Incorporating GIS Building Data and Census Housing Statistics for Sub-Block-Level Population Estimation

Shuo-Sheng Wu
Texas State University—San Marcos and U.S. Geological Survey

Le Wang
University of Buffalo, State University of New York

Xiaomin Qiu
Missouri State University

This article presents a deterministic model for sub-block-level population estimation based on the total building volumes derived from geographic information system (GIS) building data and three census block-level housing statistics. To assess the model, we generated artificial blocks by aggregating census block areas and calculating the respective housing statistics. We then applied the model to estimate populations for sub-artificial-block areas and assessed the estimates with census populations of the areas. Our analyses indicate that the average percent error of population estimation for sub-artificial-block areas is comparable to those for sub-census-block areas of the same size relative to associated blocks. The smaller the sub-block-level areas, the higher the population estimation errors. For example, the average percent error for residential areas is approximately 0.11 percent for 100 percent block areas and 35 percent for 5 percent block areas. **Key Words: block population, dasymetric mapping, population estimation, population interpolation, sub-block.**

这篇文章提出了一个确定性模型以作为人口分板块级估算。该研究的数据来自地理信息系统(GIS)和三个板块级人口房屋普查。为了评估这个模型，我们总计了普查里的板块区并计算各自住房统计来制造几个人造板块。接着，我们运用模型来估计人造板块区的人口并以这些地区的现实对照的人口普查来评估概算。我们的分析指出人造分板块区的人口估计平均百分点误差能媲美那些相同大小的分普查板块区。分板块级的区越小，人口估计的误差越高。举例来说，住宅区百分之一百的板块平均百分点误差是大约0.11%，而百分之五的板块只有35%的误差。
关键词：人口板块，分区密度制图，人口估算，人口插值，分板块。

En este artículo se presenta un modelo determinista para el cálculo de la población a nivel de sub-bloque con base en el volumen total de edificios derivado de datos sobre edificios obtenidos con el sistema de información geográfica (geographic information system, GIS) y tres estadísticas sobre vivienda a nivel de bloque del censo. Para evaluar el modelo, generamos bloques artificiales agregando áreas de bloques del censo y calculando las estadísticas de vivienda respectivas. Entonces aplicamos el modelo para calcular las poblaciones de las áreas de los sub-bloques artificiales y evaluamos los cálculos con las poblaciones de las áreas del censo. Nuestro análisis indica que el porcentaje medio de error del cálculo de la población en áreas de sub-bloques artificiales es comparable con el de las áreas de sub-bloques del censo del mismo tamaño en relación con los bloques asociados. Cuanto más pequeñas sean las áreas a nivel de sub-bloque, más grandes serán los errores en el cálculo de la población. Por ejemplo, el porcentaje medio de error en áreas residenciales es de aproximadamente 0.11 por ciento en áreas con bloques de 100 por ciento, y un 35 por ciento en áreas con bloques de cinco por ciento. **Palabras clave: población en bloques, mapeo dasimétrico, cálculo de la población, interpolación de la población, sub-bloque.**

The Professional Geographer, 60(1) 2008, pages 121–135
Initial submission, July 2006; revised submissions, November 2006 and April 2007; final acceptance, July 2007.
Published by Taylor & Francis, LLC.

In the United States, the most fine-grained census population data available to the public is at the block level. According to the U.S. Census Bureau (2003), "census blocks are areas bounded on all sides by visible features, such as streets, roads, streams, and railroad tracks, and by invisible boundaries, such as city, town, township, and county limits, property lines, and short, imaginary extensions of streets and roads." In fact, the sizes of census blocks vary greatly. Using the city of Austin, Texas, as an example, the boundary on one side of a census block can range from fifty meters in downtown to 10 km in the suburb; the population of a census block can range from zero to 3,300.

For various purposes, people might need to estimate population for areas not coinciding with census block boundaries or for areas smaller than a census block. For example, floodplains usually do not share the same boundaries as those of census blocks, yet local governments might need to estimate floodplain populations for flood hazard planning. Redevelopment subdivisions might not have the same boundaries as those drawn for census blocks, yet city planners and developers might need to estimate the number of local residents for planning and resource management purposes. Transportation engineers might need to estimate populations within a half-mile buffer of a proposed railroad or highway to assess the potential impact. Demographers might need to estimate sub-block-level population in suburban areas where census blocks are usually large for studying urban sprawl.

This article presents a deterministic model for sub-block-level population estimation. The model can estimate population for an arbitrary area based on total building volumes within the area and three housing statistics for the area. Using building volumes derived from geographic information system (GIS) building data and housing statistics derived from the U.S. Census 2000 block-level data for our case study area in Austin, Texas, we applied a simulation approach to assess the model for sub-block-level population estimation. The simulation approach generates artificial blocks by aggregating census block areas, calculates housing statistics for the artificial blocks, estimates populations for sub-artificial-block areas based on the deterministic model, and compares the estimates with census populations of the areas. Using the simulation approach, we also assessed whether the average percent error of population estimation for sub-block-level areas of the same size relative to associated blocks varies with the block size, so that we can apply the error statistics based on artificial blocks to those based on census blocks. Furthermore, we assessed how the rescaling for block population preservation improves sub-block-level estimates, and how incorporating land use information affects the estimation accuracy. The results of this study have implications for researchers and practitioners who need population estimates at finer spatial resolution than the census block as well as accuracy assessment for the estimates.

Review of Past Population Estimation Studies

Past population estimation studies can be grouped into three categories depending on the data required for input in the estimation. The first category estimates areal population from census zone-based population data using certain mathematical interpolation functions. The second category infers population from population-relevant physical or socioeconomic variables. The third category disaggregates census unit populations into zones that are partitioned by population-relevant variables. The first category of studies requires only census population data as the input. The second category of studies requires only population-relevant variables as the input. The last category of studies requires both census population data and population-relevant variables as the input, and therefore the resulting estimates are generally more reliable.

Mathematical interpolation of census population can be divided into point-based approaches and area-based approaches (Lam 1983). In point-based approaches, a control point is assigned to represent each zone-based census unit and its associated population. Then, using a certain mathematical function of interpolation, a grid map of population is generated with grid point values estimated from the control points (e.g., Martin 1989, 1996; Bracken 1991). In contrast, area-based approaches directly use census zones as the unit of operation and transform the original

zone-based population data to a representation of fine grids based on certain mathematical functions (e.g., Tobler 1979; Rase 2001).

