

Cadastral-based expert dasymetric system (CEDS) using census and parcel data

Georgianna Strode, Victor Mesev

* Florida State University

Abstract. Population censuses are commonly aggregated and mapped as uniformly distributed areal units. This leads to generalization of geographical patterns where occupied land is indistinguishable from unoccupied land. Many applications, including demographic profiling, calculation of vulnerable populations requiring access to healthcare, and electoral districting can be measured more precisely if cartographic representations are more disaggregate. An example of disaggregate mapping is the dasymetric principle, which has been widely used to estimate the population of census tracts which are classified as occupied. These methods include simple areal weighting, centroid-based moving windows, land use interpolation from remote sensing and other more complex mathematical models. Each attempts to combine two or more data sets, redistribute population across known occupied land use, and some even abide to the smoothing and mass-preserving principles known as pycnophylactic interpolation. However, some of these methods are rigid and at best only replicate the output from basic choropleth maps by assuming uniformity across occupied space or by overly-reapportioning population, neglecting the underlying population distribution, and operating at single scales of census collection.

In contrast, the cadastral-based expert dasymetric system (CEDS) is a non areal weighting algorithm that interpolates census data using cadastral parcel data at multiple scales. In demonstrating, this paper will explore the United States population census at three scales—tract, block group and block level—as well as cadastral parcel data (sometimes referred to as taxlot data) containing information on land use type, number of residential units (RU), and number of square feet of living area (RA). The objective of our CEDS example is to estimate the population of a parcel, where:

$$POP_1 = POP_c * U_1 / U_c$$

POP_1 = dasymetrically-derived cadastral parcel level estimated population

POP_c = census population (at the tract, block group, or block level)

U_l = the number of proxy units at the cadastral parcel level (RU or RA)

U_c = the number of proxy units at the census level (RU or RA per tract, block group, or block level).

The CEDS method calculates error rates for the number of residential units and square feet of living area. It then chooses the model that best fits each individual census polygon (across all three scales). The CEDS method has been tested on cadastral parcel data for the county of Hillsborough in the US city of Tampa, Florida. It has shown to perform well against many common methods of population estimation, and offers flexibility in terms of choice of predictive factor and scale of population information. Our research found no significant differences between final population estimates using three predictive factors but found significant differences between two scales of census population data disaggregation.

Keywords: cadastral, CEDS, population census, dasymetric mapping

1. Introduction

Precise population is important in planning, health, crime, and emergency management fields because it provides knowledge of where services are needed. As with most nations, the United States Population Census is the most commonly used because it is readily available, affordable, and is in a consistent format across the years. However, the traditional drawback is that it is tied by confidentiality and is required to aggregate census returns and represent them cartographically using areal units, that commonly bear little relation to the underlying geography.

Dasymetric mapping has been used as an alternative vehicle for estimating census populations. Its main variation is that spatial data are re-sampled into finer-scale units using additional data. For population estimation, dasymetric methods usually combine census population data with any type of local ancillary information that can indicate residential or non-residential areas. Many types of ancillary data are used, such as areal photographs, satellite sensor images, LiDAR, topographic land use maps, emergency response databases, street networks, electrical hookup databases, soil impact information, cadastral data, and others.

We chose cadastral data, also referred to as *parcel* or *taxlot* data, to demonstrate the utility of the ancillary data principle, and the cadastral-based ex-

pert dasymetric systems (CEDs). The CEDs method has high accuracy when tested against areal weighting, filtered areal weighting, and centroid-based methods. CEDs has also proven to be robust when tested for pycnophylactic interpolation (Maantay 2007). Figure 2 shows an example of the population estimation results of the CEDs method. Maantay et al (2009) recommended more testing of proxy units and disaggregation scales, to which our work will evaluate three proxy units and two census disaggregation units through pycnophylactic testing.

1.1. Pycnophylactic principle

The pycnophylactic principle is widely used for testing population estimates. It focuses on “volume-preserving,” or maintaining values for known populations (Tobler 1979). With vector data, the principle can be applied by using three levels (scales) of data: high, medium, and low (Figure 1). The high and medium populations are known values and the low level is the estimated population. High level data are disaggregated to the low level, then re-aggregated to the medium level for the purpose of testing. The known medium-level counts are compared to the estimated medium-level counts. The closer the estimated value is to the observed or known value, the better the method is thought to perform (Maantay, 2009). Figure 2 illustrates using the pycnophylactic testing on vector data.

