Municipal Area Population Estimation Through Areal Interpolation of Census Data

Chad Ramos, Kennedy Jackson

Texas State University Department of Geography

Abstract

Small area population estimates play a critical role in a broad range of local, regional, and private sector planning procedures related to the allocation of funding and location of services. Traditionally, population estimates are done with the Housing Unit Method in which populations are estimated based on the number of housing units in an area and the average number of persons living in each unit. The Housing Unit Method, while conceptually simple, requires gathering precise information on both housing units and people per household across a study area, which can be tedious and time consuming. The research presented here provides an alternative method for small area population estimation based on spatial disaggregation and reaggregation of Census Bureau population data. We create a methodology for the estimation and then an ArcGIS Pro Python script tool to carry out the methodology. The tool is tested on many cities in Texas and the estimates are compared to authoritative estimates provided by the Texas Demographic Center-the state agency responsible for demographic data. The results show that the methodology can be highly accurate, but further testing is necessary to limit sources of error.

Keywords: Population Estimation, Housing Unit Method, Urban Planning

1 Introduction

Small area population estimates play a critical role in a broad range of local and regional planning procedures (Deng, Wu, and Wang 2010; Hoque 2012; Smith and Cody 2013). In the public sector, population estimates are used in decisions regarding the allocation of funding and for the locations of public facilities such as schools, housing developments, hospitals and public water, wastewater and transportation infrastructure (Deng, Wu, and Wang 2010; Hoque 2012; Smith and Cody 2013). In the private sector, population estimates are used for market area delineation, site-location analysis, etc. (Hoque 2012; Smith and Cody 2013). There are many methods for population estimation and of them, the Housing Unit Method is the most widely used for estimating small area populations in the U.S. (Deng, Wu, and Wang 2010; Hoque 2012; Smith and Cody 2013). The Housing Unit Method is an estimation technique used to determine the population of an area based on the number of housing units (HU) that exist in an area and the average number of people living in each unit (PPH) (Deng, Wu, and Wang 2010; Hoque 2012; Smith and Cody 2013). This is a popular technique because it is conceptually simple and can adapt to various data sources but, it is a general approach rather than a specific methodology and obtaining precise data for the number of people per household across a study area can be tedious and time consuming (Smith and Cody 2013). Therefore, the goal of this study is to create a widely applicable script tool that uses spatial methods and the 2020 U.S. Census Bureau Block population data (by default) to quickly and accurately estimate the population of user-input study areas. The methods will rely on authoritative census data but will also allow for alternate population sources provided by the user. The research will be targeted to municipal governments, but will be adaptable to a variety of public and private sector uses.

2 Literature Review

Small area population estimations are necessary for planning processes and to understand the growth and decline of a population (Deng et al. 2010). It also determines where resources are allocated for state and local governments Deng et al. 2010). There are multiple techniques for small area population estimation using census data. The component method II (CMII), the ratio-correlation method (RCM), and the Housing Unit Method are a few commonly reviewed in literature. The component method estimates population by including major components of local demographic change (Deng et al. 2010). The population is then estimated by taking the most recent census population, adding the estimated number of births, subtracting the number of

deaths, and then adding net migration and changes in group-quarter population (Deng et al. 2010; Hoque 2012). This method is reliable on the county level if birth and death data is available at the county level, but there are limitations (Hoque 2012). Birth and death data, private school enrollment data can be difficult to find especially in small places (Hoque 2012). The ratio-correlation method estimates population by taking into account school enrollment, car registration, workforce, and occupied housing units (HU) data (Deng et al. 2010). The RCM for place level is especially difficult on the place level because of the availability of data. On the county level, like the CMII, it can be reliable but also depends on how much data is available. In current literature, the housing unit method (HUM) has been reported as one of the most reliable method to use for small area population estimation (Deng, Wu, and Wang 2010; Hoque 2012; Smith and Cody 2013). This is also a method that the Census Bureau in Texas uses for population estimation (Hoque 2010).

