International Journal of Remote Sensing
Vol. 31, No. 21, 10 November 2010, 5673–5688

Taylor & Francis
Taylor & Francis Group

Improving the housing-unit method for small-area population estimation using remote-sensing and GIS information

CHENGBIN DENG†, CHANGSHAN WU*† and LE WANG‡
†Department of Geography, University of Wisconsin-Milwaukee, P.O. Box 413,
Milwaukee, WI 53201., USA
‡Department of Geography, State University of New York at Buffalo,
Buffalo, NY 14261, USA

Small-area population estimates for a non-census year are essential for supporting a wide variety of planning processes. Many demographic or geographic-information-based models have been developed for generating small-area population estimates. Little research, however, attempted to integrate these two types of models to achieve a better estimation. This study explores the feasibility of incorporating geographic information system (GIS), remote-sensing and demographic data into the housing-unit (HU) method, a popular demographic model, to estimate small-area population in Grafton, WI, USA. In particular, two major components of the HU method, HU counts and persons per household (PPH), are obtained by modelling their relationships with demographic and geographic factors using a sequence of ordinary least-squares (OLS) regression models. Analysis of results indicates that spatial factors derived from remote sensing and GIS datasets, together with demographic information, can significantly improve the accuracy of small-area population estimation.

1. Introduction

Accurate estimates of small-area population are essential for supporting a wide variety of planning processes. The size and distribution of the population are often key determinants for resource allocation for state and local governments (Smith *et al.* 2002). Population estimates are critical in decisions about when and where to build public facilities such as schools, libraries, sewage treatment plants, hospitals and transportation infrastructure. Population estimates are also often used by private sectors for customer-profile analysis, market-area delineation and site-location identification (Martin and Williams 1992, Plane and Rogerson 1994). In addition, population information is an important input in many urban and regional models, such as land-use and transportation-interaction models, urban-sprawl analysis, environment-equity studies and policy-impact analysis (Rees *et al.* 2004). Clearly, accurate and timely population estimates are of great importance (Smith *et al.* 2002). Accurate population data, however, is only available for every decade through the national census survey. It is obvious that this frequency does not meet the needs for rapid-growth areas where noteworthy local intercensal population changes occur. Thus, appropriate estimation methods for such geographical areas are extremely necessary.

Numerous methods have been proposed for population estimates in demography, such as the component method II (CM-II), the administrative-record (AR) method, the ratio-correlation (RC) method and the housing-unit (HU) method (Ghosh and Rao

*Corresponding author. Email: cswu@uwm.edu

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online © 2010 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/01431161.2010.496806

1994). The CM-II updates population estimates by accounting for the major components of local demographic change. The population is calculated by starting with the recent census population, adding the estimated number of births, subtracting the number of deaths, then adding net migration and the changes in group-quarter population. The AR method, similar to the CM-II, additionally uses administrative records for estimating net migration for population under age 65. These records include federal income-tax returns, immigration and naturalization service records and military-movement records. The RC method relates population changes to the changes in several symptomatic indicators, such as school enrolment, car registration, work-force information and occupied HUs. A regression model is used to construct the relationship between population changes and the changes of symptomatic indicators between two census years. In the HU method, population estimates for a small area are calculated as the product of the number of occupied HUs and household size plus the population in group quarters. Among all these methods, the HU method is the most commonly used and considered one of the most accurate and cost-effective methods for small-area population estimation (US Census Bureau 1998, Smith and Cody 2004). In fact, the HU method has been used by US Census Bureau as the single method for estimating population of subcounty areas (e.g. incorporated places and minor civil divisions) since 1996 (US Census Bureau 1998, 2005, Smith and Cody 2004).

In addition to these demographic models, remote-sensing and geographic information system (GIS) techniques provide an alternative for small-area population estimation. One early application of the remote-sensing technique is the house-counting approach. In particular, dwelling units are manually counted from high-resolution aerial photographs, and then these counts are multiplied by the surveyed household size to derive population estimates (Lo 1986a,b). This approach, although relatively accurate, requires a tremendous amount of time and labour and is rarely employed by state and local agencies. To address this issue, automatic approaches have been proposed for HU and population estimation. With these approaches, either spectral radiance/reflectance or urban physical parameters are extracted from remote-sensing imagery to represent housing information. With these parameters, regression models are typically applied to derive population estimates (Lo 1995, Webster 1996, Chen 2002, Harvey 2002a,b, Li and Weng 2005, Wu and Murray 2007). Similarly, existing GIS datasets, such as transportation network and land-use land-cover data, were also applied to derive population estimates (Qiu *et al.* 2003, Wu *et al.* 2005).

These demographic models and remote-sensing-/GIS-based approaches for population estimation have been developed almost in parallel. Each of them, however, has its own issues. The HU method, a popularly applied demographic approach, can only produce population estimates at an aggregated geographical level (e.g. town, city or county), instead of a finer local level (e.g. census-block level). This is mostly due to the problems of data acquisition. Currently, commonly used data sources of HU information are building permits and electric-customer information, most of which can only be obtained at an aggregated geographical level. Although some methods have been developed to derive population and HU estimates at the block-group level using simple interpolation and step-down techniques (Perry and Voss 1996), such estimates, however, are unreliable. These techniques assume that population count in a census block is a linear function of its geographic area, thereby distributing population from an aggregated unit to a block accordingly. Such assumption, however, is problematic due to the heterogeneous population distribution within a geographic area. On the other side, remote-sensing-/GIS-based automatic techniques can reveal detailed spatial information (either radiance/reflectance or physical parameters). This information, however, is not directly associated with HUs, and may produce large errors when applied as the solo data source to estimate HUs or population (Harvey 2002a,b, Li and Weng 2005, Wu and Murray 2007). Therefore, state and local agencies seldom used remote-sensing/GIS-based automatic approaches for producing small-area population estimates.

The other problem of both approaches is that the relationship between population and HUs (or indicative parameters extracted from remote-sensing imagery) is always assumed to be unchanged from the most recent decennial census (US Census Bureau 2005). Several researches have pointed out that the persons per household (PPH) in the US has had a tremendously downward trend during the past two centuries, falling from 5.8 in 1790 to 4.8 in 1900, then to 2.6 in 2000 (Kobrin 1976, Bongaarts 2001, US Census Bureau 2001). This drop is most likely due to declining birth rates and the tendency for adults to head separate households. Acknowledging the problems of using historical PPH values, Starsinic and Zitter (1968) developed PPH estimation methods through extrapolating historical trends. Smith (1986) adjusted the estimated PPH in small areas by examining changes in larger areas (e.g. states) where current PPH estimates are available from other sources. These simple approaches may generate biased results when PPH trends are not stable (Smith *et al.* 2002). Smith *et al.* (2002) estimated the county-level PPH by introducing population-age structure variables, births, school enrolment and Medicare enrolment (for age 65 and older). Currently, studies on the PPH estimation are rare, and it is of great necessity to estimate the PPH at the detailed level (e.g. census block).

