The Cadastral-based Expert Dasymetric System (CEDS) for Mapping Population Distribution and Vulnerability in New York City

Juliana Astrud Maantay [(1) (2) (3)]
Andrew R. Maroko [(1) (2)]

(1) Lehman College, City University of New York, Department of Environmental, Geographic, and Geological Sciences; and Geographic Information Sciences Program
(2) City University of New York Graduate Center, Earth and Environmental Sciences Ph.D. Program
(3) Lehman College, City University of New York, Department of Health Sciences, Public Health Graduate Program

Corresponding Author's E-Mail: juliana.maantay@lehman.cuny.edu

ABSTRACT:

We discuss the importance of determining an accurate depiction of total population and specific sub-population distribution for urban areas to develop an improved "denominator," enabling the calculation of more correct rates in GIS analyses involving public health, crime, hazard and risk assessment, and urban environmental planning. Rather than using data aggregated by arbitrary administrative boundaries such as census tracts, we use dasymetric mapping, an areal interpolation method using ancillary information to delineate areas of homogeneous values. Previous dasymetric mapping techniques (often using remotely-sensed land-cover data) are contrasted with our technique, Cadastral-based Expert Dasymetric System (CEDS), which is particularly suitable for hyper-heterogeneous urban areas. CEDS uses specific cadastral data, land-use filters, modeling by expert system routines, and validation against census enumeration units and other data. The CEDS method produces a more accurate estimation of population density and distribution, resulting in more robust analyses of environmental justice, health disparities, and hazard vulnerability.

KEYWORDS:

Dasymetric, Geographic Information Systems (GIS), cadastral, expert systems, population distribution, population mapping, environmental justice, health disparities, hazard vulnerability

Project Background and Purpose

It is very important to be able to accurately depict population distribution for urban areas in order to develop an improved "denominator," allowing for more correct rates in GIS analyses involving public health, crime, and urban environmental planning. Rather than using data aggregated by arbitrary administrative boundaries such as census tracts, accuracy is improved by the use of dasymetric mapping, an areal interpolation method using ancillary information to delineate areas of homogeneous values. Specifically, a new methodology called the **C**adastral-based **E**xpert **D**asymetric **S**ystem (CEDS) was designed and implemented in order to provide vital population data at the tax-lot level, a geographic unit roughly 350-times smaller than the census tract in New York City. This model is particularly suitable for urban areas, especially hyper-heterogeneous urban areas. CEDS uses specific cadastral data, land use filters, modeling by expert system routines, and validation against

various census enumeration units and other data. Previous and traditional disaggregation techniques were reviewed during the development of the CEDS method (Bhaduri et al, 2002; Bielecka, 2005; Bracken and Martin, 1989; Cai, et al, 2006; Eicher and Brewer, 2001; Flowerdew and Green, 1992; Goodchild, Anselin and Deichmann, 1993; Holt et al, 2004; Kyriakidis, 2004; Langford and Unwin, 1994; Liu, Clarke, Herold, 2006; Mennis, 2001; Moon, and Farmer, 2001; Reibel and Bufalino, 2005).

The CEDS method differs from these existing disaggregation methods in two major ways. Firstly, the ancillary data used is very detailed cadastral data, more appropriate to estimating population distribution in hyper-heterogeneous urban areas in a continuous (non-binary) way. Secondly, the CEDS method also uses an expert system to determine which of several formulae to use, calculating which method fits the data best. In this way, each source record within the area of interest can be customized as to method of disaggregation, which when validated, yield more accurate results.

CEDS Methodology and Analysis

Our method of using cadastral data as the ancillary data appears to be an innovative and progressive approach to dasymetric mapping. Cadastral data is used in recording property boundaries, property ownership, property valuation, and of course, for the all-important purpose of property tax collection. The type of cadastral level data used in our CEDS method is commonly available for most urbanized areas in the United States, western Europe, and other more-developed regions. The data is usually organized by township, municipality, or county, and, less often, by metropolitan region.

However, in many parts of the world, census and cadastral data may not be readily available, current, or accurate. Baudot makes the point that for urban areas in less-developed countries, very often there are no census, property tax records, or city planning data on population to work with, and even when such data are collected, the exponential growth rates of these cities makes the census data obsolete almost immediately. This is why satellite data are most often used for dasymetric mapping – they are available for almost all parts of the world, and are very current. However, "urban environments are often considered too complex to be analyzed by satellite remote sensing, and indeed, the spatial resolution of current satellite sensors means that they are not particularly well-suited to the task," (Baudot, 2001: 266). In urban areas where census and cadastral data are available, the CEDS method will be an improvement. For instance, municipalities where property tax records are linked to a digital spatial database (e.g., most larger cities and towns in the United States and more

developed nations) the cadastral data required by the CEDS method will be available. Although this data may not be available to the general public for free, it still tends to be less expensive for the end user than high-resolution remotely sensed images for the equivalent spatial extent.

