	Wayne County (Southeast)-Downriver Area (South) PUMA	521388458.1						
1474	2636540	Wyandotte City School District	2603207	Wayne County (Southeast)-Downriver Area (North) PUMA	148034101.5						
1475	2636570	Wyoming Public Schools	2601001	Kent County (Southwest) PUMA	311501854.2						
1476	2636600	Yale Public Schools	2601600	Tuscola, Sanilac & Huron Counties PUMA	120770984.1						
1477	2636600	Yale Public Schools	2603100	St. Clair County PUMA	4201470775						
1478	2636630	Ypsilanti Community Schools	2602703	Washtenaw County (East Central, Outside Ann Arbor City) PUMA	999323994.9						
1479	2636660	Zeeland Public Schools	2600802	Ottawa County (East) PUMA	1453401583						
1480	2636660	Zeeland Public Schools	2600900	Allegan County PUMA	349091794						
1481	2636660	Zeeland Public Schools	2600801	Ottawa County (West) PUMA	851108261.8						
1482	2615754240	Millington			0						
1483	2610122320	Dickson			0						
1484	2614369560	Roscommon			0						
1485	2603180880	Tuscarora			0						
1486	2601765980	Portsmouth			0						
1487	2601787380	Williams			0						
1488	2610944360	Lake			0						
1489	2604106040	Bay de Noc			0						

Alternatively, the filter option can be activated on the ribbon in Excel or by pressing **Ctrl+Shift+L**. This will bring up buttons on all the column headers that can be used to filter out the entries that are not part of the final process. Click on the button in the column labeled at *GEOID10*, and uncheck the top box next to the words “(Select All)”, and scroll to the bottom of the list and check the box next to “(Blanks)” as seen in the image below. This will filter out all rows except for the rows that are blank in the *GEOID10* field which is the field that identifies the PUMA of the polygon. If this field is blank, it is not part of the exhaustive mosaic of the state. The user will highlight and

delete these rows. Once that is complete the user can remove the filter either by going back to the filter and clicking on the “(Select All)”, or by pressing Ctrl+Shift+L again to turn off the filter. Once that is completed, the user can refilter the results to delete all the entries that have no data in the *GEOID* field. Once those two filters are complete, the user should be left with the complete, exhaustive, mutually-exclusive dataset that is the end set that we need to move forward to incorporating the weighting data, producing the weights and ultimately making the end estimates.

Figure C-22 Creating Weights

GEOID	NAME	GEOID	NAME	fml_area
2607369300	Rolland	Isabella, Gratiot & Clare Counties PUMA	999956538.7	
2607373200	Sherman	Isabella, Gratiot & Clare Counties PUMA	994950042.5	
2602310880	Bronson	St. Joseph & Branch Counties PUMA	967983470.5	
2602343260	Kindertrook	St. Joseph & Branch Counties PUMA	594964503	
2602357860	Noble	St. Joseph & Branch Counties PUMA	594479209.2	
2614530220	Frankenmuth	Saginaw County PUMA	902832643	
2614530220	Frankenmuth	Tuscola, Sanilac & Huron Counties PUMA	0	
2609139720	Hudson	Lenawee & Hillsdale Counties PUMA	61165974.79	
2609155500	Morenci	Lenawee & Hillsdale Counties PUMA	59117303.07	
2608149540	Lowell	Kent County (Southeast) PUMA	86240671.05	
2616122160	Dexter	Washtenaw County (West, Northeast & Southeast) PUMA	55217701.5	
2609179120	Tecumseh	Lenawee & Hillsdale Counties PUMA	164597553.9	
2608169080	Rockford	Kent County (North) PUMA	92308179.72	
2608181520	Vergennes	Kent County (Southeast) PUMA	985831525.9	
2600761520	Ossineke	Northeast Lower Peninsula PUMA	2988753828	
2612919260	Cumming	Iosco, Gladwin, Roscommon, Ogemaw & Arenac Counties PUMA	985486542.3	
2612925020	Edwards	Iosco, Gladwin, Roscommon, Ogemaw & Arenac Counties PUMA	993989022	
2612929860	Foster	Northeast Lower Peninsula PUMA	0	
2612929860	Foster	Iosco, Gladwin, Roscommon, Ogemaw & Arenac Counties PUMA	2504535839	
2616153920	Milan	Washtenaw County (West, Northeast & Southeast) PUMA	33001530.86	
2616171140	Saline	Washtenaw County (West, Northeast & Southeast) PUMA	120638464.4	
2615740440	Indianfields	2602701 Tuscola, Sanilac & Huron Counties PUMA	932299291	
		2601600 Tuscola, Sanilac & Huron Counties PUMA	932299291	