To address the above issues associated with the HU method and remote-sensing-/GIS-based automatic approaches, we propose to integrate GIS and remote-sensing techniques into the HU method for deriving better small-area population estimates. In particular, the first objective of this paper is to redistribute new HUs at an aggregated geographic level (e.g. village or town) to census blocks with the help of remote-sensing and GIS information. The second objective is to develop a model for better PPH estimation at the census-block level through incorporating remote-sensing, GIS and demographic data. The remainder of this paper is organized as follows. The next section introduces the study area and data. The third section details model development and accuracy-assessment techniques. Analysis of results and accuracy assessment are discussed in §4, and finally conclusions and future research are given in §5.

2. Study area and data

The village and town of Grafton, Wisconsin, USA (see figure 1) is selected as the study area for this research. Located in the north of Milwaukee City, Grafton has a geographic area of 66.1 km^2, including a variety of land-use types, such as residential, commercial, transportation, forestry and agricultural, as well as other rural lands. Since 1990, Grafton has been experiencing rapid population growth, with a significant amount of residential and commercial developments. Its population was 13 330 in 1990 and rose to 14 444 in 2000, with an increment of 7.7%. In addition, the HU number rose from 4827 in 1990 to 5773 in 2000, with 946 new HUs constructed, or an increment of 16.4%. Due to the noteworthy population growth in this area, detailed population estimates are essential for supporting urban and rural planning.

For the study area, population data at the census-block level in 1990 and 2000 were acquired from the US Census Bureau. In Grafton, there were 266 blocks in 1990 and 313 blocks in 2000. In fact, a number of census blocks in 1990 had been sub-divided into several

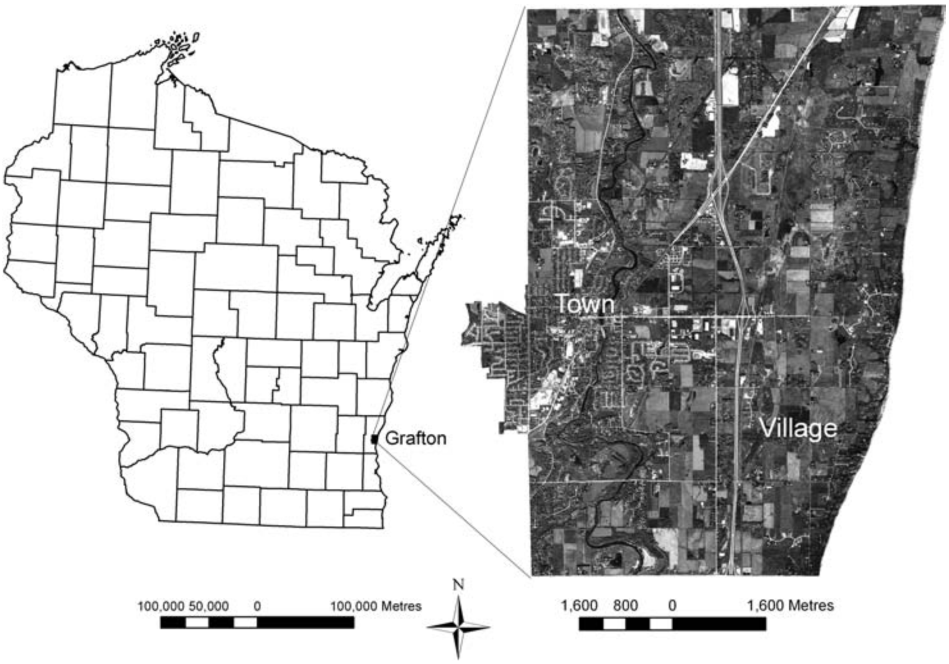


Figure 1. Study area as the town and village of Grafton, WI, USA.

small blocks in 2000. Moreover, the boundaries of several census blocks have been modified. To address this data inconsistency problem, an area-weighted spatial interpolation (Goodchild and Lam 1980) was applied to the 1990 block data and associated information was transferred to the 2000 block boundary. In addition to census data, detailed land-use data for 1990 and 2000 (see figure 2) were obtained from the Southeast Wisconsin Regional Planning Commission. A Landsat Thematic Mapper (TM) image acquired on 1 August 1989 and a Landsat Enhanced Thematic Mapper Plus (ETM+) image acquired on 8 September 2000 were downloaded from the WisconsinView project website (http://www.wisconsinview.org/). Both images were reprojected to the Universal Transverse Mercator (UTM) projection (zone 16, datum WGS84), and a further georeference was conducted to reduce geometry misregistration. The residual mean square (RMS) of the georeferenced image was within 0.2 pixel. Moreover, atmospheric correction was performed using the algorithm developed by Richter (1996a,b, 2005). In particular, this method involves two steps: (1) calculating the reflectance of every pixel based on standard atmospheres, aerosol types and visibility and (2) correcting the adjacency effect using weighting functions. More details about this algorithm can be found in Richter (1996a, 2005).These Landsat images will be used to extract urban biophysical information for better estimation of HUs and PPH. In this research, demographic and spatial data in 1990 are employed for model development, and data in 2000 are applied for model calibration and accuracy assessment.

3. Methodology

3.1 *HU estimation*

With traditional demographic methods, the number of new HUs is typically estimated from building permit and electrical-customer information. This information, however,

