Transactions in GIS, 2009, 13(4): 401–414

Research Article

A GIS Approach to Estimation of Building Population for Micro-spatial Analysis

KoKo Lwin
Graduate School of Life and Environmental Sciences
University of Tsukuba

Yuji Murayama
Graduate School of Life and Environmental Sciences
University of Tsukuba

Abstract

Population data used in GIS analyses is generally assumed to be homogeneous and planar (i.e. census tracts, townships or prefectures) due to the public unavailability of building population data. However, information on building population is required for micro-spatial analysis for improved disaster management and emergency preparedness, public facility management for urban planning, consumer and retail market analysis, environment and public health programs and other demographic studies. This article discusses a GIS approach using the Areametric and Volumetric methods for estimating building population based on census tracts and building footprint datasets. The estimated results were evaluated using actual building population data by visual, statistical and spatial means, and validated for use in micro-spatial analysis. We have also implemented a standalone GIS tool (known as 'PopShape GIS') for generating new building footprint with population attribute information based on user-defined criteria.

1 Introduction

Research into micro-spatial analysis has increased due to the emergence of high spatial resolution satellite images for urban areas and the availability of fine-scale GIS data with enhanced attribute information (e.g. building footprints with the number of floors, building use type and building name). In view of the advances in modern geospatial information technologies, this is a good time for studying the world at a micro level. Population count is a key anchor for much of the official statistical system and the benchmark for many commercial and research surveys and analyses (Cook 2004). GIS plays a critical role in population studies and analyses by means of mapping the spatial

Address for correspondence: KoKo Lwin, Division of Spatial Information Science, Graduate School of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8572, Japan. E-mail: kokolwin@geoenv.tsukuba.ac.jp

© 2009 Blackwell Publishing Ltd
doi: 10.1111/j.1467-9671.2009.01171.x

extent and analyzing it along with other GIS datasets. However, estimating and mapping the population is not an easy task due to the nature of human activities which change over space and time. Normally, population can be estimated using statistical and spatial (remote sensing and GIS) approaches. For example, night-time city lights imagery has been shown to demonstrate a reasonable correlation with population (Sutton 2003). Developments in computer hardware and mapping software have already encouraged many statistical and census offices to move from traditional cartographic methods to digital mapping and geographic information systems (GIS) (Rhind 1991, Ben-Moshe 1997, United Nations 1997). The benefits of geographic data automation in statistics are shared by the users of census and survey data. The data integration functions provided by GIS, which allow linking of information from many different subject areas, have led to much wider use of statistical information. This, in turn, has increased the pressure on statistics agencies to produce high-quality spatially referenced information for small geographic units. The types of applications for such data are almost limitless. Examples include planning of social and educational services, poverty analysis, utility service planning, labor force analysis, marketing analysis, voting district delineation, emergency planning, epidemiological analysis, floodplain modeling and agricultural planning (United Nations 2000). The main objective of this study was to apply GI science theory and practice to produce a smaller geographic unit of population data (i.e. at the building level) for improved accuracy in the decision-making process at the micro level.

2 Problems in Micro-spatial Analysis

Openshaw (1992) identified the following sources of error in micro-spatial analysis: errors in the positioning of objects; errors in the attributes associated with objects; and errors in modeling spatial variation (e.g. by assuming spatial homogeneity between objects). Population data exhibits spatial variation, especially in areas with a mix of high- and low-rise buildings such as Tokyo, and residential areas mixed with unpopulated spaces (paddy fields, parks, playgrounds or government institutions) such as in Tsukuba City. Moreover, the population data used in GIS analyses is generally at the level of census tracts, townships or prefectures, since building population data is not available for public use due to privacy concerns. For spatial information users, population data has generally only been available in township polygons or city point features with aggregated population data. For non-spatial information users, population data (text and tables) can be obtained from the National Census Bureau or local government offices. All of these datasets are suitable for local and regional analysis, but not for micro-spatial analysis and decision-making processes.

In general, population mapping has two purposes: firstly, to cartographically portray the extent and density of population across an area of interest; and secondly, to derive a quantitative estimation of population density for use in subsequent spatial analytical modeling tasks (Bielecka 2005). Common cartographic forms of population mapping are the choropleth map and the dasymetric map. Choropleth maps provide an easy way to visualize how a measurement varies across a geographic area. However, choropleth maps have limited utility for detailed spatial analysis of population data, especially where the population is concentrated in a relatively small number of villages, towns and cities. Moreover, choropleth maps cannot express statistical variation within the administrative areal units, such as changing population density. One way to avoid this limitation is by transforming the administrative units into smaller and more relevant map units through a process known as dasymetric mapping (Bielecka 2005). Dasymetric maps use ancillary information to help delineate new zones that better reflect the changing patterns over space. Recent research suggests that dasymetric mapping can provide small-area population estimates that are more accurate than many areal interpolation techniques that do not use ancillary data (Mrozinski and Cromley 1999, Gregory 2002). The U.S. Geological Survey (USGS) has refined and extended automated processes for improving spatial accuracy and visualization in mapping population distribution using dasymetric mapping (Sleeter 2008). This technique aims to refine the spatial accuracy of aggregated data by using ancillary information to partition space into zones that better reflect the statistical variation in population. However, most census boundaries do not coincide with or intersect the boundaries of geographic features such as land use/land cover, soil type, geological unit, and floodplain and watershed boundaries – an issue that is known as "spatial incongruity". This may introduce inaccurate population results for environmental assessment and emergency preparedness. Moreover, under the GIS domain, spatial analysis functions performed within the census tract do not acquire any significant changes in population. To overcome these problems, this article introduces a GIS approach for estimating building population for micro-spatial analysis.