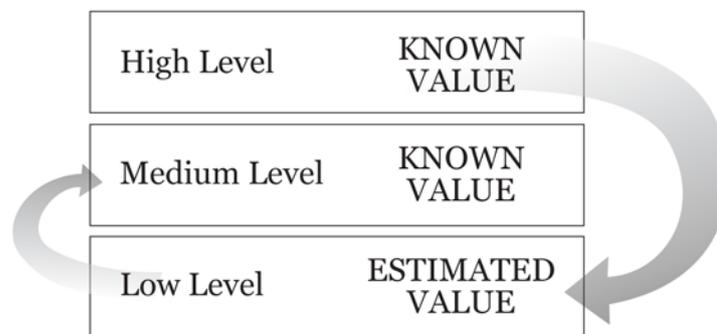


Figure 1. Illustration of the pycnophylactic testing method.

Population Distribution Downtown Tampa, Florida

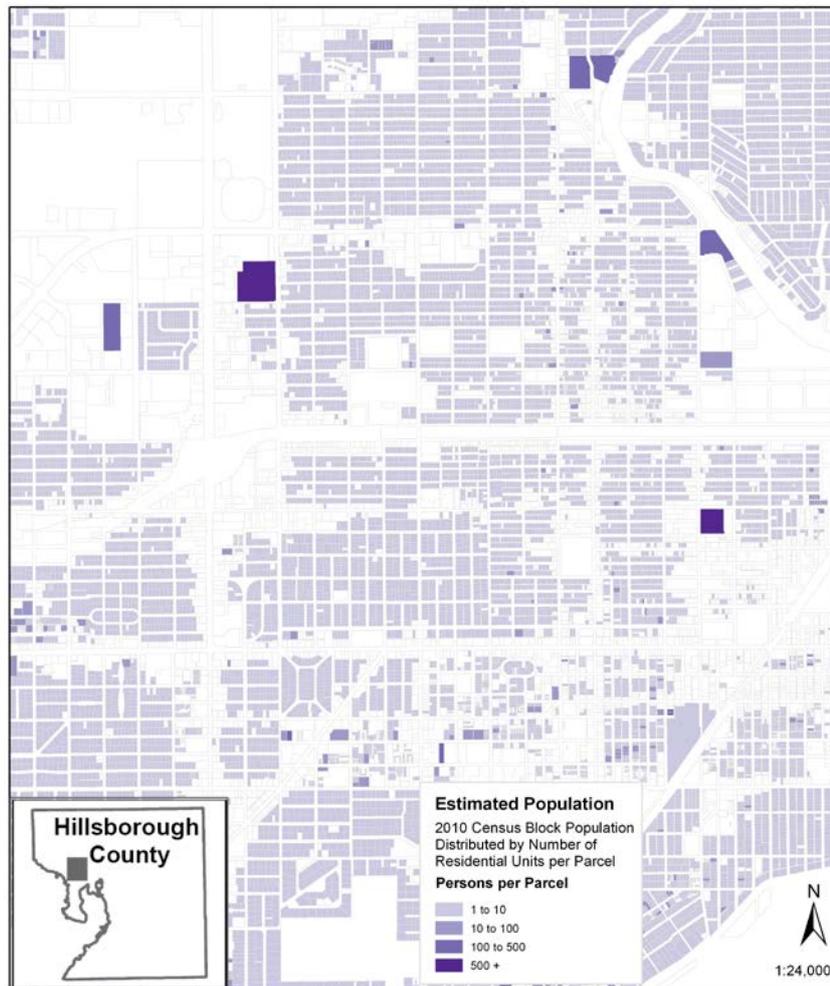


Figure 2. Map showing population in Tampa, Florida. This dasymeric map was created by combining U.S. census data with Florida Department of Revenue cadastral (*parcel* or *taxlot*) data.

1.2. Census data

The pycnophylactic principle is applied to the small-area US Population Census at three scales:

Block: The block is the smallest unit for which census data is reported. The boundaries of a block are usually set at the local level and are delineated by topography. A block can be a city block, bounded by streets, or it can be a larger area delineated by roads, city or neighborhood boundaries, water features, or any other topological feature. A block can never cross a county boundary or another census boundary. The population counts for blocks run from zero to several hundred people or sometimes over a thousand. Because of data suppression, census blocks can be overly aggregated and impossible to determine if a block's population is actually zero or if the data have been suppressed due to confidentiality.