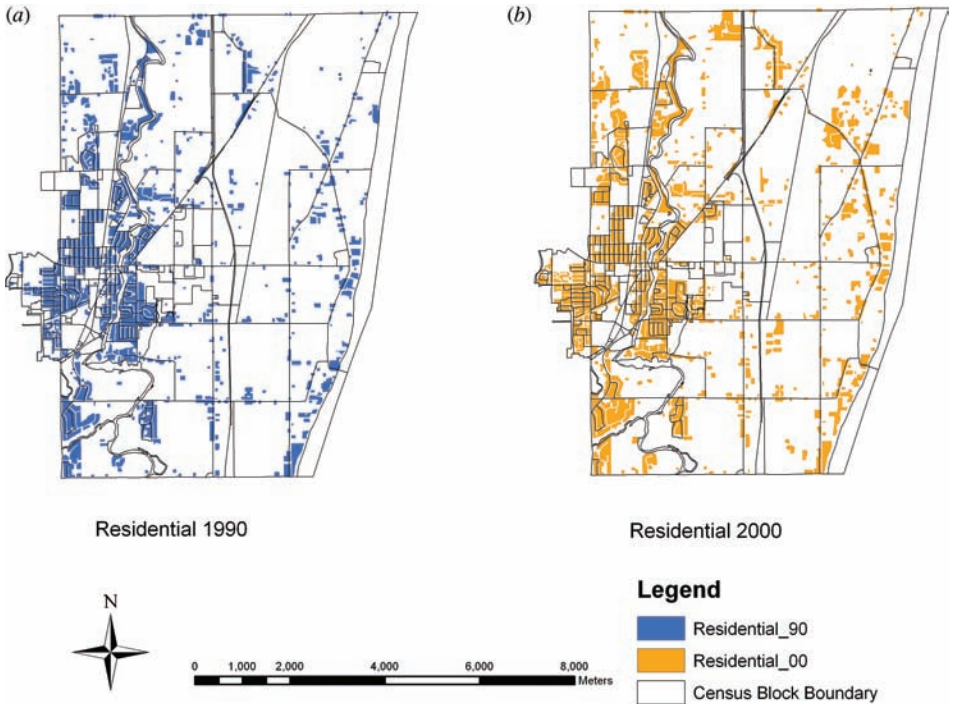


Figure 2. Detailed land-use data of (*a*) 1990 and (*b*) 2000 obtained from the Southeast Wisconsin Regional Planning Commission (residential land uses are shown in colour).

can only be obtained at an aggregated geographical level, and there is a need to assign these new HUs to individual blocks. In this paper, two interpolation methods were employed to derive HU estimates at the census-block level. This first method is the simple step-down interpolation technique (Perry and Voss 1996), with which new HUs at the aggregated level are assigned to each census block according to the geographical area of that block. The second method is the dasymetric-mapping method developed in this paper. Dasymetric mapping was created for visualizing population data consistent to land-use types to avoid the problem of ecological fallacy. Dasymetric mapping has been popularly used for population-density estimation and areal interpolation (Mennis 2003, Wu and Murray 2005, Mennis and Hultgren 2006), but not particularly applied for HU estimation. In this paper, a dasymetric-mapping method was developed to redistribute newly developed HUs to individual census blocks with the help of ancillary remote-sensing and GIS data. In particular, three variables derived from remote-sensing and GIS datasets are used. These variables include: (1) single-family land-use area change, (2) multi-family land-use area change and (3) the change of normalized difference vegetation index (NDVI), representing vegetation-cover change. It is assumed that the number of newly developed HUs is positively associated with land-use area changes (e.g. single-family and multi-family) and negatively related to vegetation-cover change. In order to quantify the weight for each variable, an ordinary least-squares (OLS) regression analysis model was developed as follows:

$$\Delta \mathrm{HU} = \alpha_0 + \alpha_1 \Delta R_{\mathrm{S}} + \alpha_2 \Delta R_{\mathrm{M}} + \alpha_3 \Delta \mathrm{NDVI}, \qquad (1)$$

where ΔHU is the number of newly developed HUs in a block from time t_1 to time t_2, ΔR_S is the areal change of single-family land use in the block, ΔR_M is the areal change of multi-family land use in the block, ΔNDVI is the change of NDVI values in that block, and α_1, α_2 and α_3 are their coefficients, respectively. After obtaining the relationship among newly developed HUs and the variables derived from remote-sensing and GIS datasets, the total HUs for the aggregated area (e.g. town and village of Grafton) can be re-distributed to each census block.

3.2 *PPH estimation*

Compared with the HU numbers, an accurate PPH estimate is more essential since a minor error of PPH may result in a significant error in the estimated population. In this paper, the current method, in which the PPH is assumed to be unchanged from the previous census and three regression models that use demographic and remote-sensing and GIS variables were developed. Demographic and economic variables include population-age structures (e.g. parentage of people with 65 and over), household income and housing values. Variables derived from remote-sensing and GIS data include distances to particular land uses (e.g. commercial centre, schools and recreational areas), vegetation-cover changes and textural information generated from Landsat TM/ETM+ imagery. With these demographic and GIS-/remote-sensing-related variables, the three regression models are described as follows.

The first model (model A) assumes that the relationship between PPH and demographic and spatial variables is unchanged over time. Therefore, it is possible to estimate the PPH of time t_2 based on the parameters obtained from the data of time t_1. Thus, this model is an 'inherit model', in which the PPH for time t_2 is estimated according to the relations obtained from the data of time t_1. The regression model is illustrated as follows:

$$\text{model A: PPH}_t = \alpha_0 + \sum_{i=1}^{m} \beta_i S_{i,t} + \sum_{j=1}^{n} \gamma_j D_{j,t}, \tag{2}$$

where PPH$_t$ is the PPH value at time t, $S_{i,t}$ indicates a spatial variable derived from remote-sensing and GIS data at time t, $D_{j,t}$ represents a demographic variable at time t, m and n are the total number of spatial and demographic variables, respectively, and α, β and γ are regression coefficients.

Model A assumes that the relationship between PPH and relevant variables does not change over time. This assumption, however, may be problematic. To have a better estimation of PPH, a special survey is needed to obtain the knowledge of PPH and demographic variables for time t_2. Therefore, the second model (model B) attempts to estimate the PPH at time t_2 with the PPH value at time t_1 and spatial and demographic variables at time t_2 as independent variables. This model is named an 'empirical model', as the PPH at time t_1 is employed as an independent variable:

$$\text{model B: PPH}_{t_2} = \alpha_0 + \alpha_1 \text{PPH}_{t_1} + \sum_{i=1}^{m} \beta_i S_{i,t_2} + \sum_{j=1}^{n} \gamma_j D_{j,t_2}, \tag{3}$$

where PPH$_{t_1}$ and PPH$_{t_2}$ are the PPH at time t_1 and time t_2, respectively, and S_{i,t_2} and D_{j,t_2} are the spatial and demographic variables at time t_2, respectively.

In addition to model B, the third model (model C) explores the relationship between the change of PPH and the changes of spatial and demographic variables from time t_1

to time t_2. Since this model represents the relationships between the changes of variables, it is named a 'change model'. The formulation of this model is as follows:

$$\text{model C: } \Delta\text{PPH} = \alpha_0 + \sum_{i=1}^{m} \beta_i \Delta S_i + \sum_{j=1}^{n} \gamma_j \Delta D_j, \tag{4}$$

where ΔPPH is the change of PPH for a census block from time t_1 to time t_2, and ΔS_i and ΔD_j are the changes of spatial and demographic variables, respectively.

