Science of the Total Environment 408 (2010) 4807–4816

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Multi-layer multi-class dasymetric mapping to estimate population distribution

Ming-Dawa Su [a], Mei-Chun Lin [a], Hsin-I Hsieh [a], Bor-Wen Tsai [b], Chun-Hung Lin [a,*]

[a] Dept. of Bioenvironmental Systems Engineering, National Taiwan University, Taipei, Taiwan
[b] Dept. of Geography, National Taiwan University, National Taiwan University, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 23 February 2010
Received in revised form 18 May 2010
Accepted 17 June 2010

Keywords:
Choropleth maps
Spatial distributions
Dasymetric mapping
Multi-layer multi-class dasymetric

ABSTRACT

The spatial patterns of population distribution are very important information for most regional planning and management decisions. But the socioeconomic data are usually published in areal aggregated format due to privacy concerns. Although choropleth maps are used extensively to display spatial distributions of these areal aggregated data, patterns may be distorted due to assumptions of homogeneous distributions and the modifiable areal unit problem. Most human activity, including population distribution, is spatially heterogeneous due to variations in topography and regional development. A multi-layer multi-class dasymetric (MLMCD) framework was proposed in this study to better redistribute the regionally aggregated population statistics into smaller areal units and reveal more realistic spatial population distribution pattern. The Taipei metropolitan area in Taiwan was used as a case study area to demonstrate the disaggregation ability of the proposed framework and the improvements to the traditional binary or multi-class dasymetric method. Assorted data, including remote sensing images, land use zoning, topography, transportation and accessibility to facilities were introduced in different layers to improve the redistribution of aggregated regional population data. The concept of multi-layer multi-class dasymetric modeling is both useful and flexible. Different levels of accuracy in this population redistribution process can be achieved depending on data and budget availabilities and the needs for different data usage purposes.

1. Introduction

Estimates of population size and distribution are vital for socioeconomic planning and management decisions such as the allocation of food and medical supplies, transportation, land use and regional development. A better understanding of the population distribution will lead to more effective management policies. Censuses are commonly used as the major source for population distribution information. Although detailed data may be recorded for each person in a census, privacy concerns prevent these data from being released. The census results are usually published in aggregate form. Data are aggregated by some spatial unit such as block group, census tract, grid square statistic and zip code area, or by administration unit such as township, county or even state (Directorate General of Budget, Accounting & Statistics, 2000; Peters and MacDonalds, 2004; Statistics Bureau of Japan, 2005; US Census Bureau, 2005).

Spatial heterogeneity in population distribution exists between natural environments and regional developments. These spatial patterns play important roles in regional decisions, resource allocation, infrastructure planning and disaster mitigation. Because homogeneity is assumed for data released in spatially aggregated forms, the spatial patterns of population distribution within the aggregated units may be lost or distorted (Fisher and Langford, 1996; Weichselbaum et al., 2005).

Gridded Population of the World, version 3(GPW3) (Center for International Earth Science Information Network, 2005) is a dataset for human population in the common geo-referenced framework displaying the human population distribution at a global scale. GPW3 was constructed from population data aggregated by assorted geopolitical spatial units worldwide. These aggregated data were disaggregated into geo-referenced grids at a resolution of 2.5′ (about 4.5 km) for use in social, economic, earth science fields. Global Rural–Urban Mapping Project (GRUMP) was built on GPW3 by incorporating more detailed population distribution information in urban and rural settlements. The GRUMP dataset has a much higher resolution at 30″ (about 1 km). LandScan dataset (Dobson et al., 2000) is another worldwide population database compiled on a latitude/longitude grid system with a resolution of 30″. The census data were redistributed into each grid cell based on a likelihood coefficient ascertained from related information such as proximity of roads, slope, land cover and nighttime light emission. The LandScan dataset provides an estimate of the worldwide ambient population at risk. This ambient population information is valuable for the purpose of emergency response, as it integrates diurnal movements and traveling habits into the population distribution estimates. Although GPW3, GRUMP and LandScan datasets provide valuable global human population information that

* Corresponding author. Department of Bioenvironmental Systems Engineering, National Taiwan University, No.1, Sec. 4, Rd. Roosevelt, Taipei City, Taiwan 10617. Tel.: +886 2 3366 3451; fax: +886 2 2363 5854.
E-mail address: jiunhong.lin@gmail.com (C.-H. Lin).

0048-9697/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2010.06.032

improves decisions about global/international resource planning and risk management, more detailed population distributions are necessary for regional management issues such as diffusion control during a disease outbreak or hazardous waste release analysis. This study aimed to establish a multi-layer multi-class dasymetric (MLMCD) model to disaggregate the aggregated population data into smaller spatial units and reconstruct the spatial distribution patterns of regional populations.

2. Pattern distortion in aggregated data

Point data are the most realistic representations of population distributions and are capable of revealing the near-true spatial patterns of regional populations, but these individual data are usually confidential. Choropleth maps are commonly employed to display spatial distribution patterns using aggregated population data. Spatial patterns may be lost or distorted during these spatial aggregations for different zones or scales. The modifiable areal unit problem (MAUP) is a potential source of error that affects spatial studies using aggregated data (Unwin, 1996). This problem was first addressed by Openshaw in 1984. When data with spatial variability are aggregated, the original underlying spatial patterns may be distorted by the choice of district boundaries. This problem is especially crucial in choropleth mapping. Applications such as spatial planning, demography, crime and disease mapping are prone to such errors. The MAUP is also closely related to ecological fallacy with the false assumption of homogeneity in aggregated data.