Block Group: Block groups are a collection of blocks, with the national average being 39 blocks per block group. Block group population counts range from 600 to 3,000 people with 1,500 persons being optimal. The number of block group housing units range from 250 to 550 with the medium count 400.

Tract: A census tract is a collection of block groups designed to be homogeneous at the time of delineation when considering population characteristics, economic status, and living conditions. Tracts contain up to 9 block groups, and tract population can range from 1,500 to 8,000 persons with the optimal number being 4,000. The spatial size of a tract can vary according to population density. A census tract contains 5000 people. Typically, there are 4 block groups in the tract, each containing around 1,250 people, and each block group contains about 1,000 parcels (cadastral units).

1.3. Cadastral data

The 2011 cadastral data are from the Florida Department of Revenue (FDOR). There are only five fields of interest: block group, block, landuse code, number of residential units, number of square feet of living area.

1.4. Overview of all data

The CEDS method relies upon multiple scales of census data combined with parcel data, and the pycnophylactic principle tests disaggregation methods by comparing known and estimated populations, thereby requiring multiple scales of census data. Traditional work using CEDS uses three data layers as shown in Figure 3. The tract data are disaggregated to the parcel level and re-aggregated to the block group level for pycnophylactic testing. Block data have not been included in studies to date because of the known issues with data suppression. The pycnophylactic test to be performed on data below is: Tract → Parcel → Block Group (T → P → BG). Our work seeks to explore block data, so an additional pycnophylactic test is performed: Block Group → Parcel → Block (BG → P → B).

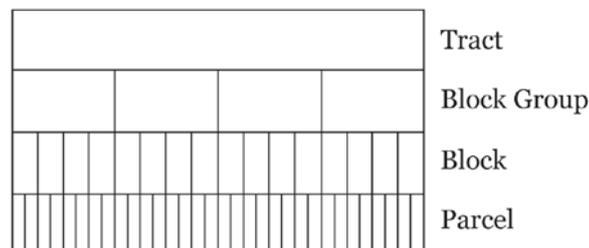


Figure 3. Hierarchical comparison of the four data layers.

2. Methodology

2.1. Study area

The study area is the county of Hillsborough in Tampa, Florida. It is the fourth most populous county in the state with a 2010 population of 1,229,000 within which Tampa is the largest city (335,000 people), but much of the county (84%) is unincorporated. The study area was chosen for its high population variation and sharp contrast between urban and rural.

2.2. Definition of parcels representing residential landuse

Land use data typically use broad categories that preclude residency in any area not zoned residential. Cadastral data more realistically resemble the underlying landscape by including a land use code and other residency-related features such as number of residential units and number of square feet of living area on each parcel. Cadastral data make it possible to identify residences not located in typically residential areas: churches can offer residence to the pastor; farmers often reside on their agricultural lands; and the owners of small businesses can reside on their income-producing property.

The definition of parcels with residential potential for the study includes parcels where one or more residential units are reported in the parcel record and the primary landuse code is not 'vacant residential.' This definition includes parcels that are not necessarily zoned residential but nevertheless include on-site living facilities, presumably for owners and employees. Examples are mixed use commercial/residential, motels, agricultural tracts, prisons, churches, homes for the aged, state parks (park rangers), and others. It does not include facilities for temporary overnight stays such as hotel rooms, vacation destinations, or hospital rooms. Hillsborough County has 386,255 of 465,848 parcels considered to have residential potential for the purpose of this study.

2.3. Proxy unit overview

A proxy unit is a predictive factor in population estimation. The CEDS method selects number of residential units (RU) and square feet of living area (RA) as predictive factors (proxy units). We also add a pre-calculated estimation of household size that is available from Social Explorer. A pycnophylactic test can show which proxy units are more accurate for individual block groups. The CEDS recommends selecting the factor that best represents each census block group individually.

Argument *in favor of* changing proxy units for each block group:

- Can claim best pycnophylactic accuracy for each block group.

Argument *against* changing proxy units for each block group:

- Data are processed in different methods across study area
- Requires extra work to change methods

2.4. Disaggregation scale overview

The pycnophylactic principle requires three levels for testing in a vector environment. Maantay et al (2009) used block group disaggregation and has already proven the CEDS method to be pycnophylactically sound. Our research seeks to further explore the CEDS method by testing the disaggregation unit.