3 Methodology

3.1 Applied Method

Here, we introduce two estimation methods: (1) Areametric (which does not require information on the number of building floors); and (2) Volumetric (which does require information on the number of floors). For improved accuracy, we also allow filtering by other categories into the computation, such as filtering by minimum footprint area and building use types, e.g. commercial, industrial, educational, and other building use types that are not occupied by residents. The calculation is demonstrated by the following mathematical expressions:

Areametric Method:

$$BP_i = \left(\frac{CP}{\sum_{\mathrm{k}=1}^{\mathrm{n}} BA_k} \right) BA_i \qquad \text{Using building footprint surface area} \qquad (1)$$

Volumetric Method:

$$BP_i = \left(\frac{CP}{\sum_{\mathrm{k}=1}^{\mathrm{n}} BA_k \cdot BF_k} \right) BA_i \cdot BF_i \qquad \text{Using number of floors information} \qquad (2)$$

Moreover, advances in remote sensing data acquisition technologies such as LiDAR can be used for extraction of building footprints, building height (Digital Height Model (DHM)) and building volume (Digital Volume Model (DVM)). Equations (3) and (4) can be used for LiDAR data:

© 2009 Blackwell Publishing Ltd

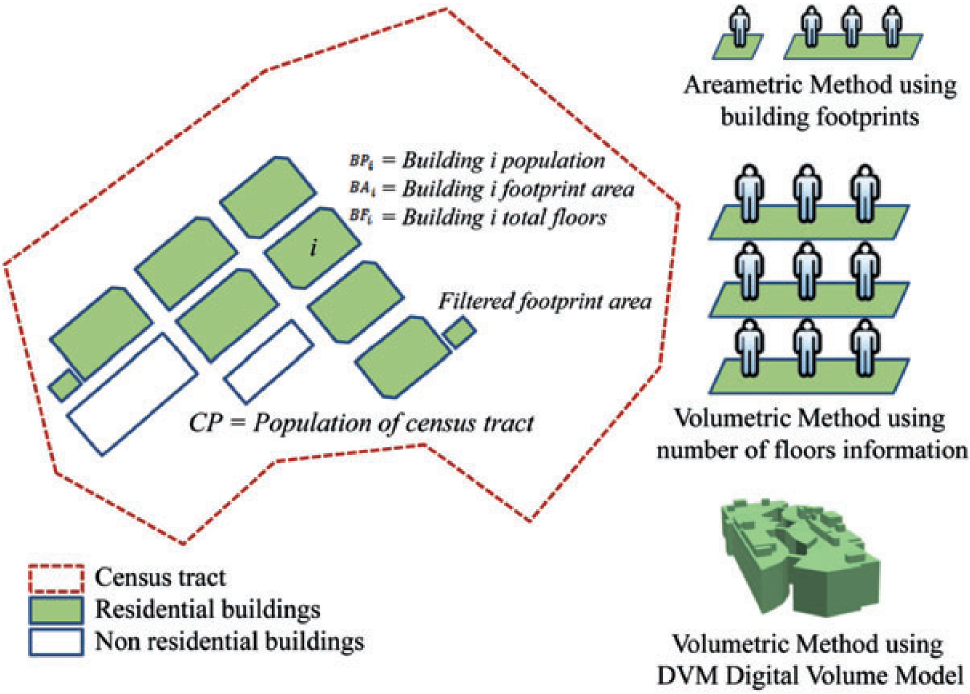

BP_i = Building i population
BA_i = Building i footprint area
BF_i = Building i total floors
i
Filtered footprint area
CP = Population of census tract
Census tract
Residential buildings
Non residential buildings
Areametric Method using building footprints
Volumetric Method using number of floors information
Volumetric Method using DVM Digital Volume Model

Figure 1 Graphical illustration of equations

$$BP_i = \left(\frac{CP}{\sum_{k=1}^{n} BA_k \cdot BH_k} \right) BA_i \cdot BH_i \quad \text{Using average building height} \tag{3}$$

$$BP_i = \left(\frac{CP}{\sum_{k=1}^{n} BV_k} \right) BV_i \quad \text{Using total building volume} \tag{4}$$

where BP_i = the population of building i, CP = the census tract population, BA_i = the footprint area of building i, BF_i = the number of floors of building i, BH_i = the average height of building i (from LiDAR data), BV_i = the total volume of building i (from LiDAR data), i, k = summation indices, and n = the number of buildings that meet user-defined criteria and fall inside the CP polygon (Figure 1).

The Areametric method is suitable for low-rise buildings especially in rural areas while the Volumetric method is suitable for high-rise buildings, especially in downtown areas.

3.2 Test Data

The two methods were evaluated using actual building population data acquired from the city administration office for study purposes. These data include detailed information about each building such as age, construction material, building type (detached, non-detached, semi-detached, flat or apartment), building use type (residential, commercial or

© 2009 Blackwell Publishing Ltd

educational, etc.), number of floors, number of households and total number of people, which is intended for use in disaster management. The test data was converted into the ESRI Shapefile format and after conversion the building footprint polygons totaled 9,913, the census tracts totaled 28 and the total population was 28,000.

3.3 Test Method

In order to identify the best results based on data availability, we employed both methods (Areametric and Volumetric) with filtering by various footprint areas such as 0, 5, 10, 15, 20, 25, 30, 35, and 40 m^2 applied to the residential building use type.

3.4 Implementation of GIS Tool

To achieve the goal, we implemented a standalone GIS tool using the Visual Basic programming language and TatukGIS DK (Development Kit). This latter tool is partially embedded in the SIOSSS system (Spatially Integrated Online Social Survey System), which makes it possible to collect, store, share and visualize spatially distributed public survey datasets for local government decision-making processes such as for public facility management, emergency response and disaster preparedness, and market analysis. We are testing this system in Tsukuba City, a planned city located in Ibaraki Prefecture, Japan. Under the SIOSSS system, the PopShape tool (Figures 2 and 3) will automatically populate survey items into building footprints for micro-spatial analysis users, since SIOSSS does not collect particulars from respondents such as room number, building number and street name due to privacy concerns.