The areal units are modifiable and can be aggregated into different partition sizes (such as census tracts, counties, or postal code zone). Although these spatial partitions are comparable in size, they can be very different from each other. For example, the epidemic situation in an area with very high disease incidence may be overlooked if adjacent districts with lower case incidence are aggregated together. Although the use of smaller spatial aggregation units may alleviate the MAUP, this problem is not completely circumvented by decreasing the aggregation area. This study attempted to mitigate MAUP by reconstructing more accurate spatial distribution of the aggregated data with the aid of ancillary data such as satellite images, terrain, land use and infrastructure, as well as public works such as traffic networks.

3. Population redistribution models

Although the MAUP is a well known phenomenon, widely available aggregated data are often used inappropriately and can yield misleading information. This is commonly observed in choropleth mapping practices, which are convenient due to the ease of mapping using Geographic Information Systems (GIS).

Some methods have been developed for redistributing these aggregated data into smaller spatial units. Areal interpolation is commonly used for disaggregating population. Areal interpolation is the process of estimating the population in a set of target polygons based on known populations that exist in a set of source polygons. The need for areal interpolation arises when data from different sources are collected in different areal units (Flowerdew and Green, 1992).

Binary dasymetric method is more or less the same as the areal weighting method but adding an extra step of filtering the data using an ancillary data set (Flowerdew and Green, 1989; Holt et al., 2004). The binary dasymetric approach defines the target area as either occupied (or populated) or unoccupied (or unpopulated) with the help of related information such as areal photographs, remote sensing images, or road buffer zones. A weighting factor of 1 is used for the populated areas and 0 for the unpopulated areas. The aggregated population is then uniformly distributed into the populated areas. Since the unpopulated areas (such as lakes, rivers or paddy fields) are excluded from the population redistribution, the results demonstrate a more accurate spatial distribution pattern of the population (Holt et al., 2004; Keping et al., 2004; Langford and Higgs, 2006; Langford and Unwin, 1994).

The multi-class weighted dasymetric model is an improved version of the binary dasymetric model. Populated areas are subdivided into additional subcategories that reflect different population densities based on information such as land use, zoning, land value, accessibility, infrastructure density, home living style, etc. Different weighting factors are applied to each category to produce a more realistic population distribution. For example, multiple family and single family zones in a residential area may have different population densities. A stronger weight is applied for multiple families because of its higher population density (Flowerdew and Green, 1992; Reibel and Bufalino, 2005; Wu et al., 2005; Wu, 2006). Contrary to the weighting factors of 0 and 1 in the binary dasymetric model, scaled weighting factors (between 0 and 1) are applied for different sub-classes in this multi-class weighted algorithm as shown in Eq. (1).(Goodchild et al., 1993; Holt et al., 2004; Langford, 2006; Relbel and Agrawal, 2005)

$$D_{ij} = \frac{P_i \times (A_{ij} W_j)}{\sum_{j=1}^{m} A_{ij} W_j} \quad (1)$$

where i and j are subscripts for areal units and sub-classes respectively, m is the number of the sub-classes, D_{ij} is the population density, P_i is the population in areal unit i, W_j is the weighting factor for subclass j and A_{ij} is the area of subclass j in unit i.

This method assumes that each class has a characteristic population density. The weighting factors W_j represent the population distribution characteristic of each class and may vary depending on the location of the area of interest and the assignment may be subjective. The calibration of these W_j parameters becomes a major problem in the application of the multi-class dasymetric model. This was not an issue for binary dasymetric model as 0 and 1 can be assigned without doubt. Although there are some alternatives proposed such as subjective choice based on local or prior knowledge (Eicher and Brewer, 2001), empirical estimation derived by sampling a subset from the study region (Mennis and Hultgren, 2006; Langford, 2006), or through the use of geo-statistical modeling (Lo, 2008), this is still a controversial issue waiting for more researches and case studies to resolve.

Another weakness of the multi-class dasymetric method is that although the differences between classes are recognized, the differences within a specific class are ignored (Eicher and Brewer, 2001). For example, the population density of the single family class may not be uniform and may vary from one part of the town to the other. This spatial non-stationary characteristic of population may be examined in more details with regional regression approach like Geographically Weighted Regression (GWR, Fotheringham et al., 1998, 2000) for better population estimations. (Lo, 2008)

Rather than using categorical land use as a proxy for population density, very high spatial resolution satellite images were used to estimate population based on image texture. Spatial units called "Homogeneous Urban Patches" (HUP) are obtained through texture-based image segmentation (Liu et al., 2006). The correlation between census population density and image texture was established, but it was not good enough to provide reliable estimates of population distribution.

Surface-generating methods are also used to model a population surface by kernel density estimation (Bracken and Martin, 1989; Martin et al., 2000; Martin, 2006). They created a population density surface by interpolating based on the population-weighted centroid. The early version of this method had problem maintaining the correct population counts in the original spatial units (Bracken and Martin, 1989), but was later revised to preserve the pycnophylactic (volume-preserving) property (Martin, 2006; Rase, 2001).

Some planning or management issues at the local level may require very detailed population distribution to yield meaningful results. Most of the methods described above mainly relied on land use/land cover data and may not be as fine-grained as necessary for many urban analysis purposes.