Argument *in favor of* block disaggregation (against block group disaggregation):

- Theoretically is more accurate because a block is smaller than a block group; however there is the possibility of block data suppression for privacy purposes. (If block data suppression has occurred, then the best scenario is to use block group disaggregation for that area, which results in multiple methods being used).

Argument *in favor of* block group disaggregation (against block disaggregation):

- Consistent data processing across study area.
- Easier to process with a single method.

2.5. Proxy units

At the block group level, there are three proxy units to be evaluated: number of residential units (RU) and number of square feet of living area (RA) (both from the cadastral database), and Pre-calculated household size estimation (from the Social Explorer table, field T064_001, defined as “average household size is a measure obtained by dividing the number of people in households by the number of households....average household size is rounded to the nearest hundredth.”) At the block level, there are only two proxy units available to estimate population because the pre-calculated estimation field is only available at the block group level. The two proxy units are RU and RA, as defined in the preceding section.

3. Proxy Unit Analysis

3.1. Block group level

Figure 4 shows a visualization of parcels with proxy units that best represented their block group pycnophylactically. Each block group has an error rate calculated for each of the three proxy units. The proxy unit with the lowest error rate is thought to have performed best pycnophylactically.

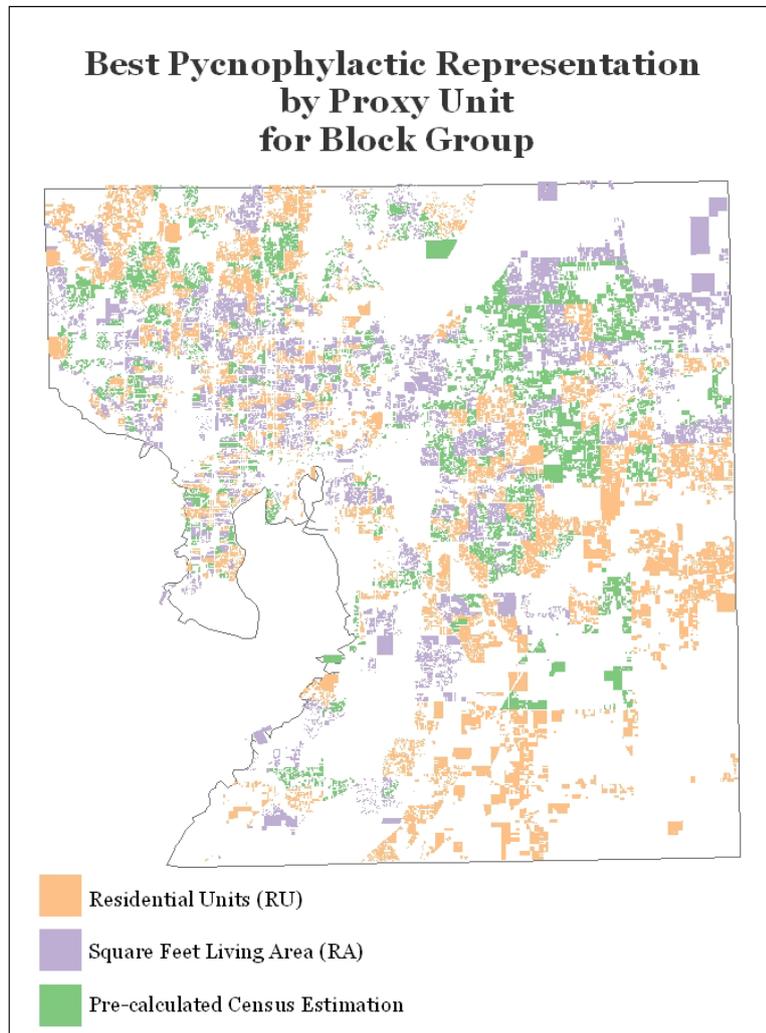


Figure 4. Visualization of parcels color-coded by the proxy unit that best represented the block group pycnophylactically.

3.2. Histograms of proxy unit error

Pycnophylactic errors are measured at the block group and block levels, not at the parcel level. Five 10-bin histograms were created to visualize the errors between proxy units and disaggregation scales. Most errors are small in size, there are a few outliers with larger errors for each of the proxy units, and the error pattern is repeats for each proxy unit and disaggregation scale. Figures 5 through 7 show the errors for the three proxy units.