Population counts can be inferred from related variables. Depending on the scale of estimation, past studies have estimated populations from urban areas (Tobler 1969; Lo and Welch 1977; Prosperie and Eyton 2000), land use areas (Kraus, Senger, and Ryerson 1974; Weber 1994; Lo 2003), dwelling unit counts (Hsu 1971; Lo and Chan 1980; Lo 1989), image pixel statistics (Webster 1996; Harvey 2002a; Liu, Clarke, and Herold 2006), and other relevant physical or socioeconomic variables (Green and Monier 1959; Dobson et al. 2000; Liu and Clarke 2002).

Census unit populations can be disaggregated into homogeneous zones delineated by spatial variables that are related to population distribution. This approach can be referred to as the dasymetric mapping method (Robinson et al. 1995). The most commonly used variable for census population disaggregation is land use and land cover data (e.g., Yuan, Smith, and Limp 1997; Mennis 2003; Holt, Lo, and Hodler 2004). Other variables that have been used include topography (Wright 1936), election district demographic statistics (e.g., Flowerdew and Green 1989, 1991), road networks (e.g., Xie 1995; Hawley and Moellering 2005; Reibel and Bufalino 2005), remote sensing image spectral and textural statistics (Harvey 2002b; Liu, Clarke, and Herold 2006; Wu, Qiu, and Wang 2006), and other relevant physical or socioeconomic variables (e.g., Dobson et al. 2000; Liu and Clarke 2002).

To disaggregate census unit population based on relevant variables, researchers need to establish a mathematical relationship between population counts and the variables. For example, when disaggregating census population based on the land use variable, researchers have to determine the population density for each land use class or the population density ratio between land use classes before redistributing census unit population to different land use zones. The mathematical relationship between population counts and relevant variables can be established from sampling (e.g., Mennis 2003), from regression analysis (e.g., Yuan, Smith, and Limp 1997; Wu, Qiu, and Wang 2006), or based on domain knowledge of researchers (e.g., Eicher and Brewer 2001).

Methods for Population Estimation and Assessment

This study presents a model to estimate population for small, sub-block areas based on the building volume variable derived from building footprint GIS data. The model can be applied under the context of the second and third categories of population estimation reviewed previously. We evaluated both the original model estimates and the rescaled model estimates (for census population preservation) that correspond to the two categories of population estimation, respectively. Compared to other population-relevant variables, building volumes have a straightforward and meaningful relationship with population counts, and building footprints provide a direct and accurate representation of where people are. Furthermore, to connect building volumes to population counts, we incorporated three housing statistics that are available from the census data and present a deterministic model:

$$\text{Pop} = \text{BdV}/\text{HuSpace}^{*}\text{OccRate}^{*}\text{HdSize}, \quad (1)$$

where Pop = population (the number of people), BdV = building volumes (e.g., cubic feet), HuSpace = average space per housing unit (e.g., average cubic feet per housing unit), OccRate = housing unit occupancy rate (percentage), and HdSize = average household size (average number of persons per household). Equation (1) states that when the total building volume within an area is divided by the average space per housing unit of the area, the derived figure is the total number of housing units within the area. Then, when the total number of housing units is multiplied by the occupancy rate of the area, the derived figure is the total number of households within the area. Further, when the total number of households is multiplied by the average household size of the area, the derived figure is the total population within the area.

The deterministic model by nature is capable of estimating population for an arbitrary area based on the total building volume and housing statistics of the area. To estimate population for sub-block areas, we used housing statistics at the block level so that the relationship between building volumes and population counts can be locally specified for each census block. For

block-level population estimation, the model should provide a very close estimate of the actual block populations. However, when the model is applied to estimate sub-block-level population, there are likely higher errors due to the heterogeneous housing statistics within blocks.

To assess the deterministic model, we first used the model to estimate populations for 720 residential census blocks and assessed the estimates as a benchmark of accuracy. We then adopted a simulation approach to infer how accurately the model would estimate population for sub-census-block areas. The simulation approach first generated artificial blocks by combining multiple census-block areas and deriving their respective housing statistics. Figure 1 illustrates this approach, in which every twenty block areas are aggregated as artificial blocks. The average household size for artificial blocks was calculated by dividing the total populations within the artificial block by the total number of households. The household occupancy rate was calculated by dividing the total number of households within the artificial block by the total number of housing units. The average space per housing unit was calculated in a similar fashion so that the housing statistic is the actual census figure.

After artificial blocks were generated, the simulation approach estimated population for sub-artificial-block areas based on the deterministic model. Then the estimates were compared with census population of the areas for

Figure 1 *Artificial blocks by every twenty-census-block aggregation.*

accuracy assessment. It is worth noting that census-defined blocks and the artificially generated blocks are in essence the same entity with the same attributes of respective block-level housing statistics and the perceived meaning of homogeneity when used as a basis to estimate their subarea populations.

To assess how the accuracy of sub-artificial-block estimates varies with the artificial block size, and also to make a connection of the accuracy assessment from artificial blocks to census blocks for extrapolating the trends of accuracy and error statistics, we compared the accuracy statistics for sub-artificial-block areas of the same size relative to artificial blocks (e.g., 50 percent artificial-block areas) based on different artificial-block sizes. Furthermore, we assessed how the rescaling of sub-(artificial-)block estimates for preserving (artificial-)block populations affects the estimation accuracy by applying a single scaling factor to all sub-(artificial-)block estimates within the same (artificial) block. To test if available land use information can improve sub-block-level population estimation, we assessed sub-block estimates for mixed-land-use blocks and compared them with those for residential blocks.

Compared to past population estimation studies, this study is unique in three aspects. First, we inferred areal population from building footprints and associated building volumes. A spatial unit of building area is a fine spatial unit that allows demarcation of population concentration areas at the sub-block level. A building is also an integral unit for counting population because people live and work in buildings. In addition, for the same type of buildings or buildings of the same land use, population counts are proportional to building volumes. Although the commonly used spectral and textural statistics from high resolution (e.g., 1-m resolution) remote sensing images offer an opportunity to infer population counts at a fine spatial scale, relevant studies (Liu, Clarke, and Herold 2006; Wu, Qiu, and Wang 2006) have not shown results as satisfactory as this study.