Some researcher used the street network data (Reibel and Bufalino, 2005), urbanization index (Mennis, 2003; Mennis and Hultgren, 2006) or cadastral data (Maantay et al., 2007) to derive weights for population interpolation. Using street network as a proxy for population distribution was reported effective "where the lack of population is reflected in the lack of roads and least in those areas (such as industrial areas) with a more developed, but non-residential transportation infrastructure" (Reibel and Bufalino, 2005).

As described above, most of the population redistribution algorithms require ancillary information. The availability and quality of this ancillary information have critical effects on the redistribution calculations. Most of the methods are either too simple like the areal interpolation or binary dasymetric, or require too much detailed data that may not be available in most of the places such as the kernel density estimation based on population-weighted centroid. A multi-layer and multi-class dasymetric algorithm for population distribution was established in this study in order to couple the availability of data from different sources.

4. Methodology

Although the true population distribution is best represented by individual points as shown in Fig. 1(b), these data are generally confidential and unavailable. The population data are usually provided as the total for a region (Fig. 1(a)) or as sums of different zoning units (Fig. 1(c)). An example of improved population distribution reconstruction of an aggregated regional population by the dasymetric model is shown in Fig. 1. The total regional population of 100 in Fig. 1(a) was used as the starting point for this example. The true populations in each spatial unit are shown in Fig. 1(c).

To reconstruct the population distribution patterns, disaggregation algorithms must be used to reallocate the total population into different subunits. A 20 m by 20 m grid system was constructed for this population distribution process. The aggregated population was distributed into each cell to reveal the spatial population distribution. For simple areal weighting, homogeneity was assumed with a uniform population density applied to the whole region, and each cell received same population amount. The results are shown in Fig. 1(d) with uniform population density in all cells. The population in each of the four subunits is equal to population sum of all the cells within that subunit. The numbers in the parentheses are the errors or deviations of the assigned population from the true values as shown in Fig. 1(c).

Dasymetric mapping is thought to better capture the true spatial distribution pattern by identifying the different spatial characteristics in the region. For example, the whole region can be classified as either populated or unpopulated as shown in Fig. 1(e). This binary classification can be accomplished with the aid of remote sensing images or areal photographs that are easily accessible. The total population was then redistributed to the populated area by the same areal weighting method and the results are shown in Fig. 1(e). Some errors in Fig. 1(e) may be worse than those shown in Fig. 1(d). These deviations are most clearly observed in the northwest and southeast regions. These are the more rural and urban regions according to the true point population pattern shown in Fig. 1(b).

Redistributed populations in each spatial unit are usually overestimated in the rural (more sparse) regions and underestimated in the urban (more dense) regions. Urban regions usually have higher population densities because there are more socioeconomic activities, public infrastructure, employment opportunities and higher land asset values.

If more relevant information is available for the region, then the populated areas can be further divided into zones with different population densities. This classification can be determined based on data of land cover and land use, topography, land value, or the density of public infrastructures such as traffic networks. As shown in Fig. 1(f), the populated area can be classified into urban, suburban and rural.

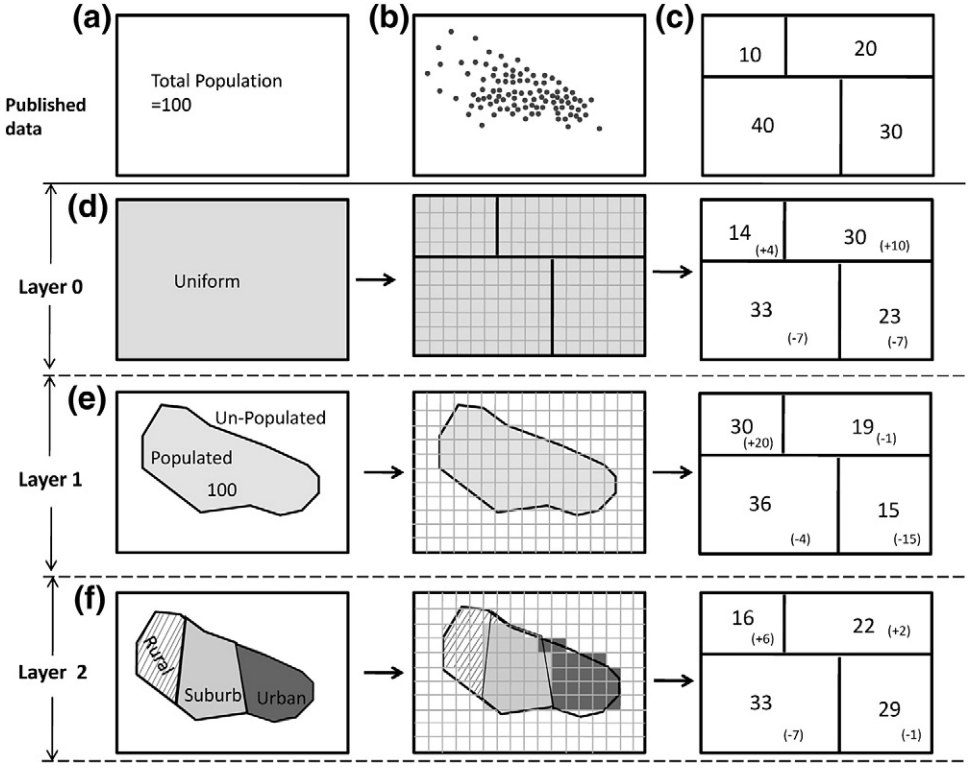

Fig. 1. Description of dasymetric models.