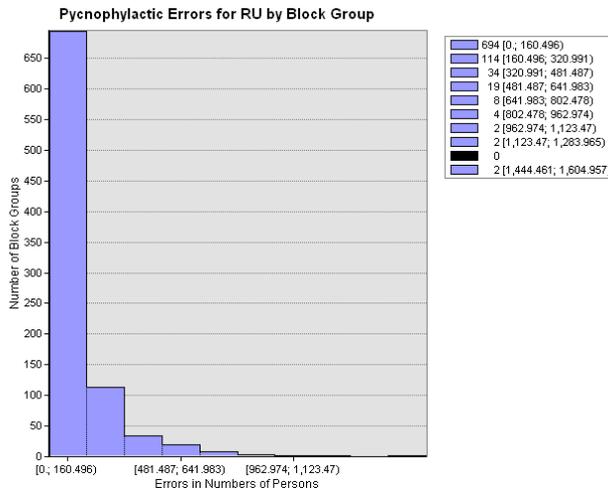


Figure 5. Histogram of errors for number of residential units at block group

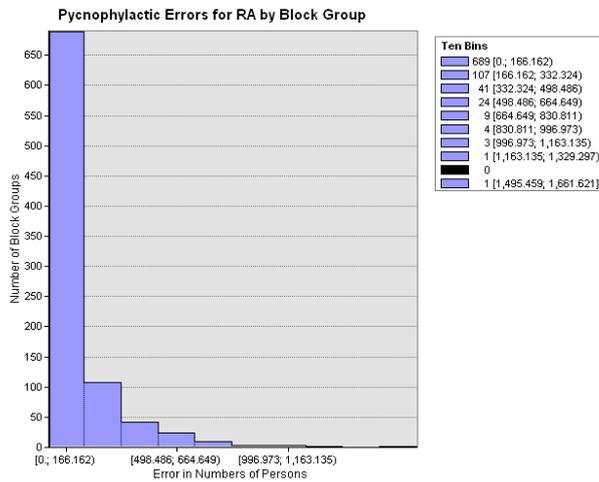


Figure 6. Histogram of errors for square feet of living area at block group.

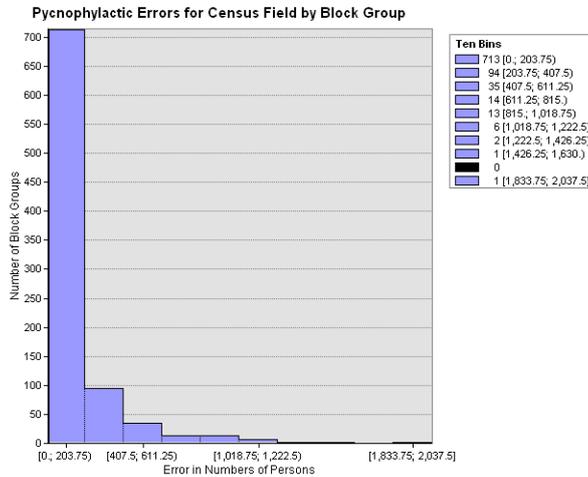


Figure 7. Histogram of errors for the pre-calculated census estimation field at block group

3.3. Statistical analysis of proxy unit error

This test measures general error at the block group level for each of the three proxy units to determine whether there are differences in proxy units. This is testing the block group level error counts for three proxy units: The testing population was chosen by randomly selecting one parcel from each block group to avoid spatial autocorrelation issues and examining the errors resulting from each proxy unit. Results reveal very significant errors between the three proxy units at block group.

- There was not quite a significant difference in the scores for all error counts at the block group level for number of residential units (M=104.34, SD=174.37) and number of square feet of living area (M=112.93, SD=177.52) conditions; $t(878)=1.8965$, $p=0.0582$.
- There was very significant difference in the scores for all error counts at the block group level for number of square feet of living area (M=112.93,SD=177.52) and pre-calculated census estimation (M=132.62,SD=210.99) conditions; $t(878)=3.2062$, $p=0.0014$.
- There was extreme significant difference in the scores for all error counts at the block group level pre-calculated census estimation (M=132.62, SD=210.99) and number of residential units (M=104.34, SD=174.37) conditions; $t(878)=5.3709$, $p=0.0001$.