The PUMA areas have a different *GEOID* and naming convention because the *PUMA10* label is used to indicate they were the PUMAs that were created following the 2010 Census instead of PUMA areas that were created after the 2000 census which were used in published, single-year datasets and tabulations until 2011 and are still part of the 5-year estimate series. There was not a shapefile in this project that would have had the 2000 label, but it is still used because the 5-year estimates series, available from the

Census, will continue to have PUMS data from the 2000 PUMA delineations until the 2016 5-year PUMS data is released in January of 2018.

The researcher now will have a file that represents the complete exhaustive, mutually-exclusive, mosaic of polygons that cover the land area of the state. As a check on this process we can sum the final areas that were recorded after the union algorithm was run, then compare to known values for the state. When this process is performed for the Michigan dataset, the sum of the final areas is 1,636,471,322,927.16 ft², and when we convert that to mi² the result is 58,700.33. As this is just a quick check and verification step, the researcher can look up land area quickly on the internet. For this example, I find that the land area of Michigan is 58,110 mi² according to the Wikipedia (2017) entry for “Geography of Michigan,” which is 590 mi² different. That difference is negligible and can be accounted for by not having removed the known inland water from the areas, though I would have expected the difference to be larger knowing there is close to 1,300 mi² of inland water in the state. Areal differences are also likely to have been created by the deletion of the sliver polygons and by the general process of cutting off the great lakes area from the shapefiles. There is also the knowledge that Wikipedia is not always a reliable source. When similar figures are downloaded from the U.S. Census Gazetteer files, the land area for the state sums to 56,546.69 mi² which puts the rounded difference at about 2,100 mi², which is much more in line with expectations considering the inclusion of inland water with the areas that I am considering land areas for interpolation purposes. I will remind the reader of the earlier discussion of the large amount of complexity involved in discounting inland water versus the relatively small increase in areal precision. There may be larger differences that can be accounted for with inland

water, depending on the source the researcher is querying for the verification figure. This step is not meant for anything other than verification that large areas are not either missing or being double counted.

It is now time to add the data that we will use to create the weights for the interpolation. As mentioned at the beginning of this section, those data are obtained from table B17001, which is found on American FactFinder. At this point, this example assumes that the researcher has summed the male and females under 5 years of age and save it with an appropriate name. In this example, the file was save as Poverty_data.xlsx. The name of the file doesn't matter, but that is the name that will appear in the formulas moving forward. To pull the data into the working sheet we will use the Excel Function *VLOOKUP*. The coding of the function is as follows:

	Location of the spreadsheet containing the data and the array containing the data. (The dollar signs [\$] anchor the array so it doesn't change when copied to all cells)	Always zero for exact matches
=VLOOKUP(A2,'E:\Dis_map\Poverty_data\[Poverty_data.xlsx]Sheet4!\$A\$2:\$F\$2128,6,0)		
Reference cell with the GEOID		Number of columns (including the reference column) between the reference and the data.