The total population is then distributed to each grid cell using Eq. (1) with different weighting factors, e.g., 0.5, 0.3 and 0.2 for urban, suburban and rural, respectively. The effectiveness in capturing the population distribution pattern can be evaluated through the errors in each zone. The results in Fig. 1(f) show less error and better represent the initial population distribution pattern than those in Fig. 1(d) and (e).

A multi-layer multi-class dasymetric framework, as shown in Fig. 2, was proposed in this study in order to better disaggregate the aggregated population data and elucidate true population distribution patterns according to several available parameters correlated with population density. Each layer in this model is composed of several classes of different population densities. Weighting factors are assigned to each class based on relative population density. Layer 0 in Fig. 2 assumes that the population is uniformly distributed over the region. Layer 1 is formed by classifying the regions into subclasses of populated and unpopulated, with weighting factors of 1 and 0, respectively. This layer represents the traditional binary dasymetric model. If more information is available and one of these classes can be further divided into more subclasses, another layer is formed. In the example shown in Fig. 2, the populated class of layer 1 can be further classified into 3 subclasses of urban, suburban and rural in layer 2. The population density within the suburban area can be differentiated by land use zoning data such as single/multiple family zones. The urban area can also be further classified as either commercial or high density residential zones. Each category in layer 3 can be further divided in higher layer according to more detailed information such as transportation network density, employment accessibility and land asset values.

This reversed tree type multi-layer framework represents a hierarchical relationship. Although more data are required to create additional layers for this framework, the population disaggregation at higher layers is more comprehensive and exhaustive and is expected to reveal more detailed spatial distribution patterns.

5. Study area

The Taipei metropolitan of Taiwan, shown in Fig. 3(a), was used as the study area to demonstrate the proposed multi-layer multi-class dasymetric framework. One third of the total national population, more than 6 million people, is concentrated in an area of about 2700 km^2. The Digital Terrain Modeling (DTM) in Fig. 3(b) reveals that the central part of the region is flatter and surrounded by a mountainous area. The region consists of 41 cities and townships and is subdivided into more than 1400 administration units called "Li" (Fig. 3(c)). The average Li population density is about 3000 persons/km^2, but ranges between 3 persons/km^2 and 230,000 persons/km^2 (Department of Civil Affairs, Taipei City Government, 2007; Department of Urban Development, Taipei City Government, 2007).

Socioeconomic data related to the population distribution were collected for the study area including population and agricultural censuses, population registration, remote sensing images, land use zoning, land use surveys, land value and transportation networks. These data were used to identify the populated and unpopulated areas and to estimate the weighting factors as described in Eq. (1), thus properly redistributing the aggregated population data into sub-regions.

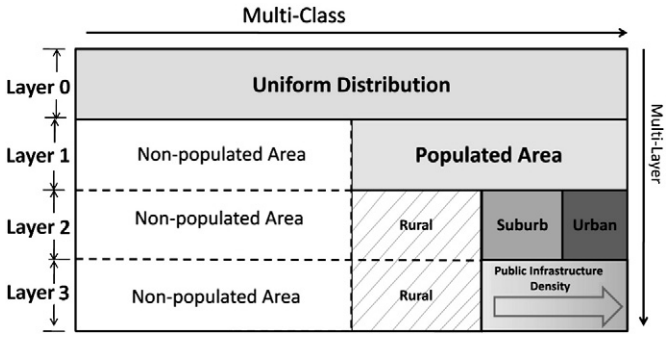

Fig. 2. Framework of multi-layer and multi-class dasymetric model.

The types of land use in the study area were roughly classified into eight categories: agricultural/forest, transportation, water conservancy, building, industrial, recreational, mining, and others (Ministry of the Interior, Taiwan, 1995). Currently, more than 80% of land is used for agricultural/forest. And the building areas represent less than 10% of the study area. While the major portion of the population is concentrated in the flat central area, a small portion of the population is scattered in the peripheral mountain-slope region. The concentration of traffic networks in the central part of the region, as shown in Fig. 3(d), also confirms this general distribution trend.

"Li" is the lowest administration level in Taiwan with an average area of about 4.6 km^2. The Li areas tend to be larger in the rural region than in the urban and suburban regions, as shown in Fig. 3(c). Most of the registered or census population data are published at this administration level. Since the true point population data are unavailable due to confidentiality reasons, the total populations at each of the smallest available areal units (Li) were collected as the true values and used as the basis for comparisons among different population redistribution results determined in this study.

These published aggregated Li population data were first aggregated for the entire study area as the initial condition for population disaggregation. The aggregated population was then redistributed into 20 m by 20 m grids to reconstruct the regional population distribution pattern using the proposed framework. The grid resolution of 20 m was chosen to match the resolution of SPOT images used in this study to identify populated areas.

6. Results and discussions

To begin disaggregation, layer 0 of the proposed multi-layer multi-class dasymetric framework was set as the uniformly distributed aggregated population across the entire study area. This population distribution, shown as the choropleth map in Fig. 4(a), is of little use for region planning and management decisions. The population density was 1.08 persons per cell.

Remote sensing images were used in layer 1 to classify the study area into populated and non-populated areas. SPOT images with a spatial resolution of 20 meters were used for this purpose. Since the population data used in this study represents the registered population who inhabit buildings, the building areas were treated as populated zones. A maximum likelihood classification method was used to identify the building areas using training samples chosen from a combination of multi-spectral images and colored ortho-images. The overall accuracy was higher than 95%. The results are shown in Fig. 4(b).