Second, this study incorporated census housing statistics to establish a mathematical relationship between population counts and the building volume variable. In contrast to sampling, regression analysis, or relying on domain knowledge to establish a mathematical relationship between population counts and population-relevant variables (reviewed in the previous section), incorporating census-block-level statistics allows us to specify the relationship between population counts and building volumes locally by each census block instead of globally across the entire data set. In addition, census data are readily available to the public and are deemed quite reliable.

Finally, past population estimation studies usually build population estimation models from higher level census population data (which are aggregated from lower level census data) and then assess the models from lower level census data. For example, population estimation models might be built from census-tract-level data, then census-block-group-level data are used to verify the models (Eicher and Brewer 2001; Lo 2003). In this study, our population estimation model (the deterministic model) is based on the most fine-grained census data at the census-block level for the purpose of estimating sub-block-level populations. Then we adopted a simulation approach to assess the sub-block-level estimates. Population estimation models based on block-level statistics are more accurate than those based on higher level statistics because block-level statistics are closer to sub-block-level statistics than statistics from block groups or census tracts.

Study Area and Data Source

We selected an area of approximately 6×14 km in the north central part of the city of Austin, the capital of Texas, as our study area. The city's three main thoroughfares, IH-35, MoPac, and Highway 183 run through it (Figure 2). The area has a total of 1,337 census blocks with a variety of residential and nonresidential land use, old and new neighborhoods, and housing patterns (Figure 3).

Data sets we used in this study include census geographic and demographic data, building footprints data, aerial photographs, elevation data, and land use data. The data are all from the year 2000. The census data were obtained from the U.S. Census Bureau's American FactFinder Web site (U.S. Census Bureau 2006). The other three data sets were obtained from the City of Austin Neighborhood Planning and Zoning Department (NPZD), either directly downloaded from their File Transfer

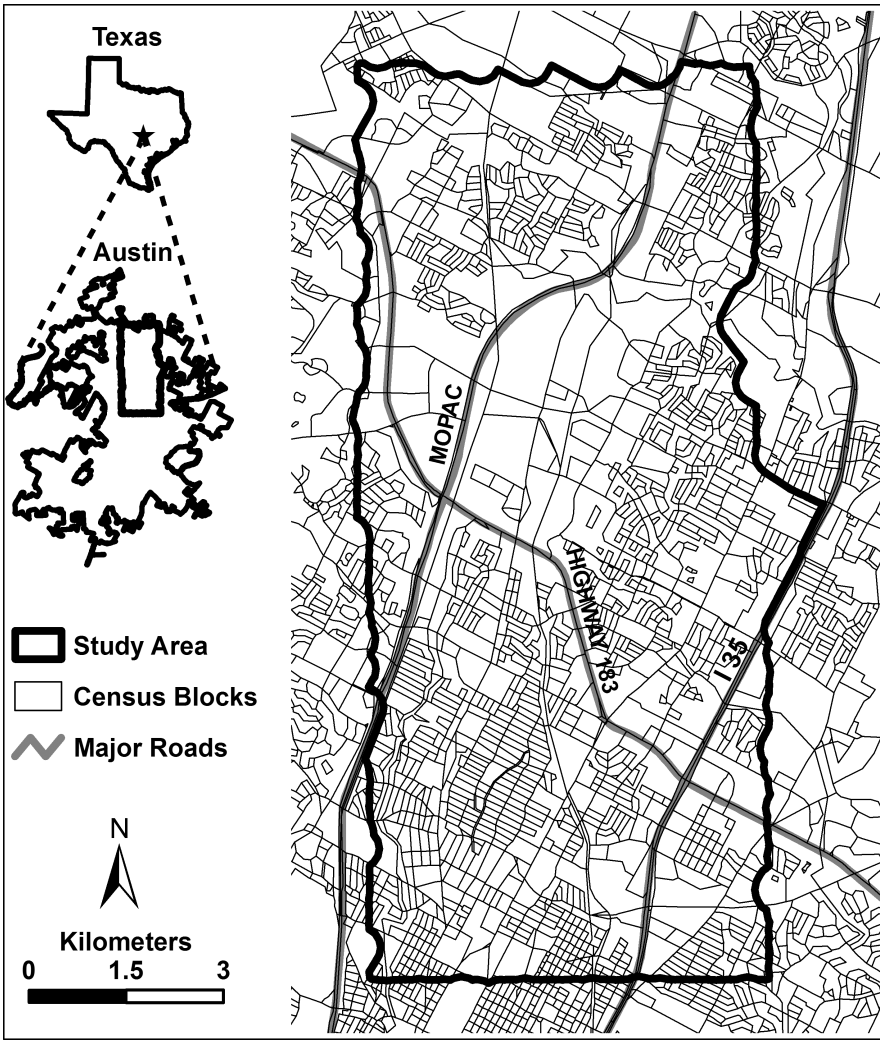


Figure 2 *Study area in Austin, Texas.*

Protocol (FTP) server (City of Austin 2005a) or acquired through personal contact.

In addition to census block geographies and populations, three block-level housing statistics were directly downloaded or computed from relevant census statistics and GIS data: housing unit occupancy rate, average household size, and average space per housing unit.

The building footprints are in vector polygon format. NPZD itself only has building footprints for the years 1997 and 2003 available. Nevertheless, we generated building footprints for the year 2000 for the core part of the city by comparing data sets of the two available years and referencing with aerial photographs from the year 2000. Specifically, for building footprints that do not change between the two years, we assume that they also exist for the year 2000. For the building footprints that are inconsistent between the two years' data sets, we visually referenced with the high-spatial-resolution (0.61 m) aerial photographs to decide which year's data set to follow. For those inconsistent building