Users can define the minimum ignored footprint size such as for porticos, garbage boxes and other unpopulated areas. They can also apply filtering by attribute field(s) such as building use type and other attribute information. Three additional approaches are available under the Volumetric method, namely Use Number of Floors, Use Average Building Height and Use Total Building Volume. After processing, the estimated building population attribute field, "EST_POP", appears in a new ESRI Shapefile. A map viewer is also provided for viewing the processed results by performing common GIS functions such as add map layer, zoom in, zoom out, get attribute information, label by attribute field and change map layer properties.

The operational steps employed (with numbers corresponding to those shown in Figure 3) were as follows: (1) Open census tracts file (Shape polygon); (2) open building footprints file (Shape polygon); (3) filter by footprint size; (4) filter by building use type(s); (5) select method (Areametric or Volumetric); (6) select approach (use Number of Floors or Building Height or Building Volume); (7) select appropriate field (Floor or Height or Volume attribute field); (8) assign output file name; and (9) start to process (see http://giswin.geo.tsukuba.ac.jp/sis/en/software.html for additional details).

4 Results and Accuracy Assessment

All estimated values were evaluated using actual building population data by means of visual, statistical and spatial approaches. We obtained the best results using the Volumetric method filtered at the 25 m^2 footprint category. Figures 4 and 5 show the results for the preferred method visually for example.

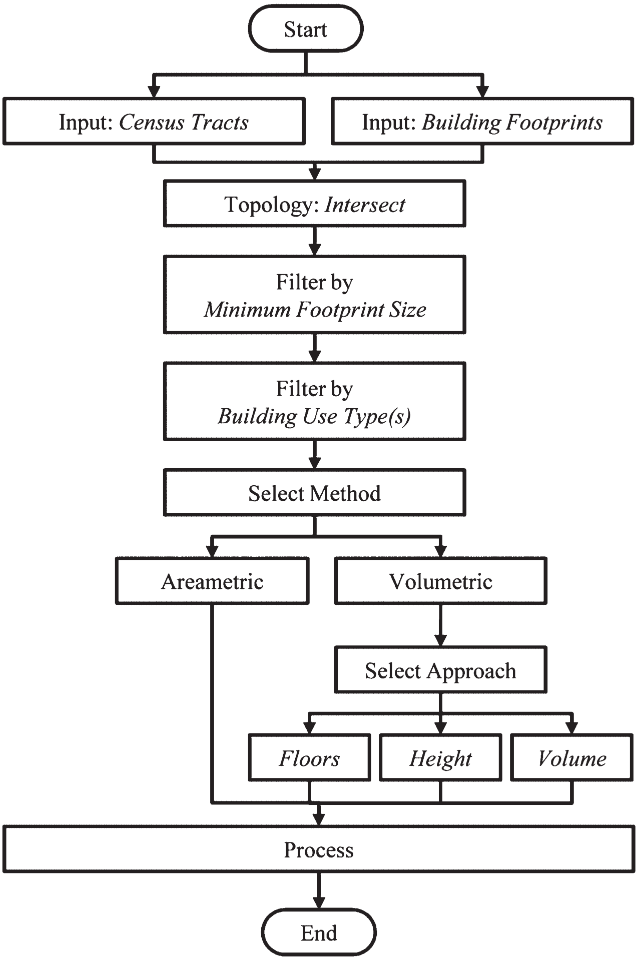
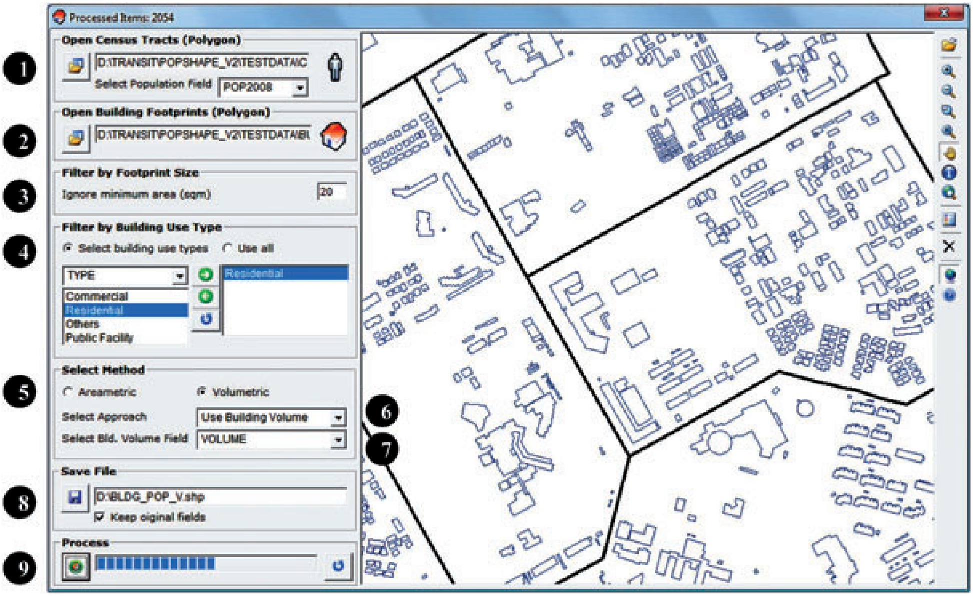


Figure 2 PopShape GIS program flow chart

To examine the statistical relationship between the two datasets (estimated and actual population), we applied linear regression analysis to determine the correlation coefficient, R^2. Tables 1 shows the results of correlation coefficients for various filtered footprint areas using two estimation methods. We also calculated the root mean square error (RMSE) for each category (Table 2) in both estimated methods:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\text{Actual} - \text{Estimated})^2}{n}} \qquad (5)$$