Although the population distribution pattern shown in Fig. 4(b) is more realistic than that of layer 0, this distribution did not discriminate based on the population densities of different land use types within the populated areas. There may be major differences in population density among different land uses. For example, the population density in a residential zone is expected to be higher than that in agricultural or industrial zones. These differences in population densities can be better captured if the land use data are incorporated into the population distribution model.

The land use zoning map was overlaid with the building area in layer 2 to better capture the population pattern varying with socioeconomic activities. The overlaying results are summarized in Table 1. Some building areas were located in traffic and water conservancy zones. These may have been mistakenly identified during satellite image processing or may represent illegal residential

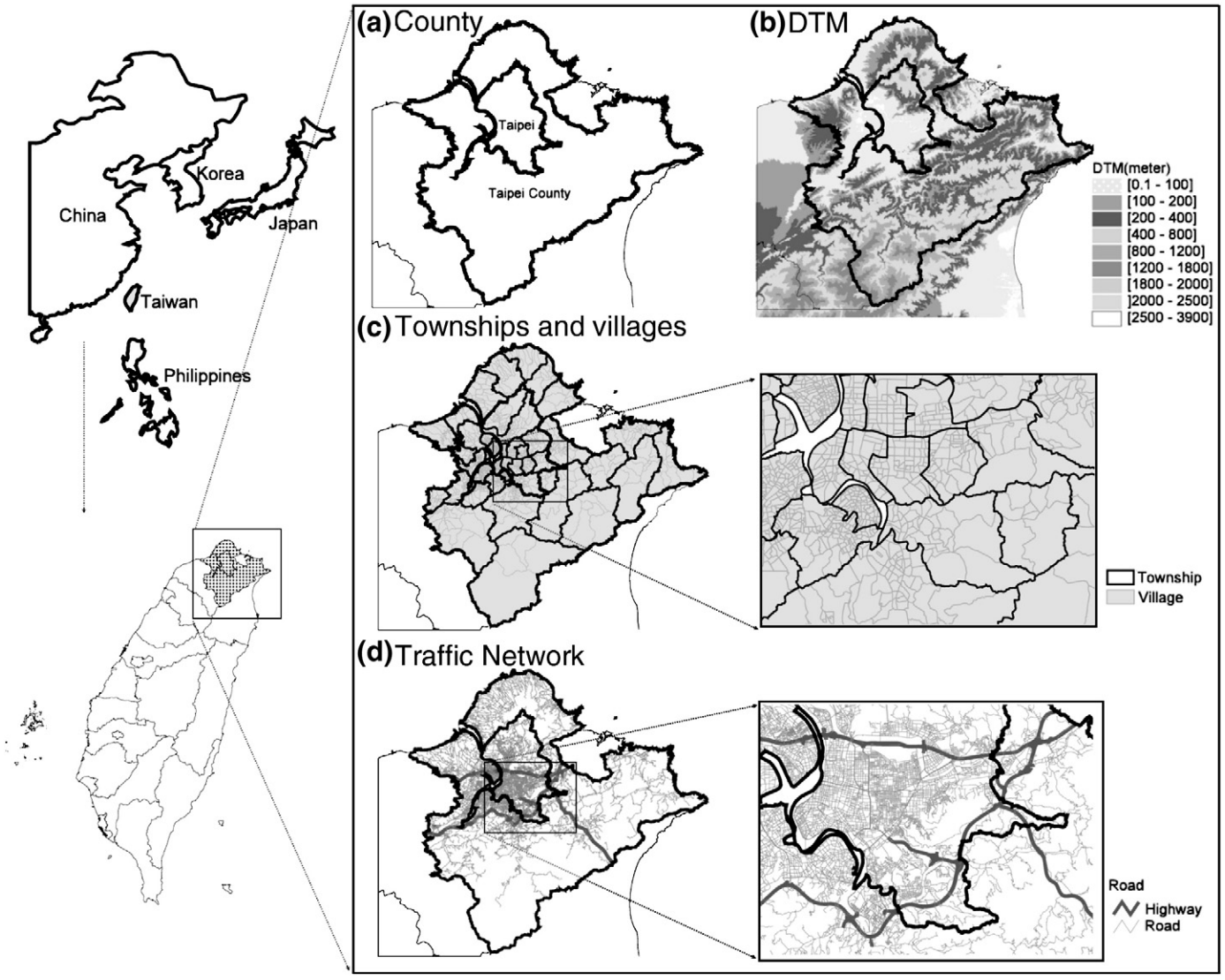

Fig. 3. Study area.

buildings where zoning regulation were not enforced. Buildings such as parking terraces, highway toll stations, pumping houses, or pavilions in the riverside parks may also be erroneously categorized as populated areas. These cells were reassigned as unpopulated areas to better capture the true population distribution.

The land use zones were aggregated into two categories of agricultural (including agriculture and forest) and non-agricultural (including residential, commercial and industrial) for this study. The population redistribution weighting factors for agricultural and non-agricultural zones were set as 0.01 and 0.99, respectively, based on the statistics from the 2005 agricultural census and the 2000 population census (DGBAS, 2007). The population distribution differences between agricultural and non-agricultural zones are captured in layer 2 of the disaggregation model, as shown in Fig. 4(c).