The above formula will be copied to all rows of the data array and that will bring over all relevant data from the poverty data table that was downloaded from the Census Bureau. One thing to note here is the total number of youth poverty will be much higher that would be the total for the state. This is due to the double counting that is a result of including data from two different, overlapping geographies. This issue was dealt with in

areally when county subdivisions were cut out of the school districts. At this point the researcher must do something similar with the poverty data. However, before that can be accomplished, the researcher must use the school district to county subdivision correspondence table that we created in step 5 of the section on preparing the shapefiles for use. To add those *GEOIDs* to the current spreadsheet, the researcher can employ, in a blank column, a formula similar to the following:

```
=IFERROR(VLOOKUP(A2,'D:\Dis_map\[CS-SD_corr.csv]CS-SD_corr'!$E$2:$U$467,17,0),-1)
```

This formula looks more complicated than it really is when you are writing it. The only difference, structurally, with this formula is the addition of an *IFERROR* function that surrounds the *VLOOKUP* function. This is included because we know there will be *GEOIDs* that will not be located as the list consists of both county subdivisions and school districts. The reason we need to add these *GEOIDs* in a new column is so that we can modify the poverty data to total the amount included from the county subdivisions and subtract that from the amount that is being allocated to the school districts.

Additionally, to accomplish the modification of the under-five poverty data, the addition of one more piece of data is necessary. The original area of the combined county subdivisions/school districts need to be saved to the working spreadsheet. This task is again accomplished with a *VLOOKUP*. That formula is structured the same as the previous two as follows:

```
=VLOOKUP(A2,'D:\Dis_map\[SDCS_area_cut.csv]SDCS_area_cut'!$A$2:$C$1007,3,0)
```

To accomplish the modification of the under-five poverty, the total number of the under-five population experiencing poverty that is represented in the school district geographies must be reduced by the number of the same group represented by the county subdivisions. In this example, the task is accomplished through the formula that follows:

```
=ROUND(IF(LEN(A2)=10,J2*(G2/H2),(J2-  
SUMIF('D:\Dis_map\Poverty_data\[Poverty_data.xlsx]Sheet4'!$C$2:$C$2128,A2,'D:\Di  
s_map\Poverty_data\[Poverty_data.xlsx]Sheet4'!$F$2:$F$2128))*(G2/H2)),0)
```

This formula looks more complicated than it is, especially to a researcher who may not spend a great deal of time working in a spreadsheet program. Note, there are no line breaks in Excel for formulas, and the breaks shown here are only due to space constraints. Also, the results of these calculations should never produce negative estimates. There may be very small fractions or zero values, but never negative numbers. If the results of this formula contain negative estimates, there has been an error in implementation.

The formula uses four Excel functions and some basic arithmetic to accomplish the task of modifying the poverty data to avoid double counting. The functions employed are *ROUND*, *IF*, *LEN*, and *SUMIF*. Before I get into the implementation of those function in the specific case above, it will be helpful to describe them in general terms. The *ROUND* function is very easy to understand, and only has two arguments. The first is the number to be rounded, and the second is the number of places. The *ROUND* function in basic terms looks like this, `ROUND([Number],[Places])`. I use this in the modification formula because there are no partial children. The next part of the formula above is the *IF* function and this is a powerful conditional function. Its basic function and format is, `IF([Logical Test],[Action if True],[Action if False])`. In the function above

the *IF* tests the *GEOID* with the *LEN* function, which returns the number of characters in a text string and is implemented as `LEN([Text])` and returns an integer. So, the *IF* function tests the length of the *GEOID* field because I am using it to separate the county subdivisions from the school districts because I want to do something different with them depending on that designation. I know that the length of a county subdivision *GEOID* is 10 characters long while a *GEOID* for a school district is only seven. If the *GEOID* is 10 characters long, meaning it is a county subdivision, I just multiply it by the proportion that is created by the $(G2/H2)$ part of the formula. In this case the G column is the column with the `fnl_area`, the final unioned area for the polygons, and the H column is the original area from the layer that was created when we combined the school districts with the county subdivisions. This is necessary because some of the polygons in the unioned dataset represent county subs that may have been split into multiple polygons in the union process. Most often this results in a polygon with a zero area, but sometimes there are separate polygons because municipalities or townships are not always contiguous. In both those cases, this multiplication resolves the multiple instances of the same municipality or other form of county subdivision in the data.