3.3 *Small-area population estimation*

With the HU and PPH estimates for each census block, it is feasible to derive the population estimates via the HU method, which can be expressed as follows:

$$P_t = \text{HU}_t \times O_t \times \text{PPH}_t + \text{GQ}_t, \tag{5}$$

where P_t is the estimated population for a small area at a non-census time t, HU_t is the number of HUs at time t, O_t is the occupancy rate at time t, PPH_t is the average size of household or PPH at time t and GQ_t is the group-quarter population (e.g. persons residing in college dormitories, military barracks, nursing homes and prisons) at time t.

With the estimated values of HU and PPH derived from §§3.1 and 3.2, it is feasible to estimate the population for each census block with information of occupancy rate (O_t) and group-quarter population (GQ_t). In this research, the occupancy rate at time t_2 is assumed to be unchanged from time t_1, and this information may also be estimated from data acquired from a special survey, if available. For an area without large group-quarter facilities, the group-quarter population for a non-census year can be assumed to be unchanged or change proportionally to the household population (Smith and Lewis 1980). In this research, group-quarter population in Grafton at time t_2 is regarded to be the same as the value at time t_1. Therefore, with the estimated values of HU, PPH, O and GQ at time t_2, the population estimate for a census block at time t_2 can be derived.

3.4 *Accuracy assessment*

To assess the estimation accuracy of HUs, PPH and population, a spatially random sampling method was used to divide all census blocks into two groups: 50% of data was used as a training dataset for modelling and the other 50% of data serves as a testing dataset to assess estimation accuracy. Several error measurements, including the mean absolute error (MAE) (see equation (6)), mean absolute percentage error (MAPE) (see equation (7)) and mean algebraic percentage error (MALPE) (see equation (8)), were employed in this paper:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{A}_i - A_i|, \tag{6}$$

$$\text{MAPE} = 100\% \times \frac{1}{n}\sum_{i=1}^{n} \left|\frac{\hat{A}_i - A_i}{A_i}\right| \tag{7}$$

and

$$\text{MALPE} = 100\% \times \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{A}_i - A_i}{A_i}, \tag{8}$$

where $\hat{A}_i$ is the estimated value for a variable, A_i is the actual value for that variable and n is the total number of samples (e.g. census blocks). For these three measurements, the MAE is popularly used in remote-sensing and GIS studies, while the MAPE and MALPE are widely employed in demographic research. Both the MAE and MAPE are measures of precision, reflecting how close the estimated values are to the actual values, while the MALPE is a measure of bias, focusing on whether the total estimate shows an upward or downward tendency (Smith *et al.* 2002, Cai 2007).

4. Results and discussion

4.1 *HU estimation*

Following the methodology described in §3.1, the simple step-down interpolation technique and the regression-based approach were implemented in the study area. For the step-down interpolation, the geographical area of each census block is used as the weight for new HU assignment; that is, the number of new HUs received by a block is a linear function of its geographical area. The MAE of the HU estimation with this approach is 8.70 and a comparison with the actual new HU distribution (see figure 3(*a*)) illustrates that the HUs in a majority of census blocks are overestimated, in particular, the blocks with larger geographical areas (e.g. blocks in the village of Grafton). Comparatively, the HUs in a few blocks at the edge of Grafton town are underestimated. This is because recent developments of HUs were along the edge of Grafton town, while the step-down technique simply assumes that the number of new HUs is linearly correlated to block size, which does not reflect the actual development pattern of new HUs.

In addition to the simple step-down interpolation technique, the regression model with remote-sensing and GIS information was also implemented. Results of this model (see table 1) indicate that the spatial pattern of new HUs can be reasonably explained by the GIS and remote-sensing variables, with the adjusted $R^2 = 0.526$, where adjusted R^2 indicates a goodness-of-fit measure of the regression model. Among the three variables, the change of single-family and multi-family areas are statistically significant at the 95% confidence level, while the change of NDVI, although negatively correlated with new housing development, does not have significant contributions. Therefore, for the HU interpolation, only single-family and multi-family residential area changes were used.

Results show that the MAE with this regression model is 3.37, which is much lower than that obtained from the simple step-down interpolation method (8.70). Moreover, when compared with the actual HU map (see figures 3(*b*) and 3(*c*)), it is apparent that the spatial pattern of HU estimates from regression analysis can accurately reflect the actual distribution of new housing development. In particular, the HU development in the town of Grafton has been almost saturated, and most of the new developments are in the fringe of the town of Grafton, with a few new developments in the village of Grafton.

4.2 *PPH estimation*

Four methods for estimating PPH, including the current method (PPH is assumed to be unchanged from the previous census) and three regression models (models A, B and C) were applied for the study area, and the results of these three regression models are reported in table 2. Table 2 indicates that the value of PPH can be modelled reasonably well by demographic and spatial variables. In fact, the adjusted R^2 values