The smallest RMSE was achieved using the 25 m^2 filtered building footprint applied in the Volumetric method. With the Areametric method, the estimated values did not agree with the actual values, especially in highly populated buildings (population >50). This may have occurred due to the presence of high-rise buildings with a large number of floors but a small footprint. We achieved the best estimated results using the Volumetric



Processed Items: 2054

1 Open Census Tracts (Polygon)
D:\TRANSIT\POPSHAPE_V2\TESTDATA\C
Select Population Field: POP2008

2 Open Building Footprints (Polygon)
D:\TRANSIT\POPSHAPE_V2\TESTDATA\BL

3 Filter by Footprint Size
Ignore minimum area (sqm): 20

4 Filter by Building Use Type
● Select building use types ○ Use all
TYPE
Commercial
Residential
Others
Public Facility
Residential

5 Select Method
○ Areametric ● Volumetric
Select Approach: Use Building Volume
Select Bld. Volume Field: VOLUME

8 Save File
D:\BLDG_POP_V.shp
☑ Keep oiginal fields

9 Process

Figure 3 PopShape GIS program graphical user interface (GUI)

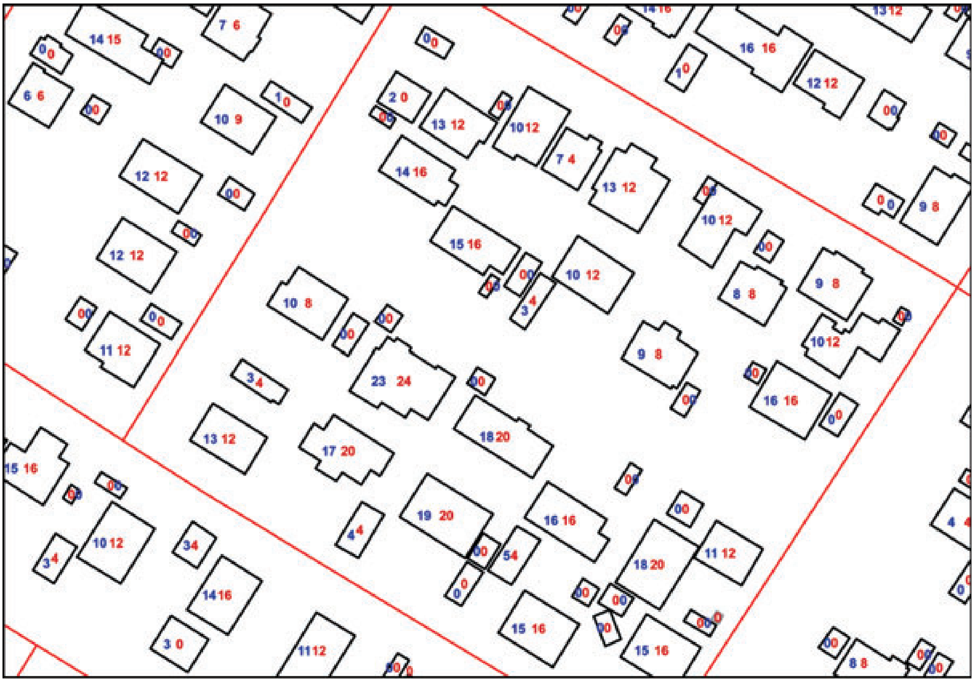


Figure 4 Estimated population (left value) vs. actual population (right value) in low-rise-building area (Filtered area 25 m^2, Volumetric method applied to residential buildings)

© 2009 Blackwell Publishing Ltd

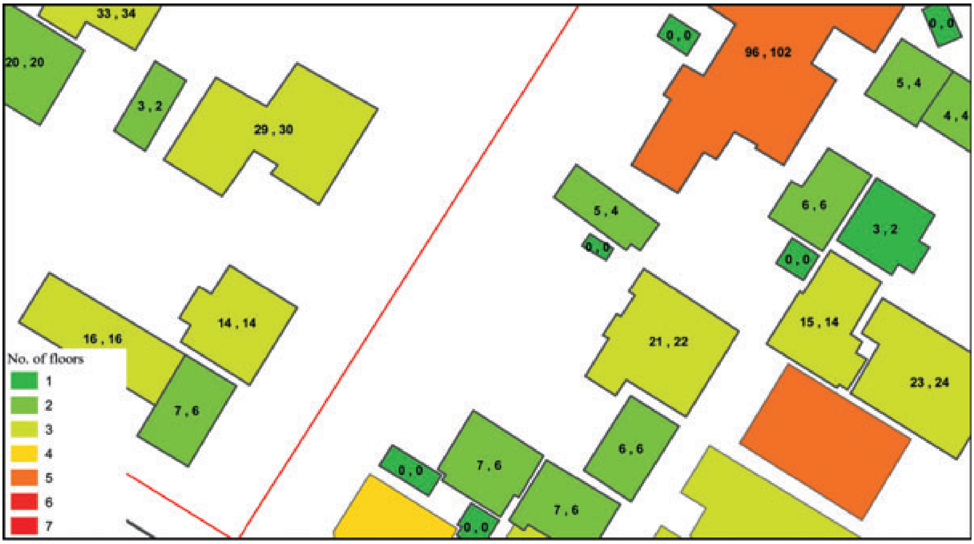


Figure 5 Estimated population (left value) vs. actual population (right value) in high-rise-building area (Filtered area 25 m^2, Volumetric method applied to residential buildings)