The *SUMIF* function is implemented as `SUMIF([Criteria Range], [Criteria], [Sum Range])`. The first argument tells the system by which criteria the entries should be considered for inclusion. In this example, the system is looking in the poverty data for *GEOIDs* that match the one in the second argument, which is the *GEOID* for the particular line in the data. The third argument is the range for the under 5 poverty that is to be summed. When those data are summed, the result is then placed in the formula to

be subtracted from the total for that geography and then multiplied by the same proportion described above.

When that formula resolved, the result is a modified poverty statistic that is one of two possible outcomes: 1- the amount of children under five years in poverty for a county subdivision that has been modified to account for split geographies, or 2- an accounting of the residual under five poverty that has been reduced by the amount registered for constituent county subdivisions and then areally disaggregated based on the proportion of the original school district area the polygon represents. This function is the whole reason so much care has been taken to find county subdivisions that are contained wholly within both school districts and PUMAS. The expectation is that by pulling portions of the population to known population centers, the number of children that were being areally interpolated would be reduced. This is also the main innovation incorporated since I produced my original estimates for the Michigan Department of Education for its internal formula testing. We will test later in this paper if the added work improves the estimates over what was produced with just an areal interpolation using only the school districts and PUMA shapefiles.

Software Choices (QGIS v. ArcGIS)

The main reason I have chosen to go through this detailed description of the interpolation process using QGIS is not because it makes the process easier, which is not the case. The choice to use QGIS for this explanation was made due to QGIS being open-source software that is free to download and use. Using the ArcGIS software would make the recording of polygon areas unnecessary and therefore the calculations

which are be the basis for the areal interpolations of the under-five population experiencing poverty.

In the previous step, I described the process of using two of the recorded areas to determine proportions of the original areas by which the under-five population was multiplied to arrive at the modified poverty data that will be the basis for the weights we will be using. This is unnecessary in ArcGIS as the user would attach the poverty data before the union process and then make the layers with the data a feature layer with the “Make Feature Layer” tool in ArcGIS. This will allow the user to enforce a “ratio policy” which would automatically distribute the data for the weighting variable spatially when the union process is run. This would make the step different in that the user would need to perform the parsing of the poverty data between the school districts and the county subdivision described in the last section and then attach the data to the combined county subdivision/school district layer prior to the union step. Once that was accomplished the system would automatically disaggregate the poverty data based on proportional area and the user could export that shapefile to a spreadsheet and proceed with the creation of the weighting variable that will be described later.

One software package is not necessarily better than another in this regard as the calculation steps to areally disaggregate the data are not difficult, but ArcGIS does eliminate several steps that can very easily introduce error into the process. If the user has access to the appropriately licensed ArcGIS software package (Analyst level license), I would suggest using that package over QGIS just to eliminate that part of the process. Though the wrangling of the data to modify the poverty data for the school districts

would still require the production of the correspondence tables, so the real benefit may be negligible.

Isolating the PUMAs and creating the weights

The process of creating the weights is straight forward as the heavy lifting has been done by the previous formula that modified the poverty data and areally distributed the school district amounts to the constituent parts of each school district. The weighting variable is created by employing some arithmetic and a SUMIF formula combined as follows:

$$=K2/SUMIF(\$E\$2:\$E\$1481,E2,\$K\$2:\$K\$1481)$$

This formula takes the modified poverty value, in column K, and divides it by the sum of the under-five poverty for all areas that match the value in column E, which is the value for the polygon's PUMA GEOID. This formula essentially provides the probability that a PUMA's youth poverty will be in the indicated polygon and will be the basis for the distribution of the target population's four-year-olds at 250 percent of poverty in the next step.