3.4. Statistical of final population estimates

Pycnophylactic errors are measured at the block group level, not the parcel. Even though error analysis at the block group level shows significant errors between proxy units, we must remember that the goal of the CEDS method is to estimate a population for each parcel. This test of population estimation is more practical as it focuses more on our final goal of population estimation. Three paired-samples t-tests were conducted to compare the estimated number of occupants of a land parcel using three different predictive factors for population estimation. The purpose of these tests is to determine if there are predictive factors that perform better than others. The predictive factors are number of residences, number of square feet of living area, and the pre-calculated household average from census data for population estimation.

We chose two random parcels from each block group in the county to avoid spatial autocorrelation issues. The following t-tests were conducted using the two randomly selected parcels from each block group in the county:

- There was not quite significant difference in the scores for number of residential units ($M=9.32$, $SD=70.83$) and the number of square feet of residential area ($M=8.7$, $SD=66.03$) conditions; $t(1757)=1.6557$, $p=0.0980$.
- There was no significant difference in the scores for number of square feet of residential area ($M=8.7$, $SD=66.03$) and the pre-calculated census estimation ($M=8.33$, $SD=62.51$) conditions; $t(1757)=0.4924$, $p=0.6225$.
- There was no significant difference in the scores for the pre-calculated census estimation ($M=8.33$, $SD=62.51$) and the number of residential units ($M=9.32$, $SD=70.83$) conditions; $t(1757)=1.5224$, $p=0.1281$.

3.5. Proxy unit visualization

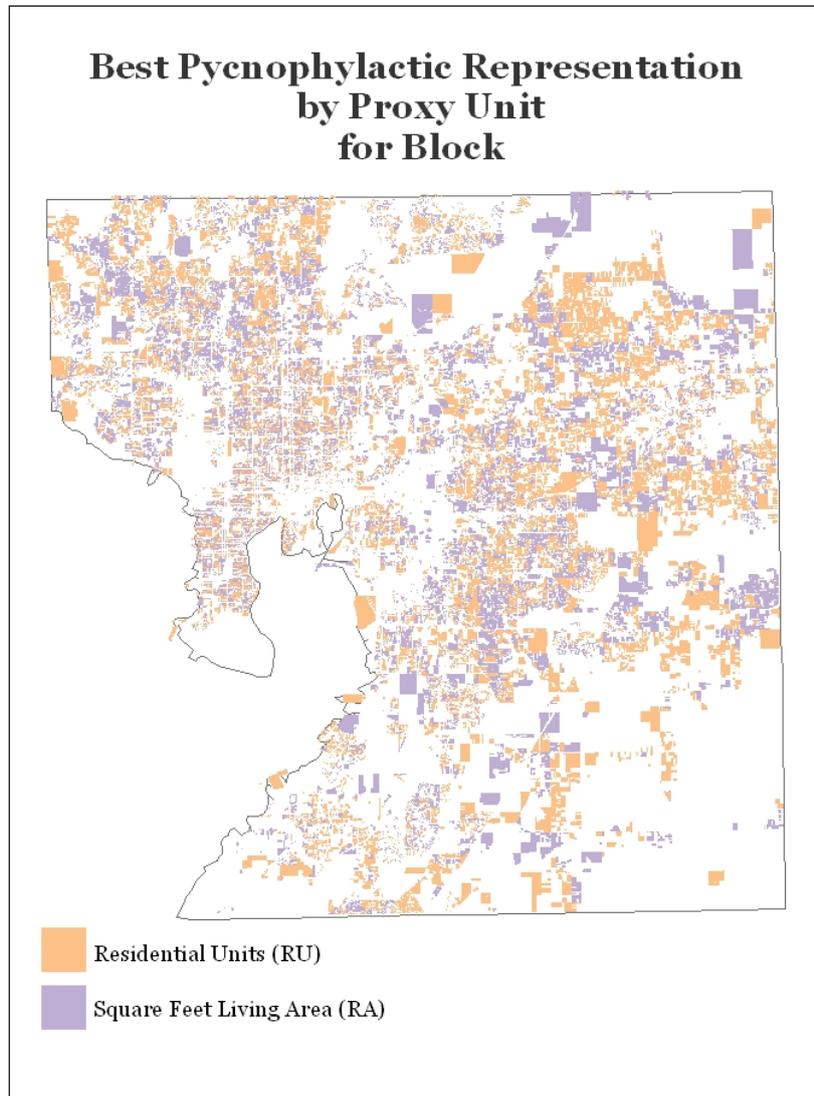


Figure 8. Visualization of parcels color-coded by the proxy unit that best represented the block pycnophylactically.