The following diagrams illustrate how the CEDS method of dasymetric mapping can be beneficial to health, environmental, crime, risk assessment, hazard and emergency planning, and other urban planning analyses. The diagrams in Figure 1 contrast standard areal weighting interpolation and filtered areal weighting dasymetric (binary) techniques with the cadastral expert dasymetric system (CEDS) method.

The CEDS method differs from most other forms of dasymetric mapping because it does not use areal weighting or the binary (filtered areal weighting or "punch-out") method alone. The ancillary data used is not remotely-sensed land cover/land use, interpreted to estimated population density classes, but rather very detailed cadastral data, more appropriate to estimating population distribution in hyper-heterogeneous urban areas. The CEDS method also uses an expert system to determine which of several formulae to use, calculating which method fits the data best. In this way, each source record within the area of interest can be customized as to method of disaggregation, which when validated, yield results that best fit the data.

Using the CEDS method, the modeled population data always preserves the pycnophylactic property, meaning that the estimated (modeled) value of the tract, when re-aggregated, must equal the original value of the tract (Tobler, 1979). Preservation of the pycnophylactic property is not always achievable with previously used dasymetric methods based on population density classes derived from land use/land cover data.

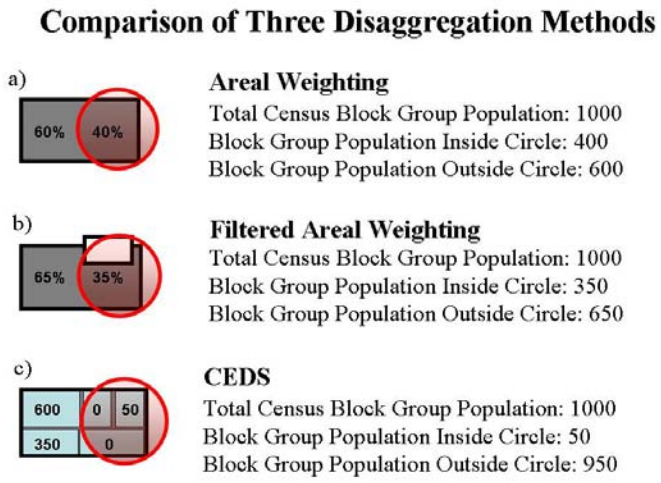


Figure 1: Methodological differences and potential improvement of population estimation of the CEDS method (c), over both Filtered Areal Weighting (b), and Simple Areal Weighting (a).

This study was designed to disaggregate the total population counts from the census block group level (5,733 in NYC) to the tax lot level (847,153 in NYC) using CEDS. Census block groups, rather than the smaller census blocks, were used due to data suppression of subpopulations in the latter.

The CEDS technique uses residential area (*RA*) and number of residential units (*RU*) as proxies for population distribution. In other words, it is assumed that where there are more potential living accommodations, there will be higher populations. As such, the population in each block group was disaggregated, or redistributed, among the tax lots based on either *RA* or *RU*. The proxy unit (*RA* or *RU*) used in the disaggregation was determined by an expert system individually for each geographic unit. The results were then validated against census data and compared to commonly used dasymetric techniques to assess predictive accuracy and possible improvement over other methods.

The CEDS disaggregation of census populations can be compartmentalized into three fundamental steps: 1) data preparation, 2) dasymetric calculations, and 3) expert system implementation. The discussion of these steps is followed by an evaluation of the results.

Data Preparation

Two datasets were used for this process: the 2000 census data (U.S. Census Bureau, 2000) and LotInfo (LotInfo, LLC, 2001). Decennial census data (2000) for New York City was downloaded via www.census.gov. LotInfo, a product of LotInfo, LLC, which combines spatial data from the New York City Department of City Planning (DCP) and attribute information from the Real Property Attribute Data (RPAD) database provided by the New York City Department of Finance (DOF), contains exhaustive data at the tax lot level in NYC (e.g. zoning, ownership, building attributes, residential area, and residential units). Although this study was done in NYC, similar data are often available from planning departments of metropolitan areas or urbanized counties.