The weighting variable has an easy confirmation step. When the total weights are summed they will total the number of pumas that are contained in the total area. When these weights are totaled in this case, they sum to 68, which corresponds to the number of PUMAs in Michigan.

Making the Estimates

This step in the process will require that the user has familiarity with the PUMS data and knows how to use them to make estimates for a target subpopulation. The process is not hard, but the description of that procedure is beyond the scope of this discussion. Suffice it to say the process requires the user to have downloaded and combined the data into a statistical package capable of using a large dataset to produce estimates. To accomplish this, I used STATA to perform the operation with the 2015 1-year estimates. The command I used was as follows:

```
table puma if st==26 & agep==4 & povpip<=250 [pw=pwgtp]
```

This command provided an accounting of four-year-olds at 250 percent or less of poverty by PUMA area in Michigan. These will be the data that will be used to bring the target population counts into the active sheet. The process for bringing these data in to the active dataset is achieved through the simple VLOOKUP as follow:

```
=ROUND(VLOOKUP(D2,Sheet1!$A$1:$B$66,2,0)*L2,0)
```

This formula looks at the data that was obtained from the PUMS data and placed in sheet1 and then multiplies those data by the individual polygon's weight that was produced earlier in this step. The result of this calculation is the ultimate estimate of four-year-olds for the individual polygons in the union file. These can then be summed to the target geography or if the geographic units in the active sheet represent an intermediate geography, as they do in this example, a final correspondence table will be employed to add the appropriate codes to allow for aggregation to the ultimate geography.

To complete the estimates for this example, the intermediate school district (ISD) designation for each geography was added to the active data sheet from a correspondence table that was constructed from publicly available data. Once those codes were added through a simple *VLOOKUP*, they were then aggregated using the pivot table function in Excel. The product of that operation represents the completion of the estimation process and results in the table here with the “Row Labels” representing ISDs and the “Sum of Est” representing the number of four-year-olds at 250 or less of poverty:

Table C-1 Target Population Estimates

Row Labels	Sum of Est
3	1134
4	206
8	161
9	303
11	1249
12	197
13	1253
14	153
15	357
16	486
17	449
18	290
19	148
21	295
22	73
23	221
25	3107
27	55
28	1020
29	399
30	297
31	116
32	212
33	1718
34	237
35	161
38	1186
39	1944

41	3304
44	466
46	392
47	325
50	4569
51	167
52	158
53	357
54	346
55	121
56	122
58	122
59	342
61	977
62	360
63	5736
64	223
70	1689
72	272
73	1294
74	1112
75	375
76	335
78	386
79	423
80	563
81	1704
82	16307
83	572

Review the Estimates

With the completion of the estimation process, the researcher must review the estimates for face validity and prepare to evaluate the estimates in any way possible. An inherent problem with this procedure is that there is no source that produces a product available for public consumption against which these estimates could be judged. Were there such a product, this process would be unnecessary. Presumably the researcher has

some knowledge about the subject area and geography for which the estimates were produced, and it is that knowledge that the researcher must use to make sure the estimates do not have any obvious issues. For example, when I first produced these estimates for districts in Michigan, areas in the Thumb region of the state were being allocated a disproportionate number of the youth in poverty from the PUMA data. To discover this, I mapped the estimates to see what the distribution looked like from a holistic perspective. By the very nature of a process that areally distributes populations, there are going to be misallocations due to the varied nature of settlement and concentrations of social factors. In the case of the issues with the Thumb region, it seemed that the combination of large maritime, low population density areas, and relatively high poverty pulled more than the area's fair share of the distributed estimates. The efforts outlined here, namely removing the water areas, and first pulling weighing data to known population centers, helped to ameliorate these issues and produce more realistic estimates. More rigorous techniques will be applied to this method in the context of this work, but a researcher in the field employing this method will not have a comparator and will have to rely on her knowledge and the knowledge of any experts available for consultation.