3.6. Histograms of proxy unit error

Pycnophylactic errors are measured at the block group and bock levels, not at the parcel level. Five 10-bin histograms were created to visualize the errors between proxy units and disaggregation scales. The histograms reveal that most errors are small in size. There are a few outliers with larger errors for each of the proxy units. Figures 9 and 10 show this error pattern repeats for each proxy at the block disaggregation scale.

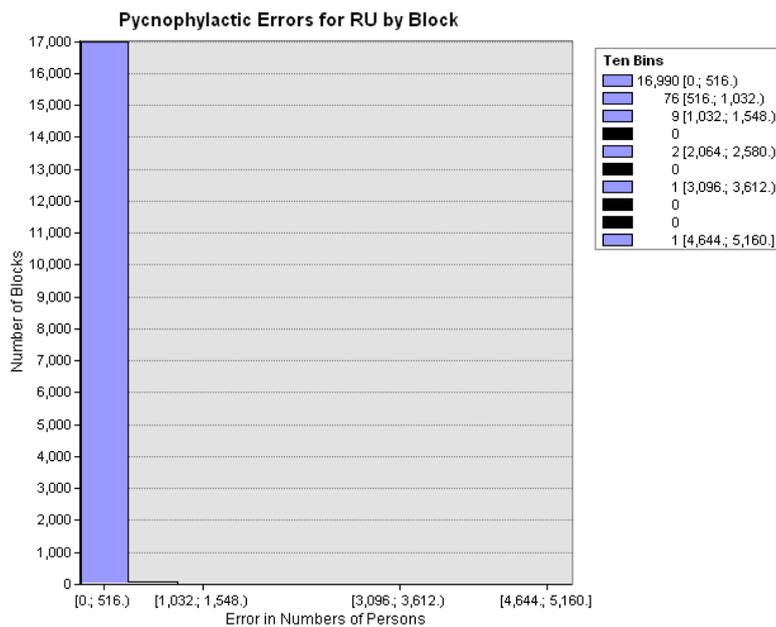


Figure 9. Histogram of errors for number of residences at the block level.

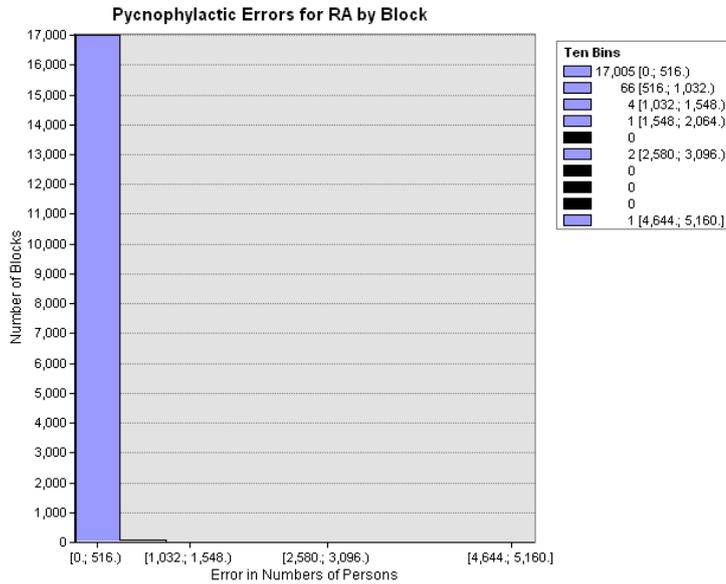


Figure 10. Histogram of errors for square feet of living area at the block level.

3.7. Statistical analysis of proxy unit error

The purpose of this test is to analyze the general error at the block level for both proxy units to determine whether there are differences in proxy units. To avoid spatial autocorrelation, the samples tested were created by randomly selecting one parcel from each block group.

The errors resulting from each proxy unit.