Residential area (*RA*) and number of residential units (*RU*) are important attributes in the CEDS process. Within the lot-level data the RU variable did not require additional processing, however there are many instances of missing data values for the RA variable in the original RPAD data from the Department of Finance. As such, a new variable, adjusted residential area (*ARA*), was created. *ARA* is identical to *RA* in many cases, however when the value for *RA* is zero and the number of residential units (*RU*) does not equal zero (i.e. there are residential units but no value for residential area), *ARA* is defined as the total building

area multiplied by the ratio of the number of residential units and the total number of units. It can be written as follows:

(Equation 1)

$$ARA = M * (BA * RU / TU) + RA$$
$$\textbf{IF } RA = 0 \textbf{ AND } RU <> 0, \textbf{ THEN } M = 1, \textbf{ ELSE } M = 0$$

Where:
ARA = Adjusted Residential Area within tax lot
BA = Building Area (residential and commercial) within tax lot
RU = Number of Residential Units within tax lot
TU = Total Number of Units (residential and commercial) within tax lot
RA = Residential Area
M = Binary variable designating ancillary data for ARA

Dasymetric Calculations

Using the GIS capabilities of ARCGIS 9.1 (ESRI, 2005) and the LotInfo dataset, the total amounts of *RU* and *ARA* were calculated for each census tract and census block group in NYC and saved in tabular form. In other words, the *RU* and *ARA* information, at the tax lot-level, was aggregated up to the block group and tract levels. This table was then used to generate a tax lot-level spatial data layer with *RU* and *ARA* values aggregated at the tax lot, block group, and tract levels, as well as the census population data at the block group and tract levels.

Several dasymetrically derived populations were then calculated. The general equation is solved by multiplying the census population with the ratio of population proxy units and can be written as such:

(Equation 2)

$$POP_l = POP_c * U_l / U_c$$

Where:
POP_l = dasymetrically derived lot-level population
POP_c = census population (block group or tract)
U_l = the number of proxy units at the tax lot level (*RU* or *ARA*)
U_c = the number of proxy units at the census level (*RU* or *ARA* per block group or tract)

Values were calculated from the block group and tract census populations using both *RU* and *ARA* as the proxy units. The process resulted in four dasymetrically derived population values for each tax lot (tract *ARA*, tract *RU*, block group *ARA*, and block group *RU*).

Expert System Implementation

The expert system was designed to determine which proxy unit, number of residential units (*RU*) or adjusted residential area (*ARA*), more accurately predict the population

distribution on a tract by tract basis. This was accomplished by re-aggregating the tax lot level population figures that were derived from the census tract data back to the block group level, resulting in an estimated block group population. In other words, tract data were disaggregated down to the tax lot and then re-aggregated up to the block group. It was necessary to use the tract-level data as a starting point so that there would be a smaller unit of aggregation (block group) available in the census data with which to compare the estimated values. Although the census data are available by census block, a unit smaller than the block group, much of the data is suppressed due to small numbers and privacy issues, particularly when dealing with sub-populations, which is one of the future goals of this study.

The absolute value of the difference between census populations and estimated populations can be written as follows:

(Equation 3)

$$POP_{diff} = | POP_{BG} - POP_{est} |$$

Where:
POP_{diff} = the difference between census and estimated populations per block group
POP_{BG} = census block group population
POP_{est} = estimated population (*RU* or *ARA*) derived from the census tract (not block group)

By comparing the estimated population to the census population for both the *RU*- and *ARA*-based techniques, it can be assumed that the process which resulted in estimates more similar to the census block group values (i.e. smaller POP_{diff} values) more accurately redistributed the data. After rejoining the POP_{diff} data with the LotInfo data, the expert system would then select the superior proxy unit as the disaggregation technique for each block group. It can be described as follows:

(Equation 4)

$$\textbf{IF } RU\_POP_{diff} <= ARA\_POP_{diff}, \textbf{ THEN } POP_l = POP_{RU\_BG}, \textbf{ ELSE } POP_l = POP_{ARA\_BG}$$

Where:
RU_POP_{diff} = the absolute difference between the census block group population and the estimated block group population derived from the census tract population based upon number of residential units
ARA_POP_{diff} = the absolute difference between the census block group population and the estimated block group population derived from the census tract population based upon residential area
POP_l = the final estimated tax lot population dasymetrically derived from the census block group population (not the census tract)
POP_{RU_BG} = the estimated tax lot population dasymetrically derived from the census block group population (not the census tract) based on number of residential units
POP_{ARA_BG} = the estimated tax lot population dasymetrically derived from the census block group population (not the census tract) based on the adjusted residential area

In essence, it is the performance of the tract-level disaggregation that defines the proxy units used for each block group disaggregation, ultimately resulting in a final dasymetrically derived value, individually tailored for each block group.

Comparison with Filtered Areal Weighting

The filtered areal weighting (binary) method was used in order to compare the accuracy of CEDS against a commonly used disaggregation technique, essentially acting as a control variable. The filtered areal weighting methodology is comparatively simple, using a combination of 'cookie cutter' overlay and areal weighting processes.

Census tract, census block group, TIGER landmark, and TIGER water body geographic files were downloaded from the U.S. Census Bureau's website. The landmark and water body data layers were then combined and processed to make an 'open spaces' layer where there is known to be no residential population (e.g. parks, airports, cemeteries, water bodies, golf courses, and national recreation areas). The open spaces layer acted as a 'cookie cutter' on the tract and block group boundaries, resulting in the tracts and block groups being geographically modified to exclude the open space regions.