Table 1 Various correlation coefficients for Areametric and Volumetric methods

Filtered Area	Areametric Method R^2 & Y	Volumetric Method R^2 & Y
0 m^2	R^2 = 0.8004 (y = 0.5785x + 1.3214)*	R^2 = 0.9461 (y = 0.8625x + 0.3699)
05 m^2	R^2 = 0.8003 (y = 0.5795x + 1.3241)	R^2 = 0.9461 (y = 0.8628x + 0.3708)
10 m^2	R^2 = 0.7995 (y = 0.5898x + 1.3022)	R^2 = 0.9467 (y = 0.8688x + 0.3765)
15 m^2	R^2 = 0.7995 (y = 0.6160x + 1.2197)	R^2 = 0.9468 (y = 0.8794x + 0.3748)
20 m^2	R^2 = 0.7990 (y = 0.6431x + 1.1348)	R^2 = 0.9479 (y = 0.8934x + 0.3402)
25 m^2	R^2 = 0.8002 (y = 0.6631x + 1.0703)	R^2 = 0.9488 (y = 0.9037x + 0.3088)*
30 m^2	R^2 = 0.7971 (y = 0.6739x + 1.0312)	R^2 = 0.9458 (y = 0.9117x + 0.2773)
35 m^2	R^2 = 0.7954 (y = 0.6823x + 1.0078)	R^2 = 0.9439 (y = 0.9189x + 0.2571)
40 m^2	R^2 = 0.7944 (y = 0.6926x + 0.9751)	R^2 = 0.9425 (y = 0.9275x + 0.2306)

* Best results

method where all R^2 values were greater than 0.9. While all R^2 values are acceptable in the Volumetric method, the best value (R^2 = 0.9488) was achieved in the 25 m^2 filtered area category (Figures 6 and 7).

We also performed spatial assessments to compare the spatial distribution patterns between the estimated and actual building population. There are several ways to perform spatial autocorrelation, the most popular being Moran's I and Geary's C. Here, we used Moran's I to compare each value in the pairs to the mean value for all features in the study area, which is also known as global Moran's I. Moran's I was computed for each filtered category in both the Areametric and Volumetric methods based on the estimated

Table 2 Root mean square error (RMSE) for both Areametric and Volumetric methods

Filtered Area	RMSE (Areametric)	RMSE (Volumetric)
0 m^2	0.0315520	0.0153020
5 m^2	0.0315193	0.0152933
10 m^2	0.0312114	0.0150973
15 m^2	0.0304115	0.0148670
20 m^2	0.0296994	0.0145070
25 m^2	0.0291742	0.0142609*
30 m^2	0.0290846	0.0145403
35 m^2	0.0290064	0.0147061
40 m^2	0.0288724*	0.0148335

* Best results

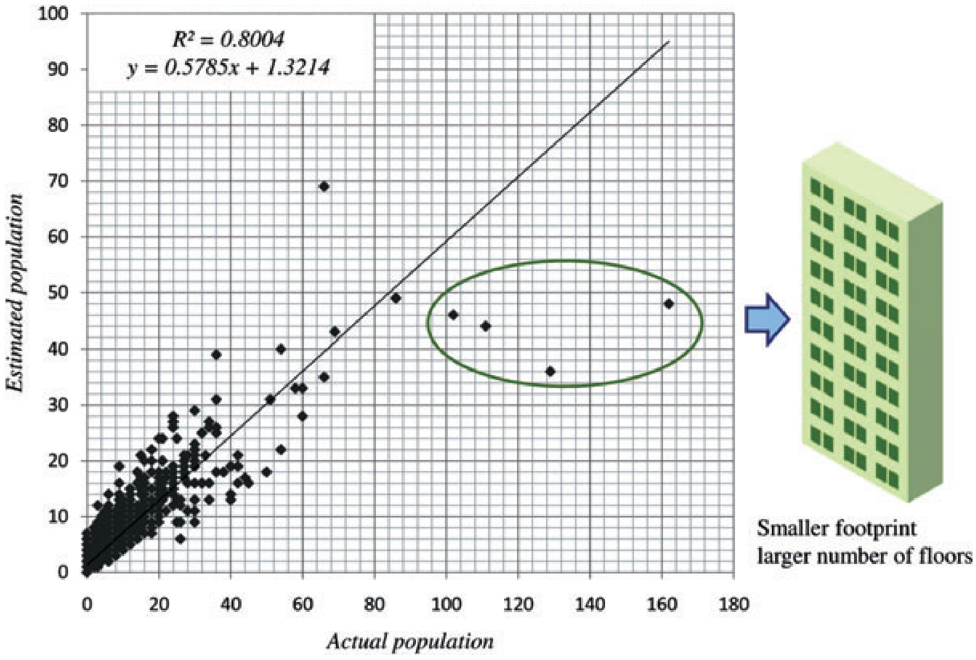


Figure 6 Scatter plot for 0 m^2 filtered area in Areametric method (Total sample size = 8,854)

population class field using ArcGIS. This tool measures the spatial autocorrelation (feature similarity) based not only on feature locations or attribute values alone but also on feature locations and feature values simultaneously. Given a set of features and an associated attribute (i.e. estimated population or actual population), it evaluates whether the pattern expressed is clustered, dispersed, or random. The tool calculates the Moran's I index value and a Z score evaluating the significance of the index value. In general, a