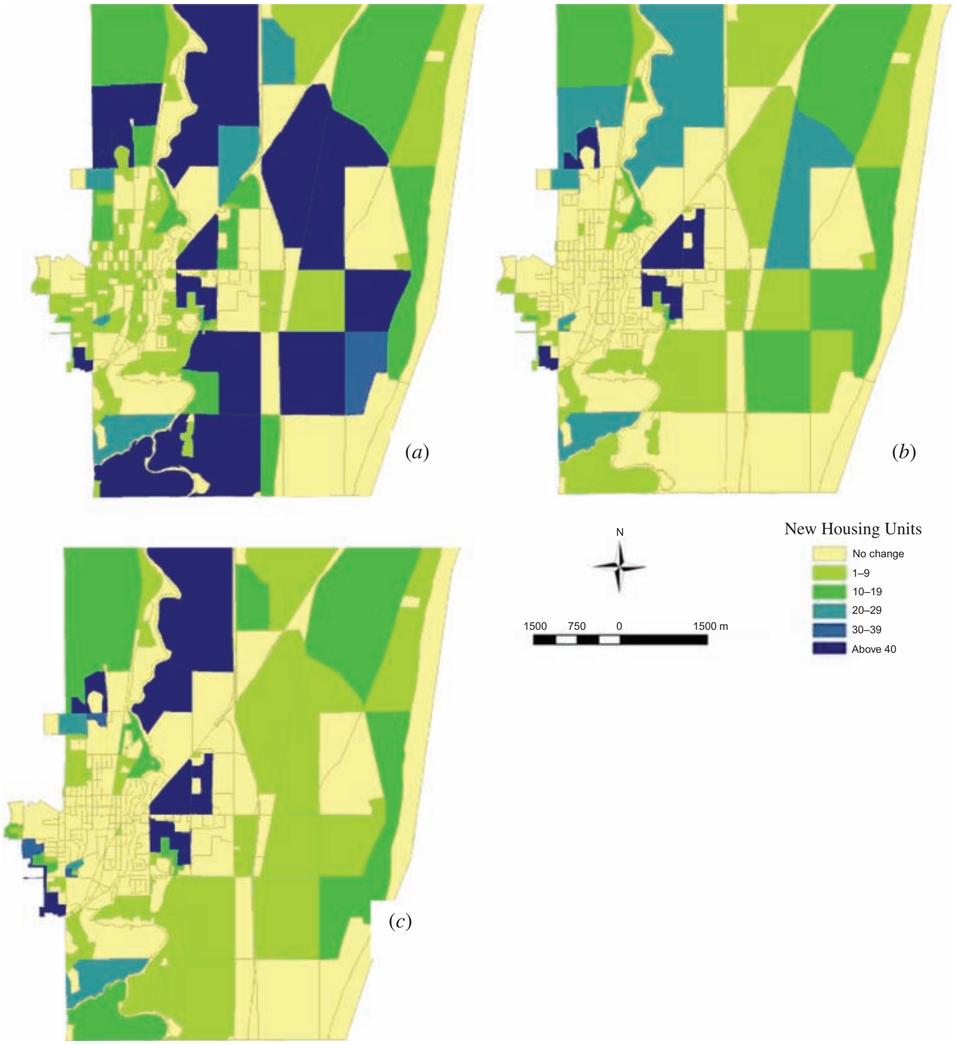


Figure 3. Comparison of HU estimates: (*a*) estimates using the simple step-down interpolation method, (*b*) estimates using the regression model and (*c*) actual HU numbers.

Table 1. HU regression-model results. ΔR_M is the change of multi-family land-use areas, ΔR_S is the change of single-family land-use areas, ** indicates statistically significant at the 95% confidence level, adjusted R^2 is 0.526, *B* stands for the coefficients for the regression model, *t* denotes the critical value in *t*-distribution, p-value represents a threshold probability to determine whether the observed outcome is statistically significant.

Coefficient	*B*	*t*	*p* value
Intercept	1.411	2.027	0.047**
ΔNDVI	–6.744	–1.400	0.163
ΔR_M	18.804	9.818	0.000**
ΔR_S	1.378	6.405	0.000**

Table 2. Coefficient summary of PPH regression models. '/' indicates the variable does not have significant contribution (at the 95% confidence level) to the regression model. PPH_90 denotes person per household in 1990.

Coefficients	Model A	Model B	Model C
Intercept	1.646	1.012	–0.215
PPH_90	/	0.208	/
Age 0_17	4.334	3.391	3.681
Age 18_39	/	/	/
Age 40_64	/	/	/
Age 65above	/	/	–0.855
Dist_Commercial	0.349	0.361	/
Dist_School	/	/	/
Dist_Recreation	–0.96	/	–0.346
R^2	0.627	0.718	0.698

for all these three models are over 0.60, indicating that over 60% of the variations in PPH can be explained. Four demographic variables, percentage of people with ages under 17 (Age 0_17), 18–39 (Age 18_39), 40–64 (Age 40_64) and 65 and over (Age 65above), were used for PPH estimates. Results indicate that the PPH has a significant and positive relationship with the percentage of young population (people with age under 17) and a significant but negative relationship with the percentage of elder population (people with age of 65 and over). These results are not unexpected and are consistent with the findings of Smith *et al.* (2002). Beside demographic variables, several spatial variables, including distance to commercial centres (Dist_Commercial), distance to schools (Dist_School) and distance to recreational areas (Dist_Recreation), were calculated using GIS land-use data. Results indicate that the PPH is positively correlated with the distance to commercial centres, but negatively correlated with the distance to recreational areas. These results imply that households with a larger size tend to choose residential locations far away from commercial centres and close to recreational areas. Spatial variables derived from Landsat data, including NDVI and textural parameters, however, were insignificant in any model. Therefore, these variables were dropped from these regression models.

The results of PPH estimates are illustrated in figure 4, with figures 4(*a*), 4(*b*), 4(*c*) and 4(*d*) showing the estimation results of models A, B, C and the current method, respectively, and figure 4(*e*) displaying the actual spatial distribution of PPH in 2000 for comparison purposes. It can be discerned that the PPH estimates from models A, B and C have similar spatial patterns when compared to the actual PPH values in 2000. Comparatively, the estimates of models A and B seem to be consistent with the spatial patterns of the actual PPH distribution, and model C clearly over-estimates the PPH in many census blocks. For the current method, though, the spatial trend is inconsistent with the actual PPH distribution, and the PPH values in a large number of blocks are visibly over-estimated, particularly in the blocks in Grafton village. The reason of this over-estimation may be that the PPH in the study area has had a downward trend in the past decades due to the declining birth rates and the tendency for adults to lead separate households.

In order to quantitatively evaluate the accuracy of the PPH estimates, the three accuracy measurements, MAPE, MALPE and MAE, were calculated and are reported in table 3. Results indicate that the current method has the lowest precision, with the highest MAPE (24.19%) and MAE (0.62). Moreover, it has the largest bias