Area of the census polygons (as calculated within ArcGIS 9.1) and total population (from census SF1, table P001) attribute data were added to the tract and block group boundary layers. Areal weighting was then utilized to complete the filtered areal weighting process by equating the estimated block group population to the census tract population multiplied by the ratio of block group area and tract area, as modified by the binary filtering. It is important to note that this weighting technique makes the assumption that the population is uniformly distributed within each census tract rather than using additional ancillary data to redistribute the population in a heterogeneous manner. It can be written as follows:

(Equation 5)

$$POP_{FAW} = POP_{TR} * AREA_{BG} / AREA_{TR}$$

Where:
POP_{FAW} = Estimated block group population from filtered areal weighting
POP_{TR} = Census tract population
$AREA_{BG}$ = Modified census block group area (open spaces excluded)
$AREA_{TR}$ = Modified census tract area (open spaces excluded)

Comparison of the Four Methods: CEDS, Filtered Areal Weighting, Dasymetrically-derived populations using both ARA and RU independently

In order to assess the accuracy and validity of the dasymetrically derived populations (as obtained by filtered areal weighting, *ARA* alone, *RU* alone, and CEDS), the results were compared to census block group populations. This can be seen very simply by comparing the estimated block group populations to the census block group populations. The absolute values for the difference between each block group population were summed, divided by the entire population on NYC, and converted to a percentage (see figure 3). This very simple

analysis suggests that CEDS, with only 6.37% difference, outperformed RU (8.69%), ARA (9.44%), and filtered areal weighting (21.91%).

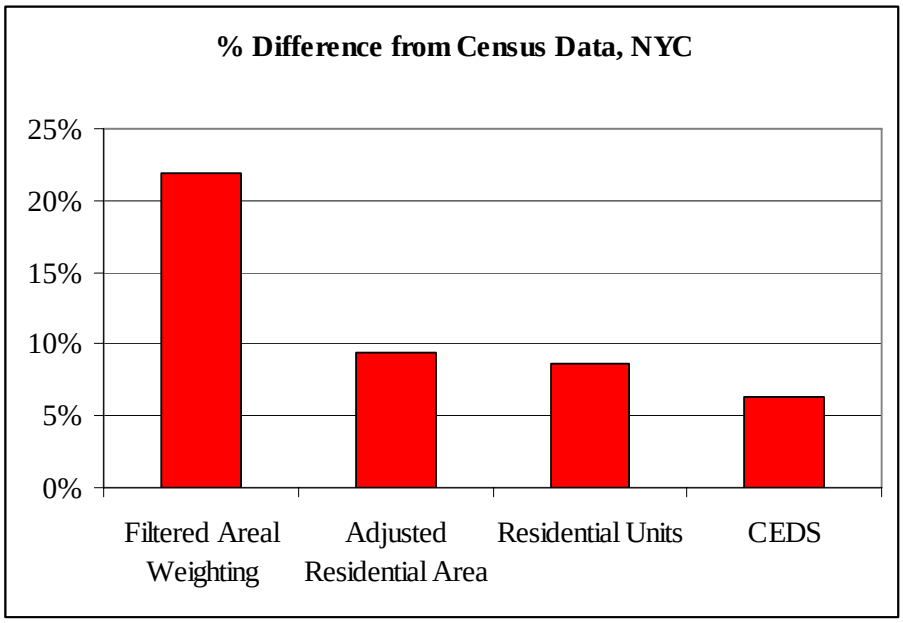
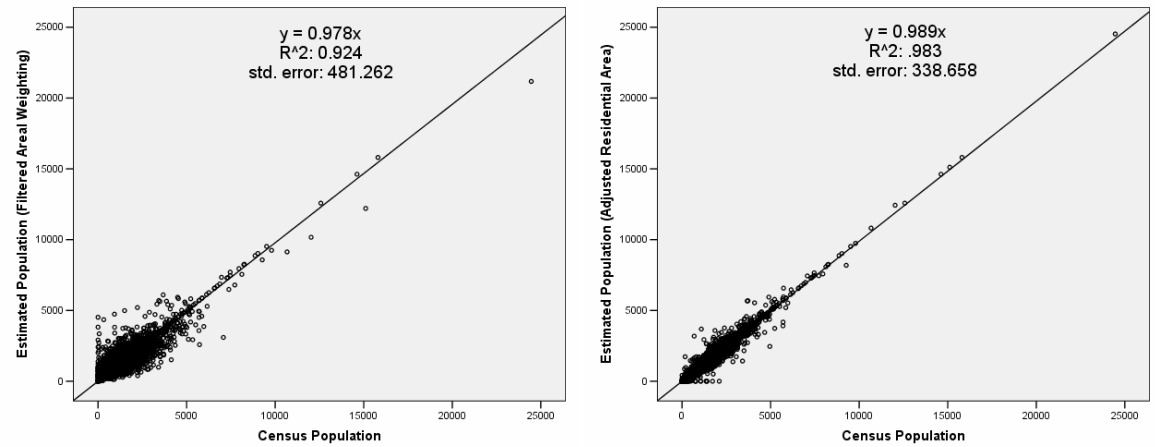


Figure 2: Percent absolute difference between census block group population and estimated block group populations in New York City for the different methods.