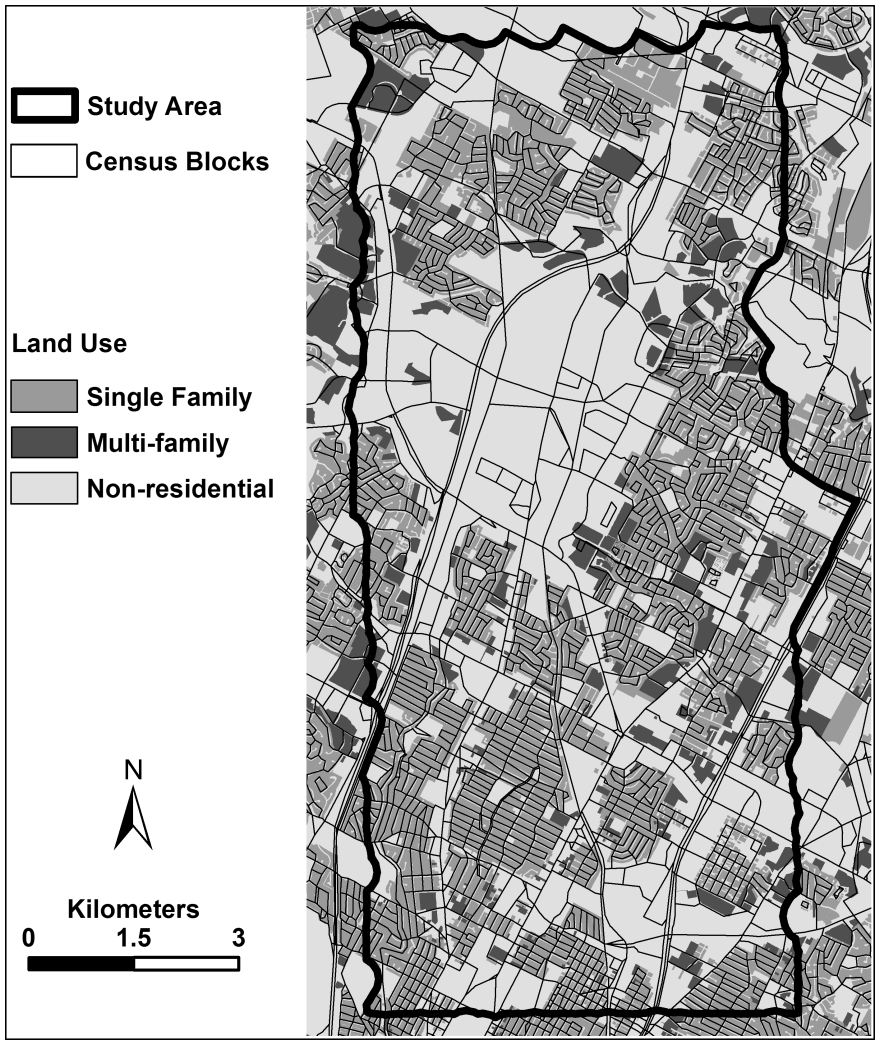


Figure 3 *Land use in the study area.*

footprints that we had to visually check, aerial photographs always matched either one of the two building data sets.

The building footprint data contain the average altitude information for individual building roofs. We inferred the building height from the elevation data of the ground surface. The building footprint data and the elevation data were both generated by Analytical Surveys Incorporated (ASI), which contracted with the city. ASI first manually digitized building footprints from high-resolution aerial photographs. Then, by referencing with digital terrains generated from remote sensing light detection and ranging (LIDAR) data (0.61-m spatial resolution), ASI estimated the altitude for individual building roofs as well as the ground surface. The elevation data of the ground surface is in 0.61-m (2-feet) contour line format. We transferred it to grid format and estimated the average ground surface elevation for individual building footprints. Then, by subtracting ground surface elevation from building roof altitude, we estimated the height for individual

buildings. Individual building volumes were then derived by multiplying the building footprint area with the building height.

It is worth noting that building volumes can also be inferred from appraisal district tax parcel data that contain information about the total acreage of building areas categorized by their floor numbers. However, the building area information is summarized by the entire parcel. Because many multifamily land use parcels are actually entire census blocks, the building information from appraisal districts is not suitable for sub-block-level population estimation. Our land use data are in vector polygon format, generated and updated by the NPZD based on a variety of sources, including historical land use data, Travis Central Appraisal District (TCAD) tax parcel data, the city parcels database, natural preserves GIS data, aerial photographs, building footprint data, and field check information (City of Austin 2005b). The land use data have a spatial unit of tax parcels and are considered quite accurate and reliable.

Assessing Block-Level Population Estimation

We first assessed how the deterministic model estimates populations for census blocks within residential land use areas, specifically single-family and multifamily land use areas, which make up more than 92 percent of the total residential land use areas within the city limit (City of Austin 2005a). By referencing the land use data, we identified 650 single-family blocks within the study area. We further picked 600 single-family blocks that are more connected in geographical boundaries, so that nearby blocks can be more intuitively aggregated into artificial blocks in the simulation analysis. As for multifamily blocks, only thirty-four of them are found in the study area. To increase the number of samples, we searched the entire Austin area and selected a total of 120 blocks that are entirely or mostly within multifamily land use.

After 720 residential blocks were selected, we calculated the total building volumes for individual blocks by overlaying with the building footprint data. The deterministic model (Equation [1]) was then applied to estimate populations for individual sample blocks. We compared the model estimates with the actual block populations from the census and calculated the average percent error as:

$$\text{average percent error} = \frac{1}{m}\sum_{i=1}^{m}\frac{|P_i - Y_i|}{Y_i} \times 100\% \qquad (2)$$

where P_i is the model-estimated population for the ith census block, Y_i is the reported census population for the ith census block, and m is the number of census blocks under investigation. The average percent error gives the average percent of the original census block population that is underestimated or overestimated. A smaller average percent error indicates more accurate estimates from the model. Because the error statistic is simple and straightforward, it is suitable for us to compare estimation accuracy under different contexts regarding block size, rescaling adjustment, and land use information.

We calculated the average percent error of population estimation for the 600 single-family blocks and the 120 multifamily blocks, respectively (Table 1). The results are quite satisfactory, with the average percent error less than 0.15 percent for both land use types. Multifamily blocks have higher estimation errors than single-family blocks. By examining the standard deviation of census-block-level population, we observe that the multifamily blocks have a more varied population distribution than that of single-family blocks, which might cause more uncertainty and errors in estimating population for multifamily blocks.