Alternate Implementation

A geographic feature not readily apparent from the diagram of geographic hierarchy presented earlier is that census tracts nest within the Public Use Microdata Areas (PUMA). It is clear tracts nest within counties, but the relationship between tract and PUMAs is not as apparent because the PUMAs created following the 2010 census were the first set created with this geographic relationship. What this means is that estimates can be produced with the weighting method described above, but much of the GIS work is unnecessary because there is no need for areal distribution of the weighting variables. The areal distribution is unnecessary because both the county and PUMA level geography are both wholly made up of complete census tracts.

With perfect correspondence between counties/tracts and PUMA/tracts all that needs to be created through application of GIS techniques are the correspondence tables similar to those that were produced above for places and school districts. With those tables in place the weighting variables can be drawn in from data obtained for census tracts and the weights can be directly produced without the need for areal distribution. This cuts much of the work that was outlined above and allows for a relatively nimble process that can be used to produce a wide variety of estimates for counties.

The trick with producing estimates with this method, regardless of the geographic level, is to determine a weighting variable that has as high a correlation with the target social characteristic as possible. This method is built on the knowledge that social characteristics are not spatially distributed in the same way as the general population. If they were there would not necessarily be a need to spend time to determine the

appropriate weighting variable. A general population count would suffice, and any target variable could be distributed accordingly.

Brief Steps for an Alternative Implementation

1. Load requisite layers, in this case: PUMA10, tracts, and counties.
2. Filter layers for study area. For example, in the case of counties, the shapefile from the Census Bureau has all counties in the nation. Limit that file to just the counties about which the study or project is concerned.
3. Ensure all layers are using the same projection
4. Use the “Join by Location” feature in QGIS, or similar function in other GIS software, to create a layer that has all the census tracts and the counties in which they are contained. This was demonstrated about when the correspondence tables for the county subdivisions and school districts were created (a correspondence table for counties and tracts is not strictly necessary as tracts are a subcounty geography, but it makes it easier later on than trying to parse *GEOIDs* on the fly). Repeat this process with the tracts and PUMAs.
5. Save both layers created in the last step to CSV files.
6. Open CSV files and delete all rows from each except for the *GEOID* fields and names for both the tracts and the joined geographies. The joined geographies will have a “_1” appended to the field names to differentiate them from fields with the same name in the list of primary geographies.
7. Pick a primary file and use the *VLOOKUP* function to bring in the values for the final geographies to create a complete geographic correspondence table. This final table will have all census tract and their corresponding county and PUMA

values. The final table need only have six columns consisting of the *GEOIDs* and names for the three types of geographies in their respective columns.

8. Download the weighting variables data for all census tracts in the study area from the Census website.
9. Use *VLOOKUP* to bring in those data to the working correspondence table. A quick check here should reveal the totals for the weighting variable should be equal between the working sheet and the data downloaded from Census.
10. At this point the weights can be created from the data that was imputed in the last step in the same manner as they were created for the school districts in the detailed description above. The *SUMIF* function is used to divide each census tract's value by that for the sum of the PUMA of which the tract is a constituent part. Remember the sum of the weights should equal the number of PUMAs in the study area.
11. With the weights in place, the data for the PUMAs can be brought in from the PUMS data. The user will use another *VLOOKUP* to multiply the tract's weight by the value derived for the PUMA containing the tract. The results of this will be the estimates for the subpopulation of interest for each tract.
12. These individual level estimates can then be summed to the county geography.
13. If the county level is the final geography of interest, the estimation procedure is complete. If the final geography is a grouping of counties, that can be completed at this point.
14. The estimates need to be evaluated given whatever method is most applicable to the subject or geography involved.