© 2009 Blackwell Publishing Ltd

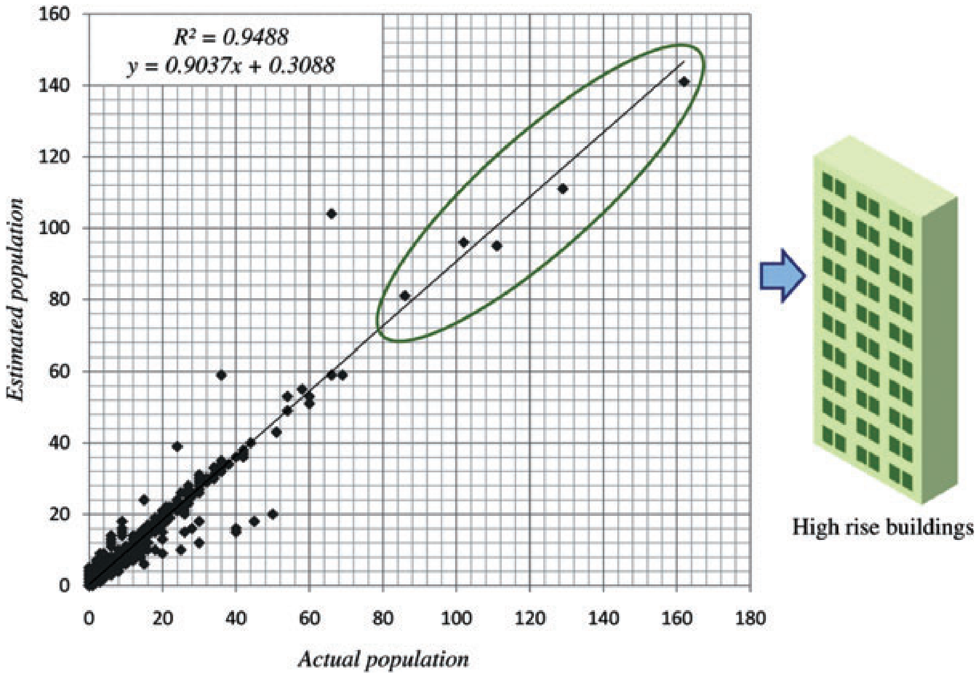


Figure 7 Scatter plot for 25 m^2 filtered area in Volumetric method (Total sample size = 8,854)

Table 3 Moran's I and Z score for actual building population

Actual Building Population				
Filtered Area	Index	Expected	Variance	Z score
None	0.03271	–0.00021	0.00000	62.05492

Moran's I index value near +1.0 indicates clustering while an index value near –1.0 indicates dispersion. A high positive Z score for a feature indicates that the surrounding features have similar values. A low negative Z score indicates that the feature is surrounded by dissimilar values. Moreover, Moran's I for actual population is also computed for comparison. Although Moran's I measures the patterns to determine whether the features are clustered or dispersed, the purpose of using Moran's I in this study was to measure the patterns for each category and then compare them with the feature patterns of the actual building population. Tables 3 through 5 show the Moran's I indexes in each filtered category for both the Areametric and Volumetric methods.

Although both indices intersect at certain filtered areas, one of the Z scores of the Volumetric method intersected at a point between the 20 and 25 m^2 footprint areas (Figure 8). This is probably the average single-unit living space in the study area.

Table 4 Moran's I and Z score for estimated building population using the Areametric method

Areametric Method				
Filtered Area	Index	Expected	Variance	Z score
None	0.02508	–0.00014	0.00000	65.48493
05 m^2	0.02529	–0.00014	0.00000	66.06920
10 m^2	0.03178*	–0.00014	0.00000	81.18271*
15 m^2	0.04734	–0.00016	0.00000	111.56468
20 m^2	0.06082	–0.00018	0.00000	130.03565
25 m^2	0.06845	–0.00020	0.00000	137.12042
30 m^2	0.07450	–0.00021	0.00000	143.36652
35 m^2	0.07615	–0.00021	0.00000	142.61562
40 m^2	0.07727	–0.00022	0.00000	141.69034

* Best results

Table 5 Moran's I and Z score for estimated building population using the Volumetric method

Volumetric Method				
Filtered Area	Index	Expected	Variance	Z score
None	0.01851	–0.00018	0.00000	39.79721
05 m^2	0.01726	–0.00011	0.00000	54.03807
10 m^2	0.01867	–0.00018	0.00000	40.15002
15 m^2	0.02128	–0.00018	0.00000	45.31663
20 m^2	0.02845	–0.00019	0.00000	58.47552
25 m^2	0.03451*	–0.00020	0.00000	68.89462*
30 m^2	0.03799	–0.00021	0.00000	73.47143
35 m^2	0.03892	–0.00022	0.00000	73.43831
40 m^2	0.04002	–0.00022	0.00000	74.06918

* Best results

5 Potential Applications

The estimated or quantitative mapping of building population is essential for micro-spatial analysis especially in terms of emergency management. Effective disaster preparedness requires quantitative spatial distribution patterns of population in order to position emergency response centers and prepare food and shelter in the event of disaster. Building population data is also required for improved accuracy in cost estimation of food and shelter for emergency preparedness and other humanitarian assistance (Figure 9). City and urban planners need to know how many local residents will benefit

© 2009 Blackwell Publishing Ltd

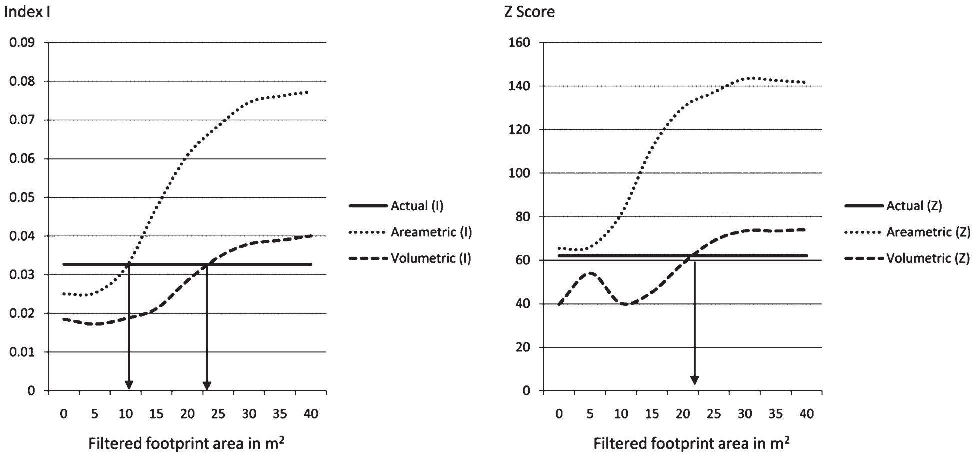


Figure 8 Moran's I (left) and Z score (right) for actual population vs. estimated population using Areametric and Volumetric methods for various filtered footprint areas