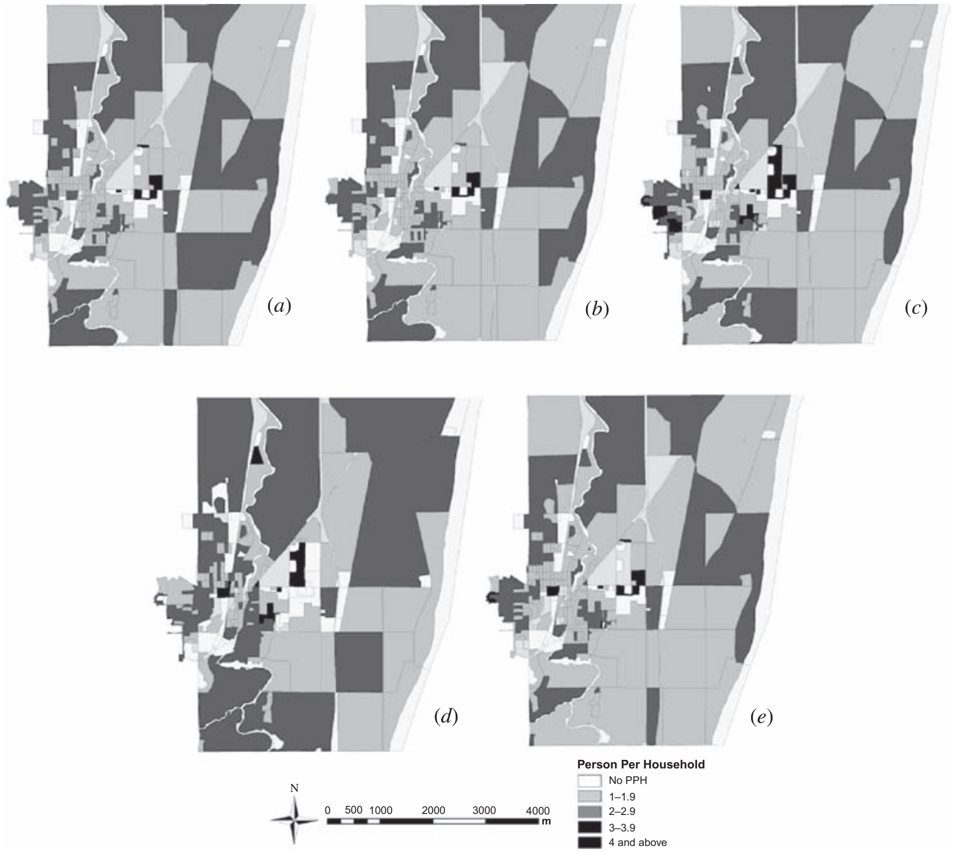


Figure 4. Comparison of persons per household (PPH) estimates among (*a*) model A, (*b*) model B, (*c*) model C, (*d*) current method and (*e*) actual PPH values.

Table 3. Accuracy comparison of estimation models of HU, PPH and population at the census-block level.

	Model	MAPE (%)	MALPE (%)	MAE
HU	Step-down method	41.08	19.26	8.70
	HU regression	35.34	6.82	3.37
PPH	Current method	24.19	13.13	0.62
	Model A	22.22	4.21	0.59
	Model B	9.98	0.21	0.27
	Model C	19.19	7.77	0.50
Population	Simple demographic method	58.27	32.68	30.02
	Regression model	25.56	5.52	11.68

and over-estimates the PPH by approximately 13%. This proves that it is inappropriate to assume that the PPH is constant over time. Comparatively, all three regression models perform better than the current method. Models A and C have slightly better precisions, but much lower bias than the current method. For example, the MALPEs of models A and C are 4.21% and 7.77%, much lower than that of the current method

(13.13%). Model B has the best overall performance. In particular, it has the highest precision, with the lowest MAPE (9.98%) and MAE (0.27) and the smallest bias (MALPE = 0.21). The lack of precision improvements with models A and C may be associated with the assumptions of these two models. In particular, model A, as an 'inherit model', assumes that the relationship between PPH and relevant variables does not change over time. Model C, as a 'change model', considers that the change of PPH can be effectively explained by the changes of spatial and demographic variables, irrelevant to the previous PPH values. These assumptions, however, are problematic, as the PPH in a census block is dependent on both the previous PPH values and the changes of spatial and demographic variables. Therefore, model B, taking both factors into account, has proven to be the most accurate model for PPH estimation.

4.3 *Small-area population estimation*

With the HU and PPH estimates for a non-census year, it is necessary to generate the population estimates using the HU method described in §3.3. In this paper, two approaches were developed for comparison. The first one is the simple demographic approach, which uses the step-down interpolation method for HU estimation and assumes that the PPH is the same as that obtained from the previous census. The second approach involves a sequence of regression analyses with demographic and spatial data. In particular, the HU estimates were achieved from the regression model using detailed land-use datasets (as described in §3.1) and the PPH estimates were generated using model B, in which demographic and spatial variables were used (as detailed in §3.2).

Results of population estimates with these two approaches, together with the actual population count from 2000 census, are displayed in figure 5. In particular, figure 5(*a*) shows the results obtained from the simple demographic approach and 5(*b*) illustrates the estimates from the regression models. In addition, detailed accuracy assessments of these population estimates are reported in table 3. Analysis of results indicates that the estimates from the simple demographic method are not acceptable. In general, it over-estimates the population for the whole study area by over 30%, largely due to the over-estimates of the PPH values. Moreover, the relative errors are also very large, as the MAPE is approximately 58% and the MAE is about 30. On the contrary, the regression-based approach only slightly over-estimates the overall population (5.5%) and is much more precise than the simple demographic approach. In fact, the MAPE (25.56%) and MAE (11.68) indicate that the regression models with demographic and spatial variables can derive small-area population estimates reasonably well.

Besides an overall accuracy assessment, analysis of relative errors for the HU, PPH and population estimates (see figure 6) illustrates the patterns of error propagation. Figure 6(*a*) shows that the HU values in seven blocks (out of 313) have been highly over-estimated (with relative errors higher than 15%). Within these seven blocks, six have highly over-estimated population counts (see figure 6(*c*)). In addition, the HU values in 19 blocks are highly under-estimated (with relative errors lower than 15%) and with them, 15 have highly under-estimated population counts. These results indicate that the errors of HU estimates have significant influences on those of population estimates. When compared to the HU, the errors of PPH do not have such strong influences on the errors of population estimates. In particular, within the 12 blocks with highly over-estimated PPH values, the population counts within nine of them are highly over-estimated. Moreover, within the 10 blocks with highly under-estimated PPH values, only four of them have highly under-estimated population counts.