Works Cited

- Beckett, Megan K. and Peter A. Morrison. 2009. "Assessing the Need for a New Medical School: A Case Study in Applied Demography." *Population Research and Policy Review* 29(1): 19-32.
- Bing. (n.d.). [Bing Map illustrating area in Michigan's Thumb Region] Retrieved December 28, 2017.
- Bakker, Bart F. M. 2012. "Estimating the Validity of Administrative Variables." *Statistica Neerlandica* 66(1): 8-17.
- DeNavas-Walt, Carmen, and Bernadette D. Proctor. 2015. *Income and Poverty in the United States*. U.S. Census Bureau. Current Population Reports P60-252, U.S. Government Printing Office, Washington, DC.
- Dinno A. 2017. tostsigrank: Test for stochastic equivalence on paired or matched data. Stata software package. <https://www.alexisdinno.com/stata/tost.html>
- Doyle, William R. and Benjamin T. Skinner. 2016. "Estimating the education-earnings equation using geographic variation." *Economics of Education Review* 53: 254-267.
- Frank, Andre I. 2003. "Using Measures of Spatial Autocorrelation to Describe Socio-economic and Racial Residential Patterns in the US Urban Areas." Pp. 146-161 in *Socio-Economic Applications of Geographic Information Science*, edited by D. Kidner, G Higgs, and S White. New York, NY: Taylor & Francis.
- Geography of Michigan. (n.d.) In *Wikipedia*. Retrieved October 2, 2017. from https://en.wikipedia.org/wiki/Geography_of_Michigan
- Horton, Hayward Derek. 1998. "Toward a Critical Demography of Race and Ethnicity: Introduction of the "R" Word." Sociology Faculty Scholarship. Paper 1. http://scholarsarchive.library.albany.edu/sociology_fac_scholar/1
- Horton, Hayward Derek. 1999. "Critical Demography: The Paradigm of the Future?" *Sociological Forum* 14(3): 363-367.
- Iceland, John and Erik Hernandez. 2017. "Understanding Trends in Concentrated Poverty: 1980-2014." *Social Science Research* 62: 75-95.
- Lazenby, Sara. 2011. "An Untapped Resource for Increasing College Attainment: Estimating the Population of Potential First-Generation Students in Wisconsin." *WISCAPE Policy Brief*. Retrieved October 13, 2016: <http://eric.ed.gov/?id=ED524265> .
- Logan, J. R., Xu, Z., and Stults, B. 2014. "Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database." *The*

Professional Geographer : The Journal of the Association of American Geographers, 66(3): 412–420.

- Males, Mike A. and Elizabeth A. Brown. 2014. "Teenagers' High Arrest Rates: Features of Yung Age or Youth Poverty?" *Journal of Adolescent Research* 29(1): 3-24.
- Merrick, Thomas. 1986. "Teaching Applied Demography." *Teaching Sociology*. 14(2): 102-109
- Michigan, State of. 2012. "Michigan School Districts - Framework V15." Lansing, MI: retrieved June 16, 2015
http://gis.michigan.opendata.arcgis.com/datasets/f40e3bf5815e4045a68c53af572690f6_10
- Morrison, Peter A. and Dean Judson. 2011. Integrating Census Data to Support a Motion for Change of Venue." *Population Research & Policy Review* 30(5): 801-815.
- Murdock, S.H. & Ellis, D. R. (1991). *Applied demography: An introduction to basic concepts, methods, and data*. Westview Press: Boulder, CO.
- Mulherin, Stephen. 2000. "Affordable housing and white poverty concentration." *Journal of Urban Affairs* 22(2): 139-156.
- Murembya, Leonidas and Eric Guthrie. 2015. *Detroit City Demographic and Labor Market Profile*. Michigan Bureau of Labor Market Information: Lansing, MI.
http://milmi.org/admin/uploadedPublications/2343_Detroit_City_Demographic_and_Labor_Mkt_Profile.pdf
- Murembya, Leonidas and Eric Guthrie. 2016. *Flint City Demographic and Labor Market Profile*. Michigan Bureau of Labor Market Information: Lansing, MI.
http://milmi.org/admin/uploadedPublications/2467_Flint_City_Demographic_and_Labor_Mkt_Profile.pdf
- New Jersey, Sate of. Department of Education. 2017. *2016-2017 Enrollment District Reported Data*. http://www.nj.gov/education/data/enr/enr17/stat_doc.htm
- Owens, Ann. 2015. "Housing Policy and Urban Inequality: "Did the Transformation of Assisted Housing Reduce Poverty Concentration?" *Social Forces* 94(1): 325-348.
- Poudyal, Neelm, Duncan Elkins, Nathan Nibbelink, H. Ken Cordell, and Buddhi Gyawali. (2016). "An exploratory spatial analysis of projected hotspots of population growth, natural land loss, and climate change in the conterminous United States." *Land Use Policy*, 51, 325-334.
- QGIS (n.d.) [Computer Software, version 2.18] Retrieved June 14, 2017. www.qgis.org
- Ranjith, Sri. And Anil Rupasingha. 2012. "Social and Cultural Determinants of Child Poverty in the United States." *Journal of Economic Issues*. 46(1): 119-141.