For a more comprehensive analysis, linear regressions, similar to Qiang Cai's approach in "Age-sex population estimation for small areas" (Cai, et al, 2006) except using all block groups rather than selected block group pairs, were performed. The estimated block group populations from the four disaggregation methods were regressed against the block group population data from the census to evaluate their relative effectiveness in New York City as a whole and separated by borough. This analysis involved linear regression with the regression line forced through the origin. The R^2, standard errors, and regression coefficients were then compared and are summarized in figure 3.

a) b)



c) d)

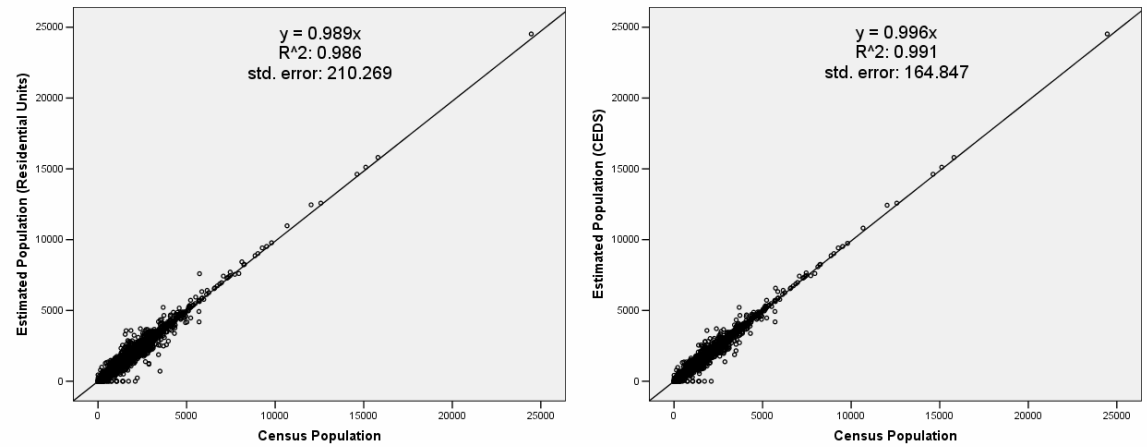


Figure 3: Simple linear regressions for NYC showing R^2, standard errors, and regression coefficients of block group populations estimated by filtered areal weighting (a), ARA (b), RU (c), and CEDS (d) versus census block group populations.

The regression coefficients for all of the methodologies were approximately '1' as would be hoped for, with the CEDS method producing the closest value (.996) and the filtered areal weighting producing the most dissimilar (.978). As can be seen by examining the differences in R^2 values, the expert system produced more highly correlated results (R^2 = .991) than by using the *ARA* (.983), *RU* (.986), or filtered areal weighting (.924). The standard errors also imply that the CEDS methodology (std. error = 164) outperformed the other three (std. error = 481, 339, and 210 for filtered areal weighting, ARA, and RU, respectively). That CEDS produced better results than *ARA* or *RU* is not unexpected since CEDS selects the better performing proxy unit on a tract by tract basis. What is more substantive is the contrast between the filtered areal weighting method, serving more or less as a control, and the expert dasymetric system. This is seen most intuitively by examining the wider spread of data points in the filtered areal weighting scatterplot (Figure 3(a)) as compared to the CEDS scatterplot (Figure 3(d)). When regression analyses were performed on a borough by borough basis, the results were similar although some spatial variation can be seen (see figures 4 and 5).

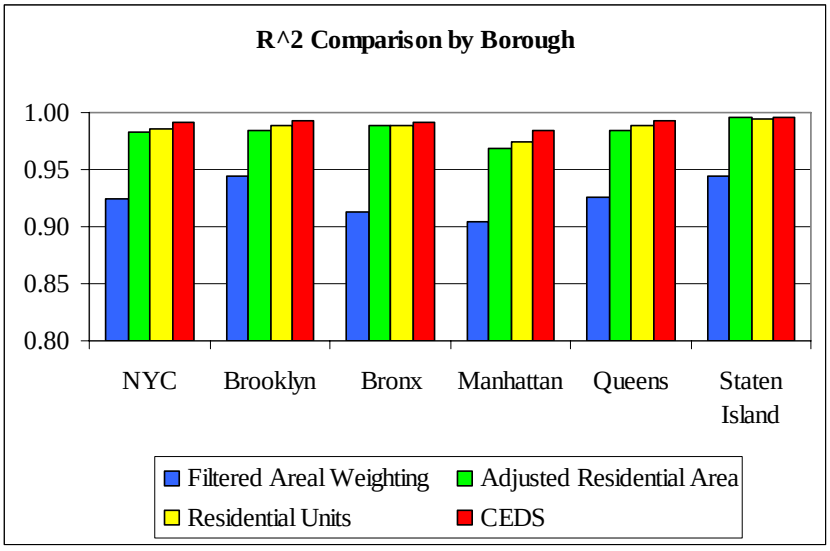


Figure 4: R^2 for linear regressions of block group populations estimated by each of the four disaggregation methods versus census block group populations.