The HUM method estimates population for each census area using this equation, Population = (HUs * PPH * VR) + (GH * PPH * VR) . HU = Housing units, PPH = persons per household, VR = vacancy rate, and GH = group housing units-apartments (Smith and Cody 2013). Each component can be found using a variety of data sources such as building permits, demolition data, and utility data (Hoque 2010; Deng et al. 2010), but there are issues with this. Some counties do not provide the U.S Department of Commerce or Texas State Data Center with building permit data or even issue them at all (Hoque 2010). Also, the HUM relies on the idea that everyone has the same housing structure, but these components are often unknown. Researchers have found that in areas losing population, the HUM overestimates the population, and underestimates in places growing at a rapid pace (Hoque 2010; Smith and Cody 2004).

For this study, a script tool will be created using spatial methods and the 2020 U.S. Census Bureau Block population data to estimate the population of user-input study areas quickly and accurately. Geographic information system techniques can provide an alternative approach for small-area population estimation (Hoque 2010).

3 Data and Methodology

To provide an alternative to the widely used Housing Unit Method of small area population estimation, in which population estimates are calculated based on the number of occupied houses in an area and the estimated people per household, this research makes use of Census Bureau population data (see Figure 1). Areal interpolation methods were performed on population data at the Census Block level. The data were disaggregated and then reaggregated based on the user-input study area using spatial estimation methods (Wu, Qui, and Wang, 2005).

3.1 Data Collection

In order to test the methods of population estimation at the municipal scale, multiple city limit shapefiles were needed. These were obtained from the websites of the City of Kyle, the city of San Marcos, and a large dataset of all Texas municipal city limits was obtained from the Texas Department of Transportation (TxDOT). The default population data for the estimation is the 2020 Census data at the Block level, which was obtained from the ESRI Living Atlas, provided by the ESRI Federal Data team. Finally, an authoritative source for city-level population estimation from the 2020 census was needed to compare against the results of the tool. This was obtained from Texas Demographics, the state agency responsible for, among other things, population estimates in Texas. See appendix 1.

3.2 Data Analysis

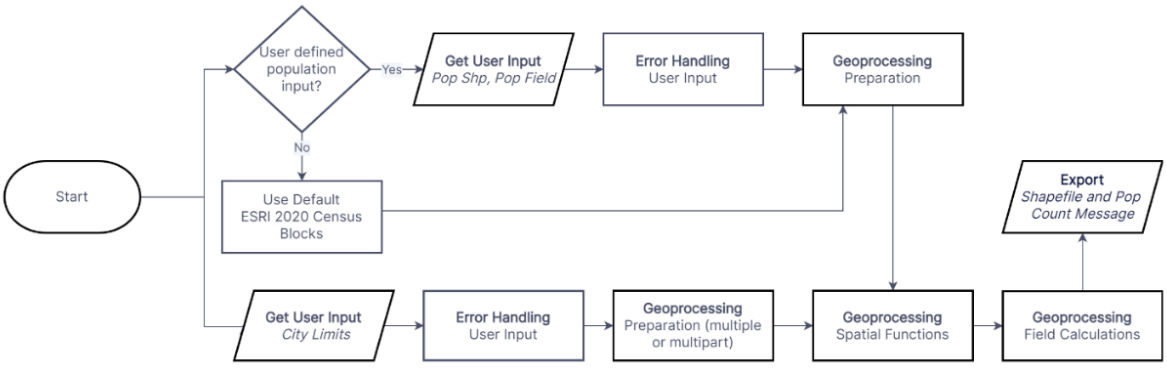

Figure 1 Conceptual Framework

The script tool uses as default 2020 U.S. Census Bureau population data at the Block level, for the input layer in the spatial disaggregation-with the option for the user to input other

population data in the correct format. The tool then takes the user input city limits (or other spatial area) as the overlay layer for the reaggregation (see Figure 1). The disaggregation is done by splitting the Census Block polygons along the boundaries of the input City Limits polygon. The reaggregation for edge-case census blocks will use the pre-split population density multiplied by the post-split area, to re-calculate the population for the portion that falls inside the City Limits overlay polygon. The final population estimation is then calculated as a summation of the populations for all census block polygons that fall completely inside the input City Limits area. Because this technique relies on population density, which is less accurate for larger Census areas, the default is to use Census Blocks.