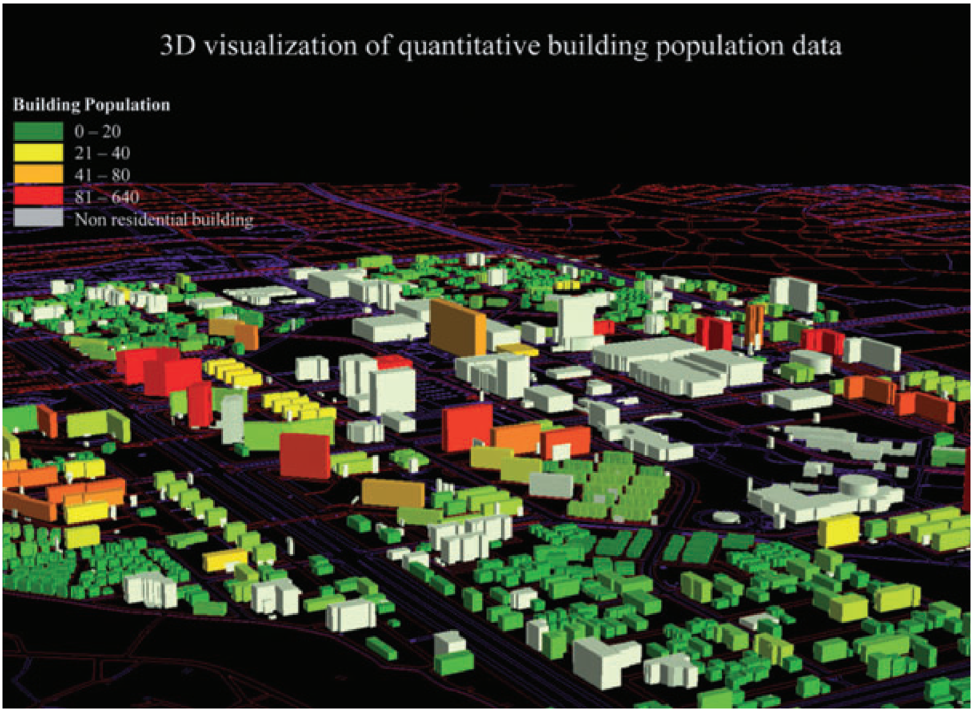
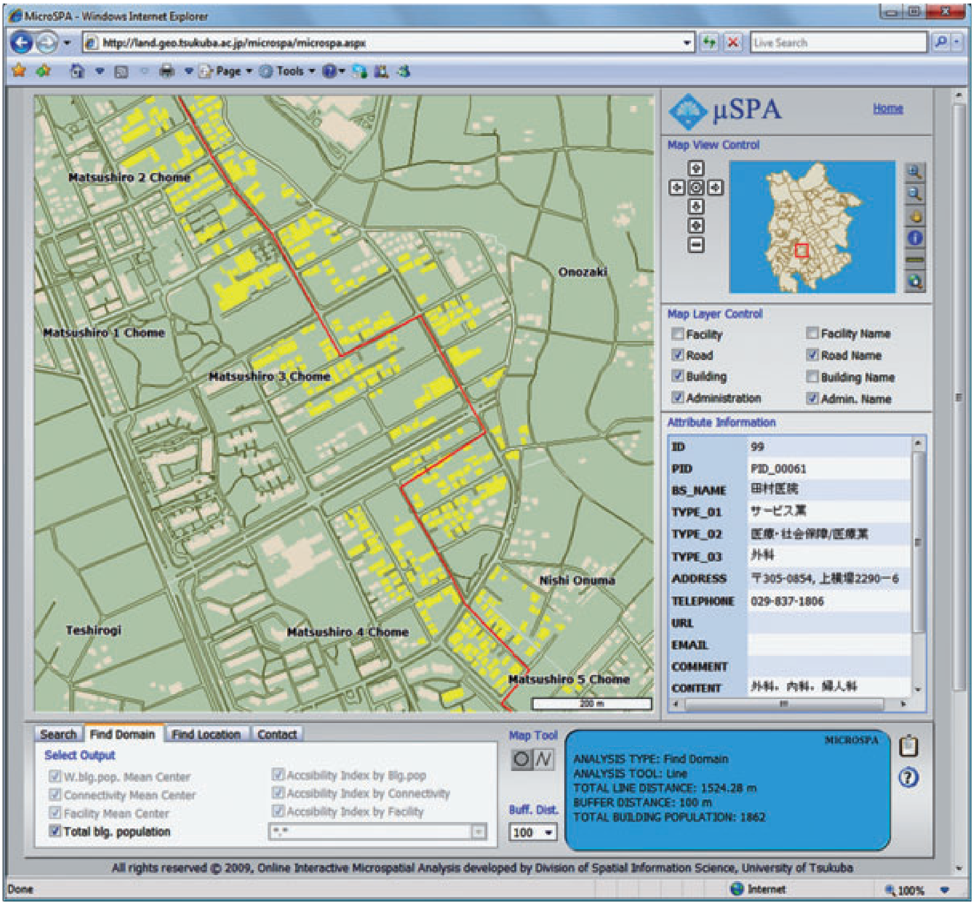


Figure 9 Example of 3D quantitative building population mapping (3D visualization is one of the techniques used for effective public facility and utility planning)

from newly constructed public facilities such as bus centers, railway stations and hospitals (Figure 10). Hydrologists require an estimate on the number of people living on a floodplain. Potential business owners can define their business location and perform consumer analysis. Quantitative building population data can be used as a weighted



Map Layer Control

☐ Facility ☐ Facility Name
☑ Road ☑ Road Name
☑ Building ☐ Building Name
☑ Administration ☑ Admin. Name