Population densities may not only be dissimilar in different land use zones but may also vary within a specific land use type. For example, the population density in an urban single family zone may be higher than that in the same zone in a suburb area. This variance can be captured by various socioeconomic factors such as land values, or by the density of utilities such as transportation networks. These discrepancies were not adequately represented in the disaggregation pattern of layer 2; therefore, the non-agricultural areas (including commercial, industrial and residential) were further classified into levels of utility accessibility. As shown in Fig. 3(d), the traffic network is denser in the central flat urban areas than in the surrounding suburban or mountain slope areas. The densities of transportation networks in each cell were used as weighting factors for layer 3. The areas with denser road networks were assumed to have higher population densities. This characteristic was introduced to formulate layer 3 of the population disaggregation model, shown in Fig. 4(d). As layers are added to the proposed framework, it is possible to better discriminate spatial variations of population distribution and better reveal the true population distribution pattern.

Since individual point data are confidential, disaggregation results of different levels of the model cannot be compared to the actual study grid. In order to evaluate the ability of the different layers of the model to capture the spatial population distribution pattern, the published population data at the "Li" level were assumed to be the true values. The grid cell population densities in each layer were used to compute the population in each Li. The Li level population density determined in each layer of the model is summarized in Fig. 5. This figure reveals that the population distribution patterns aggregated by Li are more similar to the published values (Fig. 5(a)) as layers are added to the model.

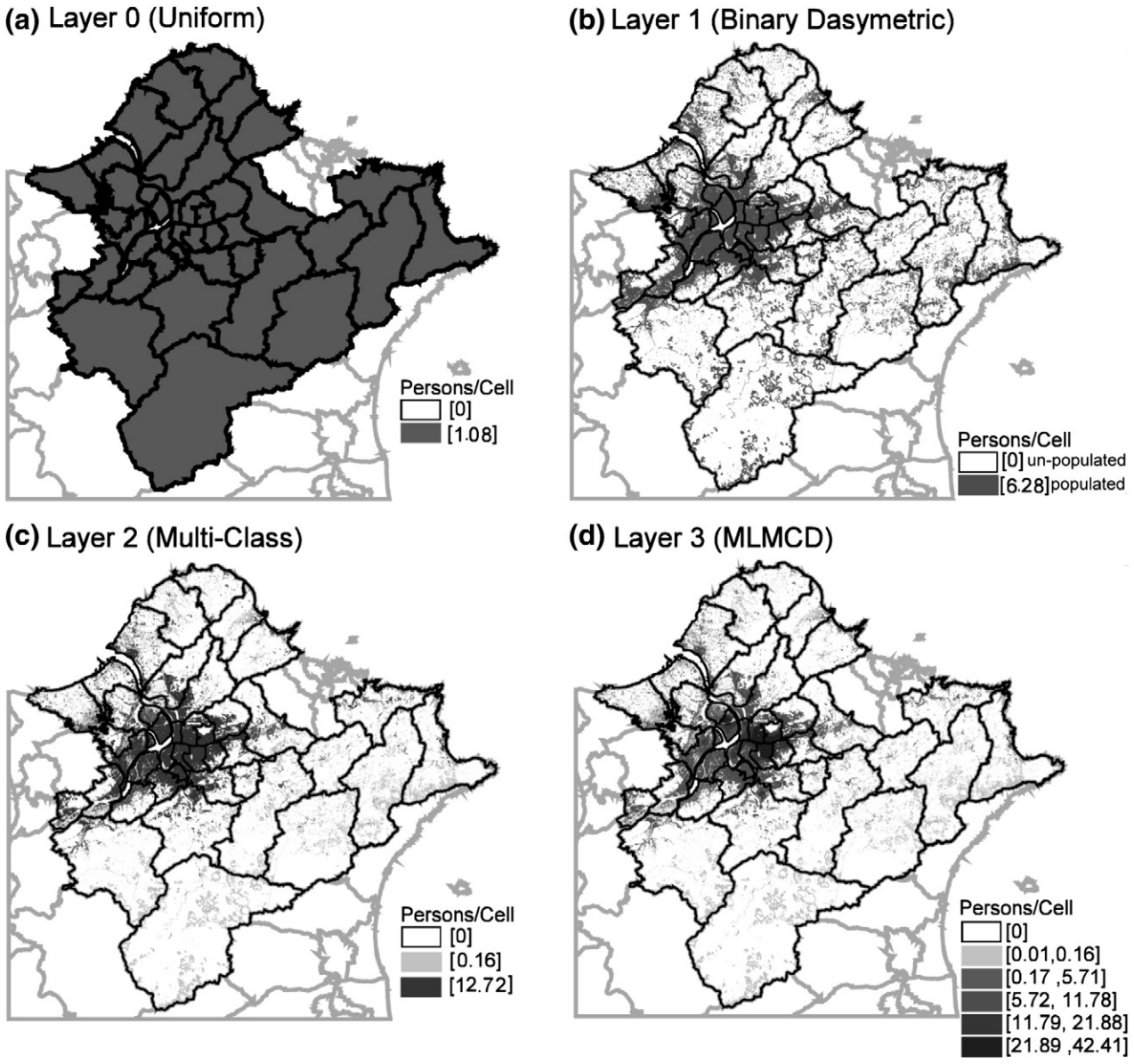

Fig. 4. Cell population densities in each layer.

The deviations of the estimated Li populations from the published values were calculated for each layer of the model. Mean Absolute deviation (MAD) and root mean square error (RMSE) were used as indices to compare error. The MAD and RMSE were calculated using Eqs. (2) and (3).

$$MAD = \frac{\sum_{i=1}^{n} \left| x_i - \hat{x}_i \right|}{n} \tag{2}$$

Table 1
The percentage of building cells in each land use type.