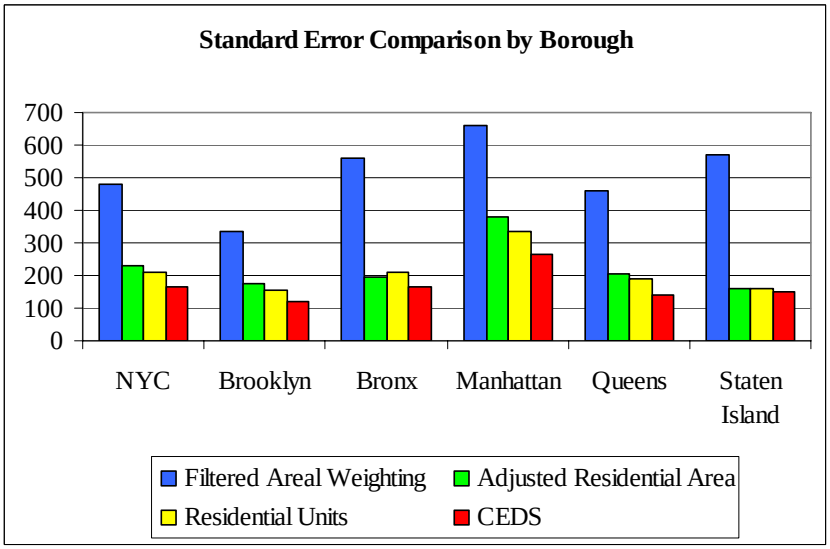


Figure 5: Standard errors for linear regressions of block group populations estimated by each of the four disaggregation methods versus census block group populations.

Even though filtered areal weighting resulted in acceptable R^2, standard error, and parameter estimates for these densely settled urban areas, the dasymetric technique used in this study is clearly superior. It is also important to note that what is being compared in this analysis section is not the end-product of the dasymetric process, but rather a validation of its efficacy at a comparatively coarse spatial aggregation. The result of the CEDS methodology is tax lot-level rather than block group-level population data, an areal unit that has approximately 150-times finer resolution. See Figure 6 for a comparison of CEDS-derived

population, CEDS-derived population density by tax lot, and traditional choroplethic population density by census block group.

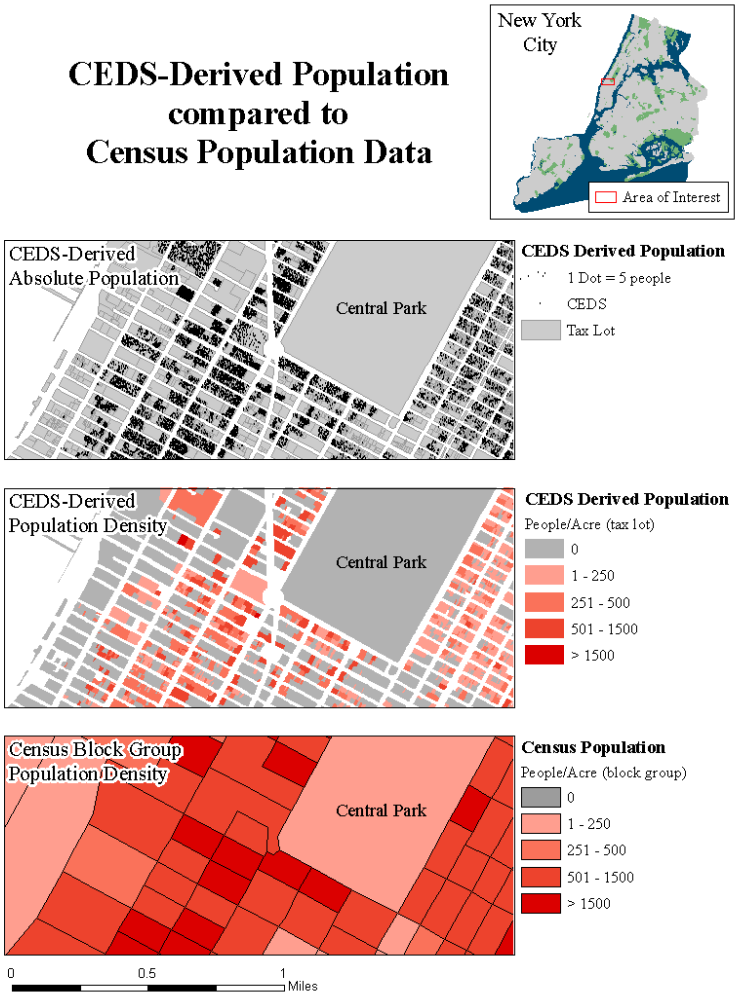


Figure 6: Visual comparison of CEDS-derived population, CEDS-derived population density by tax lot, and traditional choroplethic population density by census block group.