Assessing Sub-Block-Level Population Estimation

For sub-block-level population estimation, the deterministic model is likely to produce higher

Table 1 *The average percent error of population estimates and the standard deviation of census population for 600 single-family blocks and 120 multifamily blocks*

	600 single-family blocks	120 multi-family blocks
Average percent error of population estimates (%)	0.10	0.14
Standard deviation of census population (persons)	34	573

errors than for block-level estimation because housing statistics are not uniform within individual blocks and block-level housing statistics cannot be well applied to sub-block areas. Given that sub-block-level populations are unavailable to assess sub-block estimates from the deterministic model, we adopted a simulation approach for the sub-block-level estimate assessment. Specifically, we first generated artificial blocks by aggregating every twenty neighboring or nearby census-block areas. The three housing statistics of HuSpace, OccRate, and HdSize for the artificial blocks were also derived from the census data. Then, within each artificial block, we aggregated census-block areas of every two to every nineteen blocks, respectively, as sub-artificial-block areas. Populations for the sub-artificial-block areas were estimated based on the deterministic model. Finally, model estimates for sub-artificial-block areas were compared with census populations of the areas for accuracy assessment. We performed the simulation processes for the 600 single-family blocks, the 120 multifamily blocks, and the combined 720 residential blocks, respectively, and for different sizes of sub-artificial-block areas. The average percent error for the same size of sub-artificial-block areas relative to the associated artificial blocks are graphed in Figures 4 and 5, in which sub-artificial-block areas are represented as percentages of associated artificial blocks. For example, a sub-artificial-block area from an aggregation of ten census-block areas is 50 percent of its associated artificial block, which is an aggregation of twenty census-block areas. Figures 4 and 5 show that the average percent error has an overall increasing trend with decreasing sub-artificial-block areas. Multifamily areas have higher errors and uncertainties (represented by error bounds) of population estimation than those for single-family areas. For example, population estimation for the 50 percent artificial-block areas has approximately 18 percent average percent error, with 9 percent variation for multifamily areas, and approximately 8 percent with 4 percent variation for single-family areas. The higher errors for multifamily areas are due to the more heterogeneous population and housing characteristics than those for single-family areas.

Assessing Sub-Block Estimates at Different Block Sizes

To investigate how the average percent error of population estimates for the same size of sub-artificial-block areas (relative to artificial

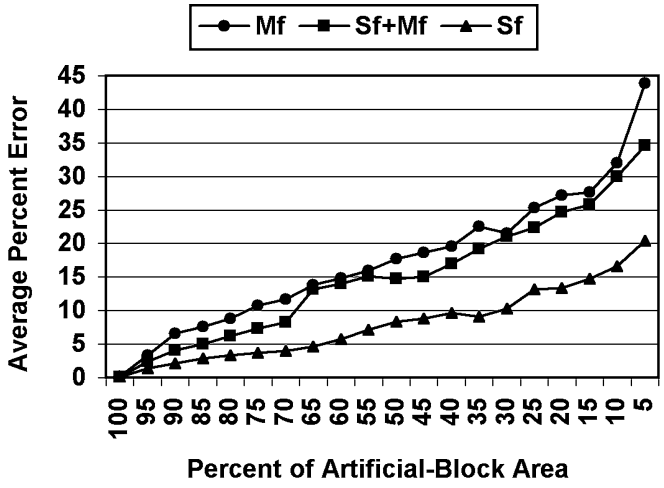


Figure 4 *The average percent error of population estimates for sub-artificial-block areas of single-family (Sf), multifamily (Mf), and combined residential (Sf + Mf) land use.*

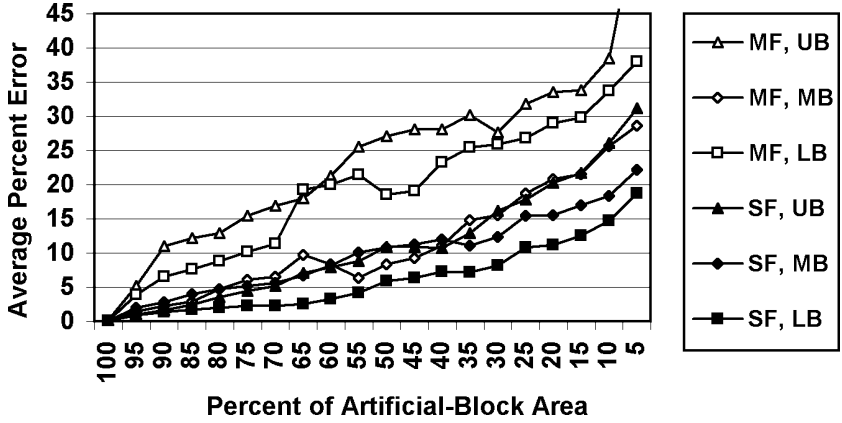
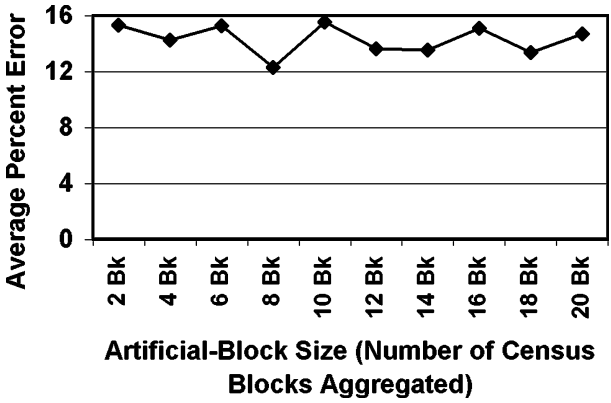


Figure 5 *Upper bound (UB), middle bound (MB), and lower bound (LB) of the average percent error of population estimates for sub-artificial-block areas of single-family (SF) and multifamily (MF) land use.*

blocks) vary with the artificial-block size, we generated artificial blocks from the 720 residential blocks based on every twenty to two census-block aggregations (twenty-block aggregations, nineteen-block aggregations, . . . , two-block aggregations) and assessed population estimates for the 50 percent artificial-block areas for all aggregation schemes. A graph of the average percent error against the artificial-block size is graphed in Figure 6, in which the artificial-block size is represented as the number of aggregated block areas. Figure 6 shows that the average percent error of population estimation for the 50 percent artificial-block areas does not increase or decrease consistently with the artificial-block size. It indicates that population estimation errors for the same size of sub-artificial-block areas would be similar (with a variation of 3 percent) regardless of the artificial-block size. Therefore, when extrapolating the constant error trend to smaller artificial blocks, such as one census block, we can logically infer the population estimation errors for sub-census-block areas. In other words, the derived error graphs for sub-artificial-block areas (Figures 4 and 5) can be applied for sub-census-block areas.