Attribute Information

ID	99
PID	PID_00061
BS_NAME	田村医院
TYPE_01	サービス業
TYPE_02	医療・社会保障/医療業
TYPE_03	外科
ADDRESS	〒305-0854, 上横場2290−6
TELEPHONE	029-837-1806
URL	
EMAIL	
COMMENT	
CONTENT	外科、内科、婦人科

Search | Find Domain | Find Location | Contact

Select Output

☑ W.blg.pop. Mean Center ☑ Accsibility Index by Blg.pop
☑ Connectivity Mean Center ☑ Accsibility Index by Connectivity
☑ Facility Mean Center ☑ Accsibility Index by Facility
☑ Total blg. population

Buff. Dist. 100

ANALYSIS TYPE: Find Domain
ANALYSIS TOOL: Line
TOTAL LINE DISTANCE: 1524.28 m
BUFFER DISTANCE: 100 m
TOTAL BUILDING POPULATION: 1862

All rights reserved © 2009, Online Interactive Microspatial Analysis developed by Division of Spatial Information Science, University of Tsukuba

Figure 10 Example of web-based interactive decision-making tool for planning local community bus route based on building population (determining the shortest route with larger building population within a specific buffer zone)

factor in spatial statistical analysis such as for determining population mean center and standard distance (Figure 11). This is important for decision making related to population such as in selecting a voting site or building a new public facility. The potential application for quantitative building population data is unlimited and we hope to increase the accuracy in various spatial decision-making processes at the micro level.

6 Conclusions and Future Work

In this article we have discussed the application of two building population estimation methods based on census tract and building footprint data which may be applied when building population data is not publicly available for privacy reasons. We have also developed a tool to generate a GIS-ready dataset with quantitative building population attribute information. The estimated results were evaluated using visual, statistical and spatial approaches. We have achieved reasonable results, confirming model suitability for

© 2009 Blackwell Publishing Ltd

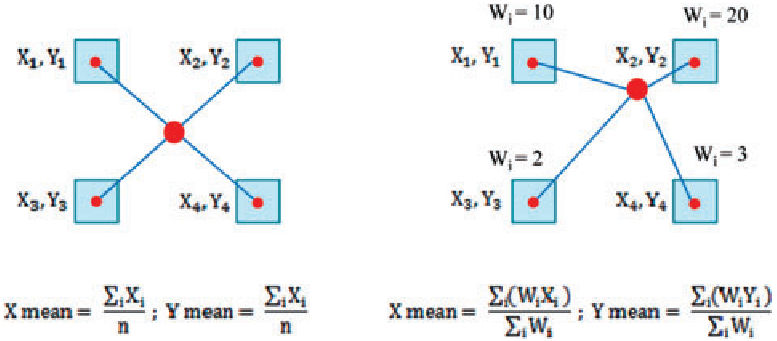

X_1, Y_1
X_2, Y_2
X_3, Y_3
X_4, Y_4
X mean = $\frac{\Sigma_i X_i}{n}$; Y mean = $\frac{\Sigma_i X_i}{n}$
$W_i = 10$
$W_i = 20$
$W_i = 2$
$W_i = 3$
X_1, Y_1
X_2, Y_2
X_3, Y_3
X_4, Y_4
X mean = $\frac{\Sigma_i (W_i X_i)}{\Sigma_i W_i}$; Y mean = $\frac{\Sigma_i (W_i Y_i)}{\Sigma_i W_i}$

Figure 11 Example of determining the mean center: (a) Mean center without weighted factor (spatially oriented); and (b) Mean center using building population as a weighted factor (population oriented)

use in micro-spatial analysis. However, further development is required to improve the accuracy by incorporating other factors such as an adjustment factor for mixed building-use type (i.e. residential buildings mixed with commercial) and building status (i.e. newly constructed or abandoned buildings). Through integration with modern spatial data acquisition technologies such as LiDAR and other high-resolution satellite imagery, we hope to achieve more accurate estimation of building population.

References

Ben-Moshe E 1997 Integration of a national GIS project within the planning and implementation of a population census. In *Proceedings of the Euro-Mediterranean Workshop on New Technologies for the 2000 Census Round*, Ma'ale Hachamisha, Israel (available at http://www.cbs.gov.il/mifkad/euromedit.htm)

Bielecka E 2005 A dasymetric population density map of Poland. In *Proceedings of the International Cartographic Conference*, A Coruña, Spain

Cook L 2004 The quality and qualities of population statistics, and the place of the census. *Area* 36: 111–23

Gregory I N 2002 The accuracy of areal interpolation techniques: Standardizing 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems* 26: 293–314

Mrozinski R D Jr and Cromley R G 1999 Singly and doubly constrained methods of areal interpolation for vector-based GIS. *Transactions in GIS* 3: 285–301

Openshaw S 1992 Learning to live with errors in spatial databases. In *Goodchild M F and Gopal S (eds) The Accuracy of Spatial Databases*. New York, Taylor and Francis: 263–76

Rhind D W 1991 Counting the people: The role of GIS. In Maguire D J, Goodchild M F, and Rhind D W (eds) *Geographical Information Systems: Volume 1, Principles and Applications*. London, Longman: 127–37

Sleeter R 2008 A new method for mapping population distribution. Washington, DC, U.S. Geological Survey Fact Sheet 2008–3010 (available at http://pubs.usgs.gov/fs/2008/3010/)

Sutton P 2003 Estimation of human population parameters using night-time satellite imagery. In Mesev V (ed) *Remotely Sensed Cities*. London, Taylor and Francis: 301–34

United Nations 1997 *Geographical Information Systems for Population Statistics*. New York, United Nations Studies in Methods, Series F, No 68

United Nations 2000 *Handbook on Geographic Information Systems and Digital Mapping*. New York, United Nations Department of Economic and Social Affairs Statistics Division