Dasymetric Mapping - Where do we go from here?

Based on the application of the CEDS methodology to New York City population data, we have demonstrated that the Cadastral-based Expert Dasymetric System can improve research and analyses that utilize population distribution information, and create more realistic models of real-world conditions (Maantay et al, 2007, Maantay et al, 2008). We have established the usefulness of the CEDS method for any analyses employing population-based rates, as is commonly the case with public health and epidemiological research, crime mapping, and risk assessment (see Figure 7), but the CEDS method is not limited to improving the development of rates alone. These methods will be useful in many disparate fields and serve many purposes. For instance, one can improve emergency management operations and implementation by providing more precise information about actual positions

of susceptible populations, thereby increasing the quality of functions such as evacuation route planning, optimal site selection for emergency shelter locations, and critical rescue and recovery prioritization for first responders (Maantay and Maroko, 2008). Obviously, this can be extended to police operations, criminal justice, fire and ambulance services, utility providers, and any other crucial public support systems dependant upon population information. (See Figure 8.)

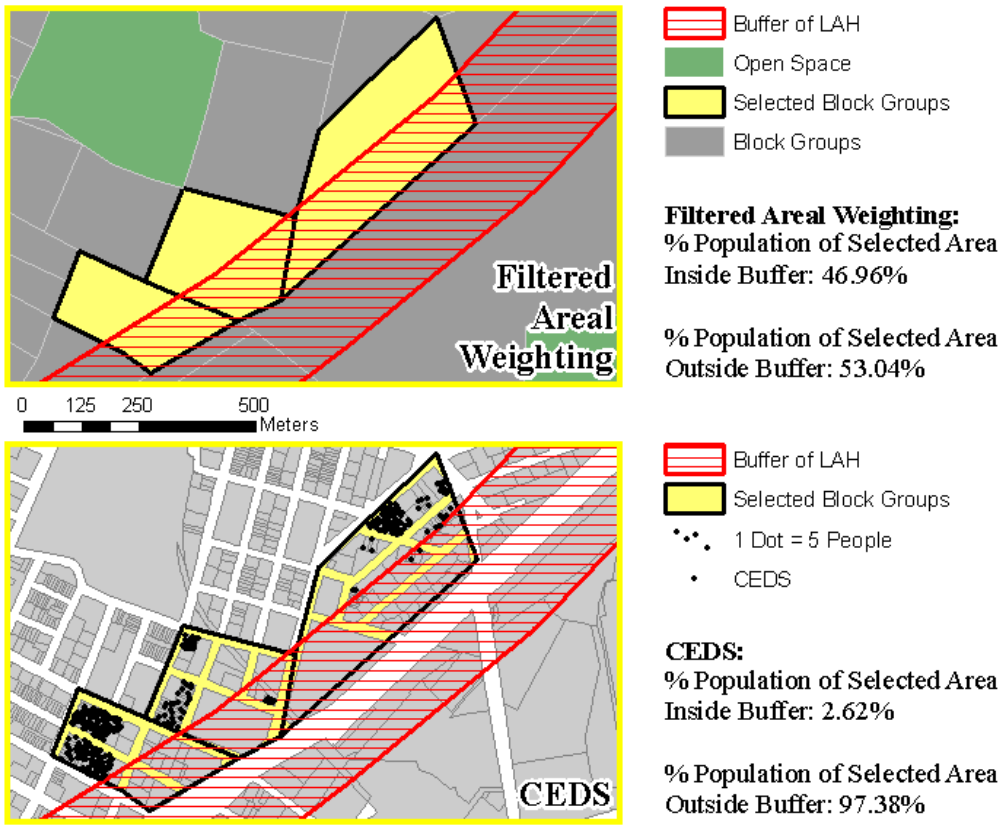


Figure 7: An example showing the differences in impacted population estimates using the conventional filtered areal weighting method vs. CEDS. In this case, the buffered area represents the distance pollutants would likely travel from a major limited access highway (LAH), and the lower population within the buffer as estimated by CEDS would result in a lower denominator, and therefore higher rates when calculating respiratory disease rates.