Assessing Effects of Block Population Preservation on Sub-Block Estimates

In the context of census-unit population disaggregation, rescaling subunit estimates is a



The average percent error of population estimates for the 50 percent artificial-block areas of different artificial block sizes (Bk = blocks). **Figure 6**

standard population estimation procedure, in which individual census-unit populations are preserved; that is, the summed total of rescaled subunit estimates within a census unit is equal to the original census-unit population. The rescaled subunit estimates are more reliable because of the population preservation. To investigate whether and how (artificial-)block population preservation improves population estimation for different sizes of sub-(artificial-) block areas, we compared the estimation errors before and after the rescaling. Block population preservation can be achieved by applying a single scaling factor to all sub-block estimates within the same block. A scaling factor is therefore the ratio between the "true" population of a block and the summed total of the sub-block estimates. From another point of view, it is a transfer coefficient or correction factor for a sub-block estimate that is multiplied to yield an adjusted sub-block estimate. After the rescaling procedure, population estimates for block areas will be the accurate estimates, whereas estimates for sub-block areas will still have errors if the sub-block areas do not have the same housing statistics (and associated scaling factors) as those of the associated blocks, which is the likely situation.

Previously we estimated sub-block-level population for residential areas and assessed the estimates (Figure 4). We then rescaled the sub-block estimates and calculated the error statistic. Specifically, we first summed up sub-block estimates to respective block boundaries. We then divided the summed figures by the actual block populations (from the census) to obtain a single scaling factor for all sub-block estimates within each block. The scaling factors were applied to upscale or downscale sub-block estimates, and the rescaled estimates were compared to actual populations of the sub-block areas to derive error statistics. Finally, we compared the error statistics of rescaled sub-block estimates with those of original sub-block estimates. The results show that the rescaling procedure improves population estimation for all sizes of sub-block areas and not in a consistent trend (Figure 7). The rescaling particularly improves population estimation for the 100 percent block areas due to the fact that the rescaled estimates for the 100 percent block areas are now the accurate estimates.

Assessing Sub-Block Estimates When Land Use Information Is Not Available

Residential land use information is not always available when estimating sub-block-level populations. To assess the sub-block estimates of mixed land use, we applied our simulation procedures to a total of 1,320 census blocks that contain residential and nonresidential land use in our study area. Specifically, we first generated artificial blocks by aggregating every twenty neighboring census blocks and calculating the housing statistics for the artificial blocks. Populations for sub-artificial-block areas of different sizes were then modeled, rescaled, and assessed using the procedures as previously described. We compared the error statistics for mixed-land-use areas with those

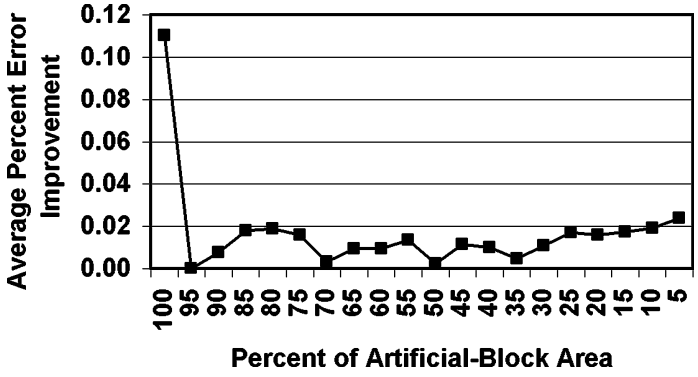


Figure 7 *The improvement of average percent error of population estimation for sub-artificial-block areas after the rescaling for preserving artificial-block populations.*

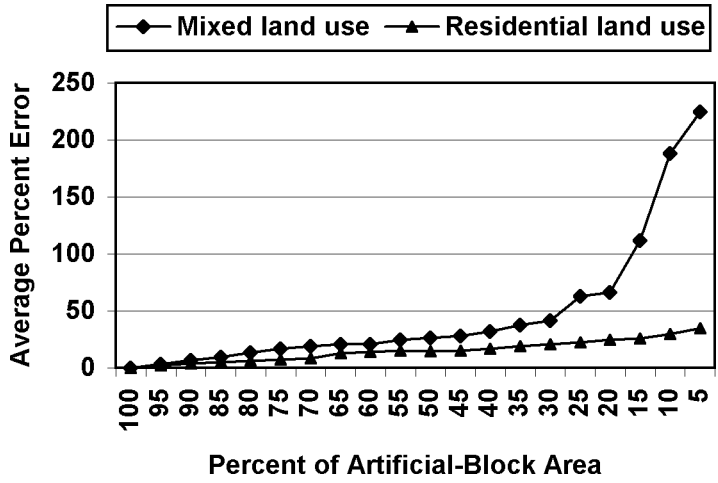


Figure 8 *Average percent error of population estimation for sub-artificial-block areas of mixed land use and residential land use.*

for residential-land-use areas. The results show that mixed-land-use areas have higher sub-block-level population estimation errors, particularly for small sub-block areas (Figure 8). The reason for higher errors in mixed-land-use areas is their relatively heterogeneous population and housing characteristics. From another point of view, some of the buildings in mixed-land-use areas are nonresidential buildings that do not contain residents, leading to inaccurate population estimation.

Discussion

The block-level estimates in this study have average percent errors less than 0.15 percent. Compared to a relevant study by Wu, Qiu, and Wang (2006) that inferred block-level population from image textural statistics and land use information with resulting average percent errors larger than 11 percent, this study has great improvements. As a matter of fact, Wu, Qiu, and Wang (2006) obtained relatively high regression coefficients ($R^2 \geq 0.67$) between population and pixel statistics of high-resolution images to use for population estimation compared to other similar studies (e.g, Liu, Clarke, and Herold 2006). The comparison indicates that inferring fine-scale population based on building volumes and census housing statistics will be more accurate than estimates based on (conventional) variables of image pixel statistics or land use information.

A common conclusion drawn from past population estimation studies is that population estimation for small areas often has higher errors than for large areas (Lo 1995; Harvey 2002a). This study also has similar findings. The deterministic model estimates block-level populations with a high degree of accuracy, but the estimation errors become higher for smaller sub-block areas. For example, the average percent error of population estimation for residential areas is approximately 0.11 percent for 100 percent block areas, 15 percent for 50 percent block areas, and 35 percent for 5 percent block areas (Figure 4). A potential way to improve the model estimates for small sub-block areas is to obtain local, sub-block-level housing statistics to use in the model. They can be obtained through field surveys.