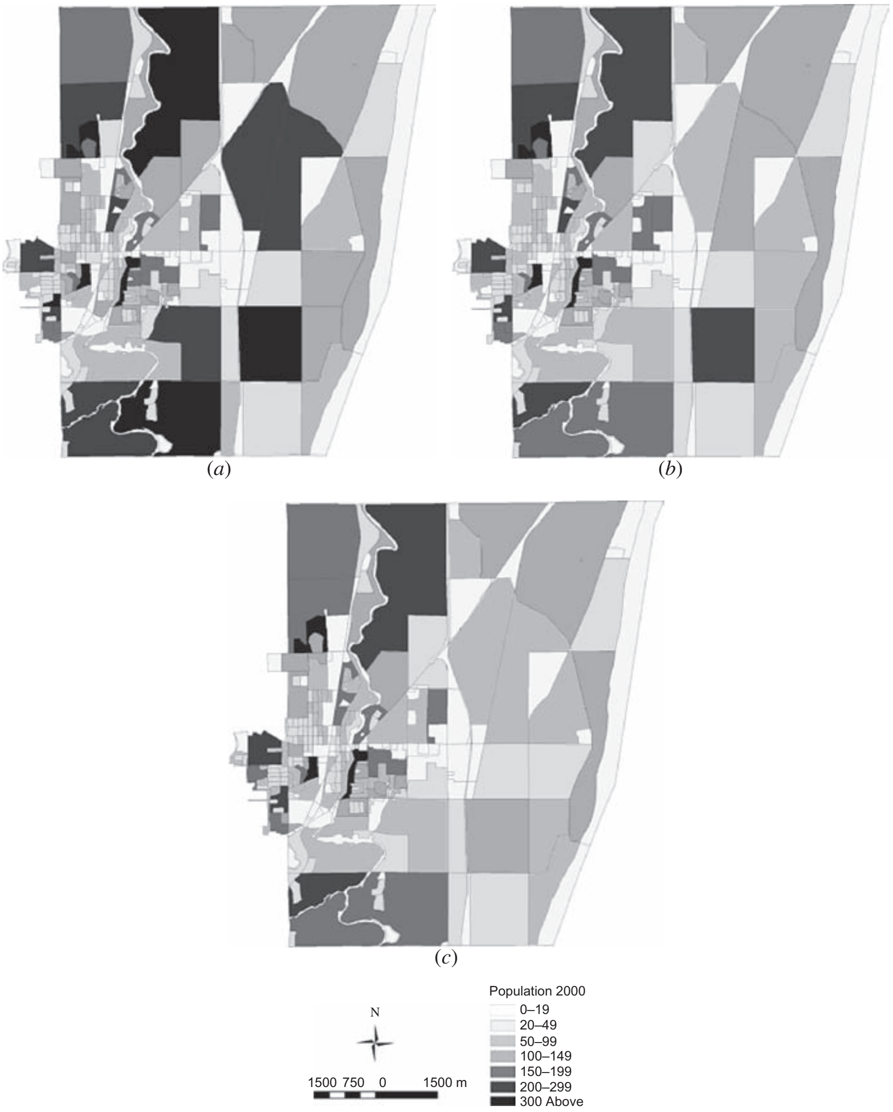
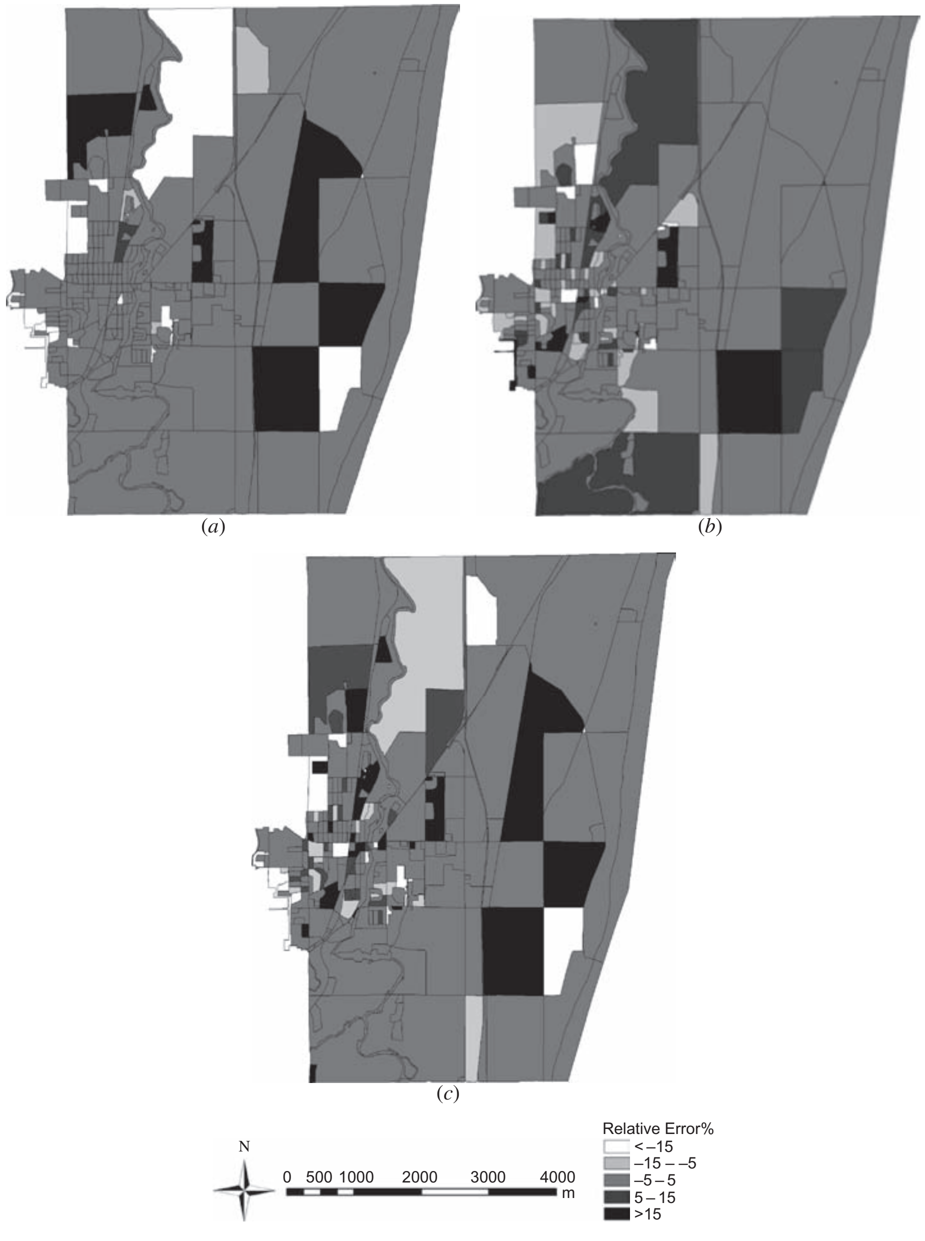


Figure 5. Comparison of block-level population estimates derived from (*a*) the simple demographic method, (*b*) regression analyses with demographic and remote-sensing/GIS data and (*c*) actual population count from the 2000 census.

5. Conclusions

In this paper, we proposed to integrate GIS and remote-sensing techniques into the HU method for generating better small-area population estimates. In particular, HUs and PPH at the census-block level were derived using regression models with demographic and spatial variables. Then these estimates were input to the HU method for deriving block-level population estimates. The accuracy of small-area estimation with this regression-based method was compared to the current method.

Analysis of results suggests two major conclusions. Firstly, the accuracy of small-area population estimates can be significantly improved through integrating


(*a*)

(*b*)

(*c*)

Figure 6. Comparison of relative errors at the block level for (*a*) estimated HUs using the regression method, (*b*) PPH using regression model B and (*c*) population using equation (5) with the estimated HUs from (*a*) and PPH from (*b*).

remote-sensing/GIS information. Detailed land-use information has proven to be the most important GIS dataset for small-area population estimation. In particular, it can be used to redistribute aggregated building-permit information for better HU generation and employed to calculate spatial variables for better PPH estimation. Secondly, this research proves that the PPH can be effectively modelled by demographic and spatial variables. In fact, several demographic variables, including the percentage of population with age under 17 and population with age 65 and over, and several spatial variables, such as the distance to commercial centres, schools and recreational areas, can explain over 60% of PPH variations.