Land Use	Percent (%)
Agricultural*	79.1
Transportation	1.8
Water conservancy	3.0
Residential/Commercial	7.5
Industrial	1.2
Others	7.4

Note: * including farming and forest.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{n}} \tag{3}$$

where x_i and $\hat{x}_i$ are true and estimated values respectively, i is the subscript for Li, and n is the total number of Li in the study region.

The related error statistics for each layer are summarized in Table 2. It clearly shows that the range of errors decreases as layers are added. The error indices shown in Table 2 also diminish with increasing layers. The RMSE are also significantly improved from layer 0 to layer 3. It can be concluded that the multi-layer multi-class dasymetric approach can better display the actual population distribution pattern by including more ancillary information. Improvements in population redistribution are more significant as layers increase from 0 to 1 and 1 to 2 than from layers 2 to 3.

Fig. 6 shows the spatial distribution of the errors. The lighter color represents underestimated regions and darker color represents overestimated regions. Generally, the populations of central urbanized areas were underestimated and those of the surrounding mountain-slope regions were overestimated, as shown in Fig. 6. This figure also indicates that errors are attenuated in higher layers as

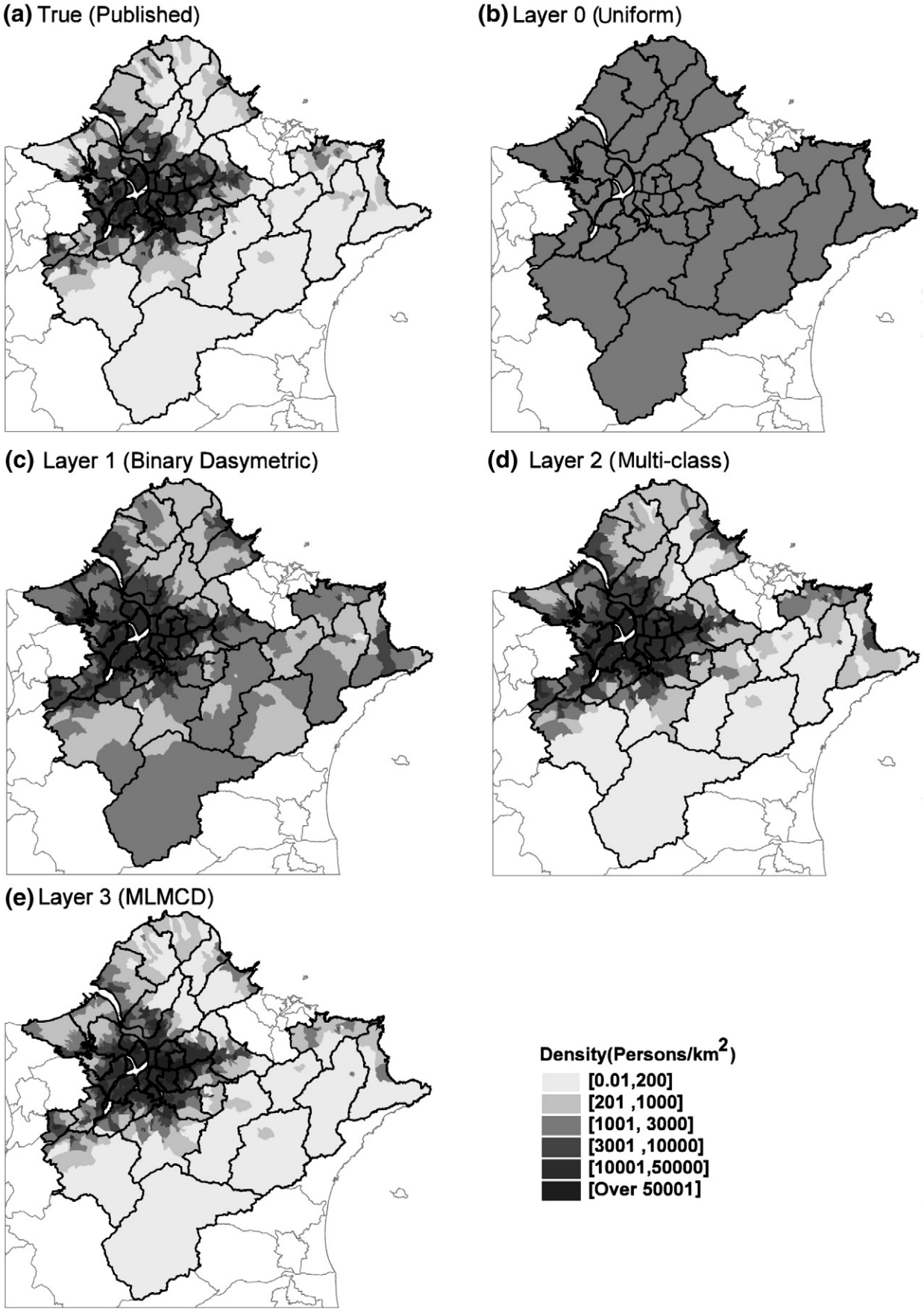

Fig. 5. Population aggregation by villages.

Table 2
Error statistics for each layer.

Error statistic	Layer 0	Layer 1	Layer 2	Layer 3
Mean	−0.1	0.0	0.0	0.0
Standard Error	488.1	253.0	114.1	110.9
Median	−3388.7	−2051.9	−1073.1	−733.4
Standard deviation	18542.6	9610.9	4333.1	4212.5
Range	502921.9	196698.1	44200.8	59743.3
Min	−23066.0	−13671.0	−10436.8	−11294.3
Max	479855.9	183027.2	33764.1	48448.9
MAD	6639.1	4210.2	2793.3	2520.4
RMSE	18536.2	9607.5	4331.6	4211.0

more ancillary data are incorporated in the population redistribution model.