There are a variety of housing patterns and characteristics in different cities and different regions. For example, a new development in a fast-growing city might contain many large multifloored apartments arranged relatively densely, and the occupancy rate is relatively low. On the other hand, an old residential neighborhood in an old city might have small houses with large yard spaces, and the

occupancy rate is relatively high. Nevertheless, the presented deterministic model is a robust model that can be applied to varied U.S. cities because it makes use of census-block-level housing statistics.

Sub-block-level population estimation will be more accurate when local census blocks are at sufficient spatial resolution to capture the variation of local housing patterns. For example, cities with well-defined zoning regulations generally have more homogeneous and large-patch housing patterns, and census blocks within the city limits will be relatively small and able to capture the variation of local housing patterns. As a result, sub-block-level population estimates based on block-level housing statistics will be relatively accurate. On the other hand, in fast-growing suburban areas, census blocks are generally large compared to housing pattern patches, and sub-block-level estimates based on block-level housing statistics will have higher errors. Because sub-block population estimation errors will vary between cities and between different regions of a city depending on the homogeneity of local land use, housing, and population distribution patterns, the error graphs for sub-block-level population estimates derived in this study might not be applicable to other cities and regions. Researchers will need to recalculate the error graphs of sub-block estimates using the proposed simulation approach.

A limitation of the presented fine-scale population estimation model is that it relies on building footprints and associated building volumes for model input. In the past, no effective way of automatically extracting building areas and heights existed. Researchers mainly rely on manually identifying and counting dwelling units from high-spatial-resolution aerial photographs, even though visual interpretation is laborious and time consuming. The building footprints and building volume data used in this study also involve intensive human interpretation and manual work in the derivation process. With the advance of very high-spatial-resolution satellite images, such as IKONOS and QuickBird, and the improvement of feature extraction techniques, automatic extraction of dwelling units from satellite images has become possible (Jin and Davis 2005; Kim, Lee, and Kim 2006). Another prospect for automatic building extraction has come with the advancement of three-dimensional object extraction techniques from LIDAR data (Chen 2007). Building footprints and volumes extracted from LIDAR have shown great accuracy improvements in recent years (Forlani et al. 2006; Zhang, Yan, and Chen 2006). With these new remote sensing data and building extraction techniques, building footprints and building volumes will become more available for population estimation.

Another limitation regarding data source is that the model relies on existing census housing statistics. In other words, if the model is to be applied for noncensus years, an assumption regarding the timeliness of input data must be made or additional up-to-date data must be acquired. For example, if we want to infer fine-scale populations for the year 2006 using the deterministic model, we can either use housing statistics from the Census 2000 data or collect up-to-date housing statistics for model input.

There could be several sources of error in this study due to data quality and accuracy issues, such as spatial misalignment between census block geographies and building footprint data, census data miscounting, and houses unoccupied or otherwise under construction. However, when the sub-block-level estimates are constrained by census block totals, all errors are also constrained within block-level estimates. Therefore, errors due to data quality and accuracy issues will only have an impact on population mapping and estimation for sub-block areas.

Conclusions

This article presents a deterministic model for sub-block-level population estimation based on GIS building volume data and three census block-level housing statistics, including the average space per housing unit, the housing unit occupancy rate, and the average household size. Model estimates for sub-block-level populations are assessed using a simulation approach by generating artificial blocks that are areal aggregations of census blocks and have corresponding housing statistics. The simulation-based assessment further shows that the average percent error of population estimation for sub-block areas of the same size relative to the associated blocks is constant regardless of the

block size. Therefore, it is reasonable to infer the error statistic based on (larger) artificial blocks to those based on (smaller) census blocks. The results show that the smaller the sub-block areas, the higher the estimation errors. For example, the average percent error for residential land use areas is approximately 0.11 percent for the 100 percent block areas, 15 percent for 50 percent block areas, and 35 percent for 5 percent block areas. Furthermore, our analyses show that the rescaling of sub-block estimates for block population preservation improves population estimates for all sub-block areas, and the improvements are not related to the sizes of sub-block areas. Population estimates have higher errors for multifamily areas than for single-family areas, and for mixed-land-use areas than for residential areas, particularly for small sub-block areas, due to the more heterogeneous housing characteristics and population distributions of multifamily and mixed-land-use areas, respectively. As a result, detailed land use information will help with more accurate and reliable sub-block-level population estimation. ■

Literature Cited

Bracken, I. 1991. A surface model approach to small area population estimation. *Town Planning Review* 62 (2): 225–37.

Chen, Q. 2007. Airborne Lidar data processing and information extraction. *Photogrammetric Engineering and Remote Sensing* 73 (2): 109–12.

City of Austin. 2005a. City of Austin GIS data sets. ftp://coageoid01.ci.austin.tx.us/GIS-Data/Regional/coa_gis.html (last accessed 6 June 2006).

———. 2005b. Land use survey methodology. http://www.ci.austin.tx.us/landuse/survey.htm (last accessed 6 June 2006).

Dobson, J. E., E. A. Bright, P. R. Coleman, R. C. Durfee, and B. A. Worley. 2000. LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing* 66 (7): 849–57.

Eicher, C. L., and C. A. Brewer. 2001. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science* 28 (2): 125–38.

Flowerdew, R., and M. Green. 1989. Statistical methods for inference between incompatible zonal systems. In *Accuracy of spatial databases*, ed. M. Goodchild and S. Gopal, 239–47. London: Taylor & Francis.

———. 1991. Data integration: Statistical methods for transferring data between zonal systems. In *Handling geographical information: Methodology and potential applications*, ed. I. Masser and M. Blakemore, 38–54. New York: Wiley.

Forlani, G., C. Nardinocchi, M. Scaioni, and P. Zingaretti. 2006. Complete classification of raw LIDAR data and 3D reconstruction of buildings. *Pattern Analysis and Applications* 8 (4): 357–74.

Green, N. E., and R. B. Monier. 1959. Aerial photographic interpretation of the human ecology of the city. *Photogrammetric Engineering* 25:770–73.

Harvey, J. T. 2002a. Estimating census district populations from satellite imagery: Some approaches and limitations. *International Journal of Remote Sensing* 23 (10): 2071–95.