Although this study showed that the integration of GIS and remote-sensing information into the HU method can greatly improve the small-area population estimates, there are many issues for future research. One direction is for better HU estimation. Detailed and more accurate HU estimation can be achieved using high-spatial-resolution remote-sensing imagery (such as IKONOS and QuickBird data) and Light Detection and Ranging (LiDAR) datasets. Moreover, the footprint and volume information generated from these data can help the estimation of PPH. Another direction is to explore whether demographic variables can be derived through GIS and remote-sensing information. Currently, a special census, or sampling, is needed for creating demographic variables. This work, however, is always time consuming and labour intensive.

Acknowledgements

This research was supported by the United States National Science Foundation grants BCS-0822155 and BCS-0822489. We are grateful to the anonymous reviewers and the Editor-in-Chief, Giles Foody, for their suggestions on an earlier draft of this manuscript.

References

Bongaarts, J., 2001, Household size and complexity in the developing world in the 1990s. *Population Studies*, **55**, pp. 263–79.

Cai, Q., 2007, New techniques in small area population estimates by demographic characteristics. *Population Research and Policy Review*, **26**, pp. 203–218

Chen, K., 2002, An approach to linking remotely sensed data and areal census data. *International Journal of Remote Sensing*, **23**, pp. 37–48.

Ghosh, M. and Rao, J.N.K., 1994, Small area estimation: an appraisal. *Statistical Science*, **9**, pp. 55–76.

Goodchild, M.F. and Lam, N.N.-S., 1980, Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing*, **1**, pp. 297–312.

Harvey, J.T., 2002a, Estimating census district populations from satellite imagery: some approaches and limitations. *International Journal of Remote Sensing*, **23**, pp. 2071–2095.

Harvey, J.T., 2002b, Population estimation models based on individual TM pixels. *Photogrammetric Engineering and Remote Sensing*, **68**, pp. 1181–92.

Kobrin, F., 1976, The fall in household size and the rise of the primary individual in the United States. *Demography*, **13**, pp.127–38.

Li, G. and Weng, Q., 2005, Using Landsat ETM+ imagery to measure population density in Indianapolis, Indiana, USA. *Photogrammetric Engineering and Remote Sensing*, **71**, pp. 947–958.

Lo, C.P, 1986a, Accuracy of population estimation from medium-scale aerial photography. *Photogrammetric Engineering and Remote Sensing*, **52**, pp. 1859–1869.

Lo, C.P. (Ed.), 1986b, *Applied Remote Sensing* (London, UK: Longman).

Lo, C.P., 1995, Automated population and dwelling unit estimation from high-resolution satellite images: a GIS approach. *International Journal of Remote Sensing*, **16**, pp. 17–34.

Martin, D. and Williams, H.C.W.L., 1992, Market-area analysis and accessibility to primary health-care centers. *Environment and Planning A*, **24**, pp. 1009–1019.

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

Mennis, J. and Hultgren, T., 2006, Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, **33**, pp. 179–194.

Perry, M. and Voss, P., 1996, Using geo-demographic methods for improving small area estimates. Centre for Demography and Ecology at the University of Wisconsin-Madison, Working Paper No. 96–15.

Plane, D.A. and Rogerson, P.A. (Eds), 1994, *The Geographical Analysis of Population with Applications to Business and Planning* (New York, NY: John Wiley).

Qiu, F., Woller, K.L. and Briggs, R., 2003, Modelling urban population growth from remotely sensed imagery and TIGER GIS road data. *Photogrammetric Engineering and Remote Sensing*, **69**, pp. 1031–42.

Rees, P., Norman, P. and Brown, D.G., 2004, A framework for progressively improving small area population estimates. *Journal of the Royal Statistical Society Series A*, **167**, pp. 5–36.

Richter, R., 1996a, A spatially adaptive fast atmospheric correction algorithm. *International Journal of Remote Sensing*, **17**, pp. 1201–1214.

Richter, R., 1996b, Atmospheric correction of satellite data with haze removal including a haze/clear transition region. *Computers & Geosciences*, **22**, pp. 675–681.

Richter, R., 2005, Atmospheric/topographic correction for satellite imagery. DLR report DLR-IB 565-01/05, Wessling, Germany.

Smith, S.K., 1986, A review and evaluation of the housing unit method of population estimation. *Journal of the American Statistical Association*, **81**, pp. 287–296.

Smith, S.K. and Cody, S., 2004, An evaluation of population estimates in Florida: April 1, 2000. *Population Research and Policy Review*, **23**, pp. 1–24.

Smith, S.K. and Lewis, B., 1980, Some new techniques for applying the housing unit method of local population estimation. *Demography* **17**, pp. 323–339.

Smith, S.K. and Mandell, M., 1984, A comparison of population estimation methods: housing unit versus component II, ratio correlation and administrative records. *Journal of the American Statistical Association*, **79**, pp. 282–289.

Smith, S.K., Nogle, J. and Cody S., 2002, A regression approach to estimating the average number of persons per household. *Demography*, **39**, pp. 697–712.

Starsinic, D.E. and Zitter, M., 1968, Accuracy of the housing unit method in preparing population estimates for cities. *Demography*, **5**, pp. 475–484.

Us Census Bureau, 1998, Subcounty population estimates methodology. Available online at http://www.census.gov/population/methods/e98scdoc.txt (accessed 17 May 2006).

Us Census Bureau, 2001, Profile of general demographic characteristics. Table DP-1 in *2000 Census of Population and Housing* (Washington, DC: US Census Bureau).

Us Census Bureau, 2005, Methodology: 2004 estimates and projections area documentation, subcounty total population estimates. Available online at http://www.census.gov/popest/topics/methodology/2004_su_meth.html (accessed 18 May 2006).

Webster, C.J., 1996, Population and dwelling unit estimates from space. *Third World Planning Review*, **18**, pp. 155–176.

Wu, C. and Murray, A.T., 2005, A cokriging method for estimating population density in urban areas. *Computers, Environment and Urban Systems*, **29**, pp. 558–579.

Wu, C. and Murray, A.T., 2007, Population estimation using Landsat Enhanced Thematic Mapper Imagery. *Geographical Analysis*, **39**, pp. 26–43.

Wu, S.S, Qiu, X. and Wang, L., 2005, Population estimation methods in GIS and remote sensing: a review. *GIScience and Remote Sensing*, **42**, pp. 80–96.