- Saporito, Salvatore and Deenesh Sohoni. 2007. "Mapping Educational Inequality: Concentrations of Poverty among Poor and Minority Students in Public Schools." *Social Forces* 85(3): 1227-1253.
- Schildberg-Hoerisch, Hannah. 2011. "Does Parental Employment Affect Children's Educational Attainment?" *Economics of Education Review* 30: 1456-1467.
- Siordia, Carlos, and Douglas F. Wunneburger. 2013. "Contiguity Principle for Geographic Units: Evidence on the Quantity, Degree, and Location of Public Use Microdata Area (PUMA) Fragmentation." *Human Geographies* 7(2): 5-13.
- Son, Le Hoang, Bui Cong Cuong, Pier Luca Lanzi, and Nguyen Tho Thong. 2012. "A Novel Intuitionistic Fuzzy Clustering Method for Geo-demographic Analysis." *Expert Systems with Applications*. 39(10): 9848–9859
- Swanson, David A., Thomas K. Burch and Lucky M. Tedrow. 1996. "What is Applied Demography?" *Population Research and Policy Review* 15(5): 403-418.
- Swanson, David. 2008. "Applied Demography in Action: A Case Study of 'Population Identification.'" *Canadian Studies in Population*. 35(1): 133-158.
- U.S. Census Bureau. 2009. *A Compass for Understanding and Using American Community Survey Data: What PUMS Data Users Need to Know*, U.S. Government Printing Office: Washington, DC.
- U.S. Census Bureau. 2011. *Final Public Use Microdata Area (PUMA) Criteria and Guidelines for the 2010 Census and the American Community Survey*. Suitland, MD: retrieved June 29, 2017
https://www2.census.gov/geo/pdfs/reference/puma/2010_puma_guidelines.pdf
- U.S. Census Bureau. 2014. *TIGER/Line Shapefiles, Michigan Public Use Microdata Areas*. Suitland, MD: retrieved June 16, 2015 <https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2015&layergroup=Public+Use+Microdata+Areas>
- U.S. Census Bureau. 2017 *U.S. Census Bureau's Budget Fiscal Year 2018: As Presented to Congress*. Suitland, MD: Retrieved July 25, 2017
<https://www2.census.gov/about/budget/FY2018-census-budget.pdf>
- U.S. Department of Commerce, Economic Statistics Administration, Bureau of the Census. 1994. *Geographic Areas Reference Manual*. Bureau of the Census: Suitland, MD.
- Wellek, Stephan. *Testing Statistical Hypotheses of Equivalence*. Boca Raton, FL: CRC Press.