Figure 8: Comparison of three methods for estimating population at risk from floods. (a) The Centroid Containment method counts only the population in those census units whose centroids fall within the flood zone. The Areal Weighting method counts the population proportional to the portion of the census unit area included in the flood zone. (c) CEDS calculates the population by tax lot, creating a more realistic estimation of population distribution and vulnerable populations.

Additionally, the knowledge of accurate population distribution can be extremely valuable in the sphere of urban planning. The understanding of the locational characteristics of target populations would allow for more equitable resource allocation in areas such as community infrastructure development, provision of open space and recreational opportunities, transportation access, and necessary environmental facilities.

As the morphology of cities becomes increasingly complex, the need continues to grow for immediate and well-informed decision-making, regarding both catastrophic and everyday events. We anticipate that advances in dasymetric mapping, such as the CEDS method, will help us to "perfect the denominator" and better our understanding of the human-urban project.

Selected References

Baudot, Y., 2001. A Method for the Geographical Analysis of the Population of Fast-growing cities in the Third World. In: Donnay, J.-P., Barnsley, M., and Longley, P. (eds.) *Remote Sensing and Urban Analysis.* Taylor and Francis, London, UK, pages 249-268.

Bhaduri, B., Bright, E., Coleman, P., and Dobson, J., 2002. LandScan: Locating People is What Matters. *GeoInformatics,* April/May: 34-35, 37.

Bielecka, E., 2005. A Dasymetric Population Density Map of Poland. In *Proceedings of the 22nd International Cartographic Conference*, July 9-15, A Coruna, Spain.

Bracken, I., and Martin, D., 1989. The Generation of Spatial Population Distributions from Census Centroid Data. *Environmental and Planning A*, 21:537-543.

Cai, Q., Bhaduri, B., Coleman, P., Rushton, G., Bright, E., 2006. Estimating Small-Area Populations by Age and Sex Using Spatial Interpolation and Statistical Inference Methods. *Transactions in GIS* 10(4):577-598.

Eicher, C., and Brewer, C., 2001. Dasymetric mapping and areal interpolation: implementation and evaluation. *Cartography and Geographic Information Science* 28:125-138.

Environmental Systems Research Institute (ESRI), 2005. *ArcGIS 9.1*. Redlands, CA

Flowerdew, R., and Green, M., 1992. Developments in areal interpolation methods and GIS. *Annals of Regional Science* 26: 67-78.

Goodchild, M., Anselin, L., and Deichmann, U., 1993. A framework for the areal interpolation of socioeconomic data. *Environment and Planning A 25*: 383-397.

Holt, J. B., Lo, C.P., and Hodler, T. W., 2004. Dasymetric Estimation of Population Density and Areal Interpolation of Census Data. *Cartography and Geographic Information Science 31*:103-121.

Kyriakidis, P. 2004. A Geostatistical Framework for Area to Point Spatial Interpolation. *Geographical Analysis 36*(3):259-289.

Langford, M., and Unwin, D., 1994. Generating and mapping population density surfaces within a geographical information system. *Cartographic Journal 31*: 21-26.

Liu, X., Clarke, K., Herold, M., 2006. Population Density and Image Texture: A Comparison Study. *Photogrammetric Engineering and Remote Sensing, 72*(2):187-196.

LotInfo, LLC, 2001. LotInfo. SpaceTrack, Inc. 304 Park Ave, 11th Floor New York, NY 10010

Maantay, J.A., Maroko, A., and Herrmann, C., 2007. Mapping Population Distribution in the Urban Environment: The Cadastral-based Expert Dasymetric System (CEDS), *Cartography and Geographic Information Science*, special issue: *Cartography 2007: Reflections, Status, and Prediction.* Vol. 34, No. 2: 77-102.

Maantay, J.A., Maroko, A.R., Porter-Morgan, H., 2008. A New Method for Mapping Population and Understanding the Spatial Dynamics of Disease in Urban Areas, *Urban Geography,* in press.

Maantay, J.A., and Maroko, A.R., 2008. Mapping Urban Risk: Flood Hazards, Race, & Environmental Justice in New York, *Applied Geography*, in press.

Mennis, J., 2001. Generating Surface Models of Population Using Dasymetric Mapping. *The Professional Geographer, 55*(1):31-42.

Moon, Z.K., and Farmer, F.L., 2001. Population Density Surface: A New Approach to an Old Problem. *Society and Natural Resources*, 14: 39-49.

Reibel, M., and Bufalino, M.E. 2005. Street-Weighted Interpolation Techniques for Demographic Count Estimation in Incompatible Zone Systems. *Environmental and Planning A*, 37: 127-139.

Tobler, W., 1979. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association* 74: 519-536.

United States Bureau of the Census, 2001(a), *Census 2000 Summary File 1, New York State*.

United States Bureau of the Census, 2001(b). *TIGER Files*. Geography Division, Cartographic Products Management Branch.