3.3 Pseudocode

1. Get input parameters
    a. Required user input study area
    b. Optional study area dissolve field
    c. Optional user submitted population data
    d. Optional field containing population data
    e. Required output table
2. Set path for ESRI Living Atlas population data
3. Set ArcGIS Pro Project as CURRENT and get map as Active Map
4. Perform error checks and provide messages on user input study and dissolve field
5. Perform error checks and provide messages on user input population data, if provided
6. Import ESRI data if no user submitted population data
7. Set variables and pass error-checked input parameters into tabulate intersections tool
8. Pass output from tabulate intersections into statistics tool to sum and group by
9. Add and drop fields to clean output table
10. Add output table to map

4 Results

The research provides an alternative to the widely used Housing Unit method of population estimation by creating a script tool that uses spatial disaggregation and reaggregation of 2020 U.S. Census Bureau Block population data (by default) to estimate the population within a user-input municipal city limits polygon. As with the Housing Unit Method, providing an

accurate and authoritative population estimate is useful for a variety of municipal government functions related to city planning, budgeting, and emergency management. Further, the user has the ability to change the source of population data, which makes it adaptable to older census data and allows the tool to be useful beyond the limitations of the default 2020 Census data.

4.1 Tool Function

The tool functions as expected, allowing the user to estimate the population of a given study area based on 2020 Census block-level data. The tool connects to and imports the population data from the ESRI Living Atlas. Warning and error messages are provided at appropriate and necessary steps, informing the user of both the tool process and if necessary, the reasons why the tool has failed. Runtimes can be excessively high if the user inputs multiple study areas in one shapefile, and the tabulate intersections system tool, which the research tool incorporates, will output a maximum of 5 features at a time. Overall, it is far faster to run the tool multiple times rather than run the tool on the maximum 5 features at a time.

4.2 Edge-Case Methodology

A large part of the accuracy of this tool is dependent on the manner in which it accounts for the disaggregation and reaggregation of census blocks before summing the population for the input study area. This is done by apportioning the population of each edge-case census block to the study area based on the proportion of the overlap between the edge-case census block and the

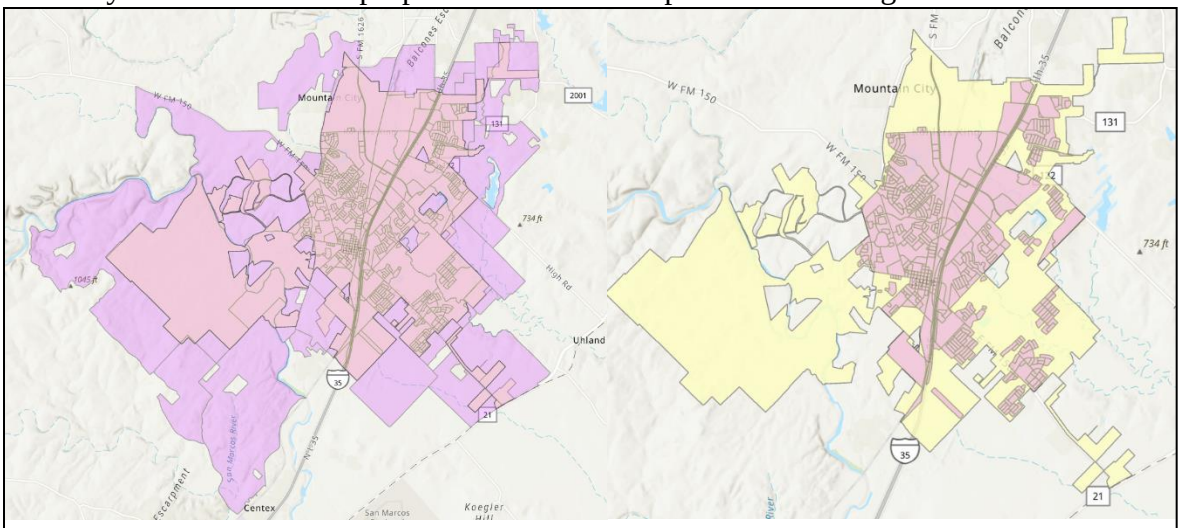

Figure 2. The two extremes of ignoring edge cases: selecting all census blocks intersecting the study area (left) and selecting only census blocks completely within the study area (right). Map created by authors.