- There was no significant difference in the scores for the number of residential units ($M=53.32$, $SD=130.64$) and the number of square feet of residential area ($M=52.44$, $SD=121.32$) conditions; $t(878)=0.5571$, $p=0.5776$.

The test was repeated using different samples and the results are different. The population was chosen by randomly selecting one parcel from each block, then choosing 2000 parcels randomly from this group:

- There was extreme significant difference in the scores for the number of residential units ($M=24.72$, $SD=52.37$) and the number of square feet of residential area ($M=26.52$, $SD=50.82$) conditions; $t(1999)=3.5497$, $p=0.0004$.

3.8. Statistical of final population estimates

A paired-samples t-test was conducted to compare the estimated number of occupants of a land parcel using two predictive factors for population estimation. This test repeats the block group level tests previously mentioned but is included to determine if a predictive factor performs better at the block level.

- There was not quite significant difference in the scores for the number of residential units ($M=15.58$, $SD=59.23$) and the number of square feet of residential area ($M=15.38$, $SD=58.78$) conditions; $t(1999)=1.7257$, $p=0.0846$.

3.9. Disaggregation scale tests

Block data are more precise, but they are not as available as block group data. To determine if there is any difference between the scales of the census disaggregation unit, two paired-samples t-tests were conducted to test the same proxy unit at different scales. At both the block and block group levels, the tests were conducted between the number of residential units and the number of square feet of living area. This is testing final population estimation, not error rates.

Population: one parcel was chosen at random from each block group to avoid spatial autocorrelation issues. Each parcel has a population estimate disaggregated from block group and from block. These were compared to note the differences.

- There was extreme significant difference in the scores for number of residential units at the block group level ($M=1395.10$, $SD=937.04$) and at the block level ($M=153.93$, $SD=282.08$) conditions; $t(878)=39.9220$, $p<0.0001$.
- There was significant difference in the scores for number of square feet of living area at the block group level ($M=10.10$, $SD=74.59$) and at the block level ($M=5.27$, $SD=44.48$) conditions; $t(878)=2.8627$, $p=0.0043$.

4. Conclusions

For all proxy units at both disaggregation scales, most errors are small. Outliers are few but they exist. At the block group level, there is significant difference between the error results for the three proxy units (predictive factors). However, these errors do not transmit to the final population estimations, as there is no significant difference in the final population estimates between the three proxy units at the block group scale.

At the block level, results are inconclusive for analyzing the error rates of the two proxy units used as predictive factors. There is no significant difference between final population estimates produced by the two proxy units at the block level. There is significant difference between the resulting population estimates from the two census disaggregation units (block group and block). In conclusion, it is likely that final population estimates are not affected by choice of predictive factor. Error rates are affected, but error rates are not the final goal of the CEDS method. More work needs to be done to determine the better disaggregation scale.

4.1. Future Work

Previous research has shown the CEDS method to have high accuracy when compared to non-dasytetric methods. The CEDS method is relatively new compared to other more established dasymetric methods based on raster ancillary data, and as such, there is scope for further exploration and development. Our research has attempted to explore predictive units and census disaggregation scales in an attempt to refine the CEDS method. There are still areas for future development:

- Automate - creating data from CEDS is laborious and repetitive. An automated program or script should be developed to simplify the process.
- Include other socioeconomic factors - the CEDS method should be expanded to include factors other than total population (e.g. income, education, etc.) More established dasymetric methods have layers of other socioeconomic characteristics and the opportunity exists to explore how the CEDS method would incorporate these additional layers.
- Improve scale limitations - the resulting data prove useful at finer scale but are less useful with coarser resolutions.
- Explore pairing with other data – the resolution of these data suggests that they could complement other types of data. More research is needed to explore novel data combinations.

We feel the CEDS method offers great potential and encourage further development and refinement among researchers.

References

Eicher C, Brewer C (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science* 28:125-38

- Maantay J, Maroko A, Herrman C (2007) Mapping population distribution in the urban environment: the Cadastral-based Expert Dasymetric System (CEDS). *Cartography and Geographic Information Science* 34:77-102
- Maantay J, Maroko A (2009) Mapping urban risk: Flood hazards, race, environmental justice in New York. *Applied Geography* 29:111-124
- Social Explorer, U.S. Census Bureau; 2010 Census of Population and Housing, Summary File 1: Technical Documentation, Issued June 2011.
- Tobler, W., 1979. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association* 74: 519-536.