———. 2002b. Population estimation models based on individual TM pixels. *Photogrammetric Engineering and Remote Sensing* 68 (11): 1181–92.

Hawley, K., and H. Moellering. 2005. A comparative analysis of areal interpolation methods. *Cartography and Geographic Information Science* 32 (4): 411–23.

Holt, J. B., C. P. Lo, and T. W. Hodler. 2004. Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science* 31 (2): 103–21.

Hsu, S. Y. 1971. Population estimation. *Photogrammetric Engineering* 37:449–54.

Jin, X., and C. H. Davis. 2005. Automated building extraction from high-resolution satellite imagery in urban areas using structural, contextual, and spectral information. *Eurasip Journal on Applied Signal Processing* 2005 (14): 2196–206.

Kim, T., T. Lee, and K. Kim. 2006. Semiautomatic building line extraction from Ikonos images through monoscopic line analysis. *Photogrammetric Engineering and Remote Sensing* 72 (5):541–49.

Kraus, S. P., L. W. Senger, and J. M. Ryerson. 1974. Estimating population from photographically determined residential land use types. *Remote Sensing of Environment* 3 (1): 35–42.

Lam, N. 1983. Spatial interpolation methods: A review. *The American Cartographer* 10 (2): 129–49.

Liu, X., and K. C. Clarke. 2002. Estimation of residential population using high resolution satellite imagery. In *Proceedings of the 3rd Symposium in Remote Sensing of Urban Areas, June 11–13, 2002*, ed. D. Maktav, C. Juergens, and F. Sunar-Erbek, 153–60. Istanbul, Turkey: Istanbul Technical University Press.

Liu, X., K. C. Clarke, and M. Herold. 2006. Population density and image texture: A comparison study. *Photogrammetric Engineering & Remote Sensing* 72 (2): 187–96.

Lo, C. P. 1989. A raster approach to population estimation using high-altitude aerial and space

photographs. *Remote Sensing of Environment* 27 (1): 59–71.

———. 1995. Automated population and dwelling unit estimation from high-resolution satellite images—A GIS approach. *International Journal of Remote Sensing* 16 (1): 17–34.

———. 2003. Zone-based estimation of population and housing units from satellite-generated land use/land cover maps. In *Remotely sensed cities*, ed. V. Mesev, 157–80. New York: Taylor & Francis.

Lo, C. P. and H. F. Chan. 1980. Rural population estimation from aerial photographs. *Photogrammetric Engineering and Remote Sensing* 46 (3): 337–45.

Lo, C. P. and R. Welch. 1977. Chinese urban population estimates. *Annals of the Association of American Geographers* 67 (2): 246–53.

Martin, D. 1989. Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers* 14 (1): 90–97.

———. 1996. An assessment of surface and zonal models of population. *International Journal of Geographical Information Systems* 10 (8): 973–89.

Mennis, J. 2003. Generating surface models of population using dasymetric mapping. *The Professional Geographer* 55 (1): 31–42.

Prosperie, L., and R. Eyton. 2000. The relationship between brightness values from a nighttime satellite image and Texas county population. *The Southwestern Geographer* 4:16–29.

Rase, W. 2001. Volume-preserving interpolation of a smooth surface from polygon-related data. *Journal of Geographical Systems* 3 (2): 199–213.

Reibel, M., and M. E. Bufalino. 2005. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A* 37 (1): 127–39.

Robinson, A., J. Morrison, P. Muehrcke, A. Kimerling, and S. Guptill. 1995. *Elements of cartography*. New York: Wiley.

Tobler, W. R. 1969. Satellite confirmation of settlement size coefficients. *Area* 1 (3): 30–34.

———. 1979. Smooth pyncophylactic interpolation for geographical regions. *Journal of the American Statistical Association* 74 (367): 519–30.

U.S. Census Bureau. 2003. 2000 Census of population and housing selected appendixes. http://www.census.gov/prod/cen2000/phc-2-a.pdf (last accessed 15 October 2006).

———. 2006. The U.S. Census American FactFinder. http://factfinder.census.gov/home/saff/main.html?_lang=en (last accessed 6 June 2006).

Weber, C. 1994. Per-zone classification of urban land use cover for urban population estimation. In *Environmental remote sensing from regional to global scales*, ed. G. M. Foody and P. J. Curran, 142–48. New York: Wiley.

Webster, C. J. 1996. Population and dwelling unit estimation from space. *Third World Planning Review* 18 (2): 155–76.

Wright, J. K. 1936. A method of mapping densities of population. *The Geographical Review* 26 (1): 103–10.

Wu, S., X. Qiu, and L. Wang. 2006. Using semivariance image texture statistics model population densities. *Cartography and Geographic Information Science* 33 (2): 127–40.

Xie, Y. 1995. The overlaid network algorithms for areal interpolation problem. *Computers Environment and Urban Systems* 19 (4): 287–306.

Yuan, Y., R. M. Smith, and W. F. Limp. 1997. Remodeling census population with spatial information from LandSat TM imagery. *Computers, Environment and Urban Systems* 21 (3–4): 245–58.

Zhang, K., J. Yan, and S. Chen. 2006. Automatic construction of building footprints from airborne LIDAR data. *IEEE Transactions on Geoscience and Remote Sensing* 44 (9): 2523–33.

SHUO-SHENG WU is a Program Faculty in the Department of Geography at Texas State University—San Marcos, 601 University Drive, San Marcos, TX 78666, and a UCGIS Postdoctoral Fellow at the U.S. Geological Survey, 1400 Independence Road, Rolla, MO 65401. E-mail: swu@usgs.gov. His research interests include GIS, remote sensing, and spatial statistics applications in land use classification, population estimation, hazard risk assessment, and watershed pollution modeling, as well as scale and resolution issues in spatial data analysis.

LE WANG is an Assistant Professor in the Department of Geography at the University of Buffalo, State University of New York, Wilkeson Quad, Buffalo, NY 14261. E-mail: lewang@buffalo.edu. His research interests include development of new methods for remote sensing, mangrove forest characterization, invasive species modeling, and urban population estimation.

XIAOMIN QIU is an Assistant Professor in the Department of Geography, Geology, and Planning at Missouri State University, 901 S. National Avenue, Springfield, MO 65897. E-mail: qiu@missouristate.edu. Her research interests include GIScience education, mapping, geovisualization, spatial cognition, applications of GIS and remote sensing, and distance learning.