study area. Given that census blocks are created to closely match real-world boundaries, it could be argued that ignoring the slight overlaps in edge cases would not greatly affect the calculation of population and that the disaggregation and reaggregation are not necessary. To test this, the population for the City of Kyle was calculated first by summing the population of all census blocks that intersect the municipal city limits, regardless of the amount of overlap, and second by summing the population of all census blocks that are completely within the municipal city limits. These two cases represent the extremes of ignoring edge-cases: summing all the blocks that intersect and summing only those that are completely within. The accuracy of these two cases, i.e. the percent difference between the estimated populations and the known, authoritative population given by the Texas Demographic Center population data, was compared to the accuracy given by research tool, which *does* account for edge cases (see Figure 2).

The calculations (see Table 1) indicate that the manner in which the tool apportions the population of edge-case census blocks makes a substantial difference in the population estimation as compared to the two extremes of ignoring edge cases. This is despite the fact that census blocks, being the smallest census designated area, typically align with real objects such as streets. Furthermore, the results indicate that a high degree of accuracy is possible, as the tool estimate shows only a 1.10% difference as compared to the authoritative estimate given by the Texas Demographic Center.

Estimation Type	**Population**	**% Difference**
Texas Demographic Center (Authoritative)	45,697	n/a
All Intersecting Blocks	54,942	20.23%
Completely Within	36,604	-24.84%
Research Tool	46,204	1.10%

Table 1. Population estimates based on the two extremes of edge cases.

4.3 Tool Accuracy

The overall accuracy of the tool was tested on a set of random and non-randomly selected cities from the TxDOT dataset of municipal city limits. The selected cities represent a range of possible city sizes and populations, as well as a mix of boundary types-cities that are landlocked

by other cities and cities whose border is clear of any other cities. There is a high variation in the accuracy of the tool population estimates across cities, as measured by the percent difference between the tool estimates and the authoritative Texas Demographic Center estimates (see appendix 2). The accuracy for the majority of the tested cities is less than 3% different than the Texas Demographic Center estimates, indicating again that the tool can be highly accurate. However, several cities have exceedingly high error above 10%, with some as high as 65%. Further testing is needed to determine the source of these high errors before the tool can be considered reliable.

5 Conclusion

The tool created from the research presented here provides an alternative to the popular, yet tedious Housing Unit Method of small area population estimation. The tool requires nothing more from the user than a feature class or feature layer of the study area and requires none of the tedious calculations associated with the Housing Unit Method. It has been shown here that the tool can be highly accurate when compared to authoritative Texas Demographic Center municipal population estimates made using the same 2020 Census Data. However, there are several test cities for which the research tool estimation was above 10% different than the authoritative data, and in some cases the percent error is as high as 65%. All test cities with an error above 8% are located along the Texas-Mexico border, which may indicate the limitations of spatial estimation methods in areas where the Census itself is difficult to complete. Further, the tool is limited to run on no more than 5 input study areas at a time, as set by the Tabulate Intersections ArcGIS Pro system tool. Until the accuracy issues are resolved and the functionality is improved, the usefulness of the tool is somewhat limited. Further research is needed into both the error and the functionality. First, the source of the high error must be determined and then, if any correlation between city properties (size, perimeter length, etc.) could be found, the relation could used to model the error to further improve accuracy. Lastly, the functionality of the tool could be improved to improve processing speed and to account for the maximum input into the Tabulate Intersections ArcGIS Pro system tool.

6 Works Cited

Deng, C., C. Wu, and L. Wang. 2010. Improving the housing-unit method for small-area population estimation using remote-sensing and GIS information. *International Journal of Remote Sensing* 31 (21):5673–5688. https://www.tandfonline.com/doi/full/10.1080/01431161.2010.496806 (last accessed 20 March 2023).

Hoque, N. 2012. Evaluation of small area population estimates produced by Housing Unit, Ratio-correlation and Component Method II compared to 2000 Census counts. *Canadian Studies in Population* 39 (1–2):91. https://journals.library.ualberta.ca/csp/index.php/csp/article/view/17838 (last accessed 20 March 2023).

Smith, S. K., and S. Cody. 2013. Making the Housing Unit Method Work: An Evaluation of 2010 Population Estimates in Florida. *Population Research and Policy Review* 32 (2):221–242. http://link.springer.com/10.1007/s11113-012-9265-2 (last accessed 20 March 2023).

Smith, S. K., and M. Mandell. 1984. A Comparison of Population Estimation Methods: Housing Unit versus Component II, Ratio Correlation, and Administrative Records. *Journal of the American Statistical Association* 79 (386):282–289. https://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10478042 (last accessed 16 March 2023).

Wu, S., X. Qiu, and L. Wang. 2005. Population Estimation Methods in GIS and Remote Sensing: A Review. *GIScience & Remote Sensing* 42 (1):80–96. https://www.tandfonline.com/doi/full/10.2747/1548-1603.42.1.80 (last accessed 15 March 2023).

7 Appendix 1: Data Sources

Data	Purpose	Source
City of Kyle municipal boundaries	Input for testing accuracy of population estimation	City of Kyle website https://city-of-kyle-maps-giskyle.hub.arcgis.com/datasets/GISKyle::jurisdiction-1/explore?location=29.986848%2C-97.776775%2C11.69
City of San Marcos municipal boundaries	Input for testing accuracy of population estimation	City of San Marcos website https://data-cosm.hub.arcgis.com/datasets/city-limits/explore?location=29.868371%2C-97.930650%2C11.91
TxDOT statewide dataset of municipal boundaries	Input for testing accuracy of population estimation	TxDOT website https://gis-txdot.opendata.arcgis.com/datasets/09cd5b6811c54857bd3856b5549e34f0_0/explore?location=31.009000%2C-100.168292%2C6.44
2020 Census population data at the block level	Population data input for tool testing	ESRI Living Atlas and the ESRI Federal Data team https://www.arcgis.com/home/item.html?id=b3642e91b49548f5af772394b0537681#overview
Texas Demographic Center population estimates	Authoritative population estimates for testing tool accuracy	https://demographics.texas.gov/data/tpepp/estimates/

8 Appendix 2: Accuracy results from select cities

City	Texas Demographic Center Estimate	Tool Estimate	Percent Difference
McAllen	142210	142207	0.01
Laredo	255205	255279	0.03
Buda	15108	15114	0.04
Lockhart	14379	14371	0.06
Pharr	79715	79658	0.07
Edinburg	100243	100148	0.09
San Juan	35294	35330	0.1
Mission	85778	85686	0.11
Benbrook	24520	24489	0.13
Taylor	16267	16295	0.17
Weslaco	40160	40089	0.18
Conroe	89956	90141	0.21
San Marcos	67553	67801	0.37
Alamo	19493	19413	0.41
Mercedes	16258	16330	0.44
La Villa	2804	2818	0.5
Kerrville	24278	24431	0.63
Palmview	15830	15716	0.72
Waco	138486	139733	0.9
Granjeno	283	286	1.06
Kyle	45697	46204	1.11
Austin	961855	943826	1.87
Sullivan City	3908	3982	1.89
Progreso	4807	4651	3.25
Elsa	5668	5470	3.49
Hidalgo	13964	13423	3.87
Rio Bravo	4450	4228	4.99
Penitas	6460	6054	6.28
El Cenizo	2540	2317	8.78
Pleasanton	10648	9679	9.1
La Joya	4457	4907	10.09
Edcouch	2732	3096	13.32
Roma	11561	9002	22.13
Rio Grande City	15317	11194	26.92
Escobares	2588	11 39	55.99
La Feria	6817	2344	65.62
Santa Rosa	2450	831	66.08