Population distributions in the study area from GPW3 (Center for International Earth Science Information Network, 2005), LandScan (Dobson et al., 2000) and the proposed multi-layer multi-class dasymetric (MLMCD) model are shown in Fig. 7 for comparison. MLMCD has better result in reconstructing population distribution as shown in Table 3. The algorithm used to build GPW3 dataset was very similar to the traditional dasymetric model. The assimilation of the LandScan dataset, which used multiple land use classifications to estimate the at-risk population, resembles the multi-class dasymetric model. As shown in the figure, GPW3 and LandScan show different population patterns as compared to MLMCD. LandScan captures population distribution more accurately than GPW3, but it shows erroneous low population density in the downtown center of Taipei (shown in the circle of Fig. 7(b)).

One controversial aspect of this proposed framework, other than its demand for comprehensive data, is the assignment of weighting factors used in the disaggregation model for the different layers. These weighting factors may vary among different regions of interest and are subject to perceived local conditions and arbitrariness of the analyst (Maantay et al., 2007). Other than local knowledge and experiences, some algorithm may be required to establish these vital parameters before implementing the model. In this study, these weighting factors were determined using the census statistics in the study region. The population living in the agricultural zone is about 1% in the study region according to the 2005 agricultural census and the

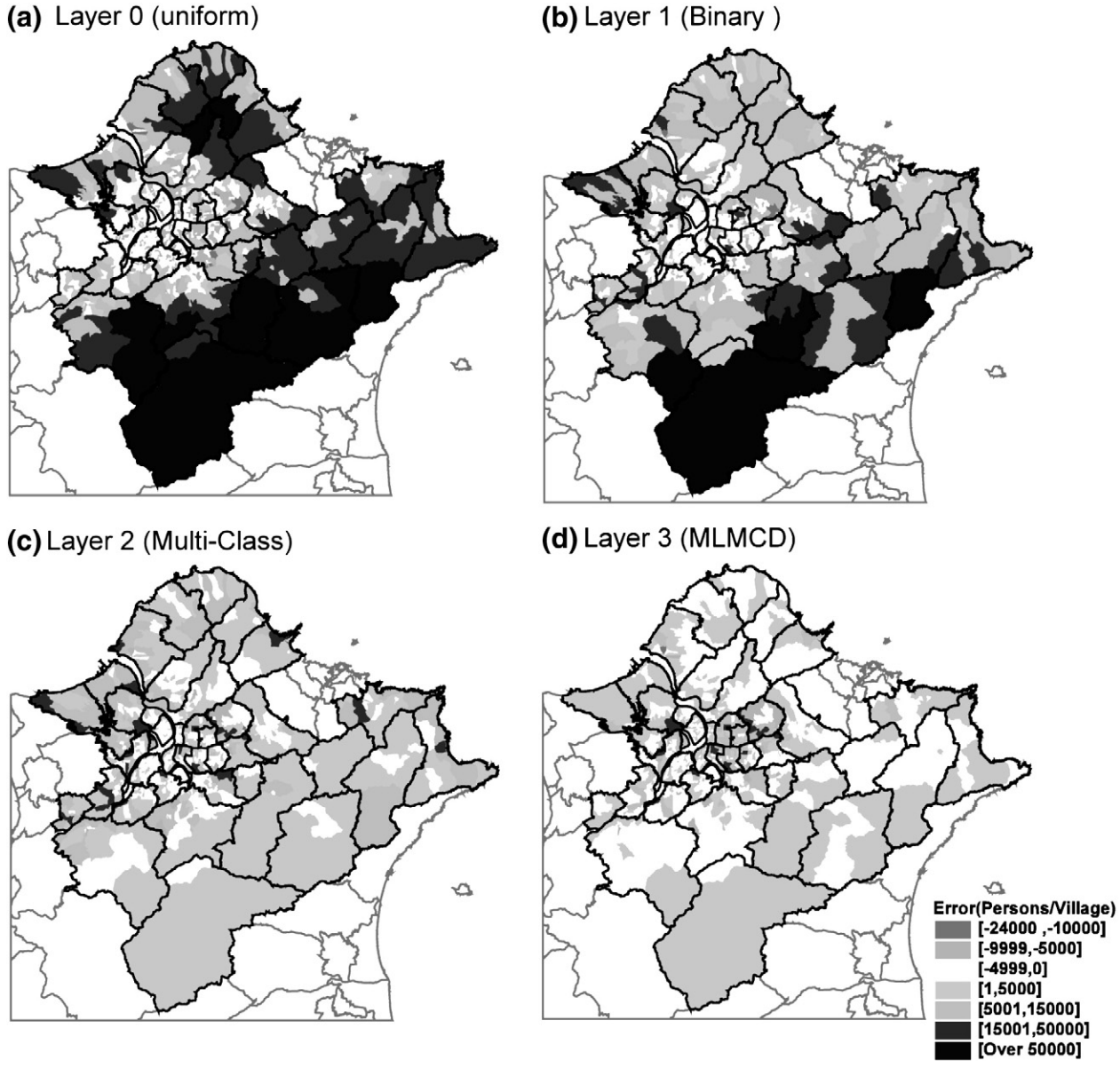

Fig. 6. Spatial distribution of errors in each layer.

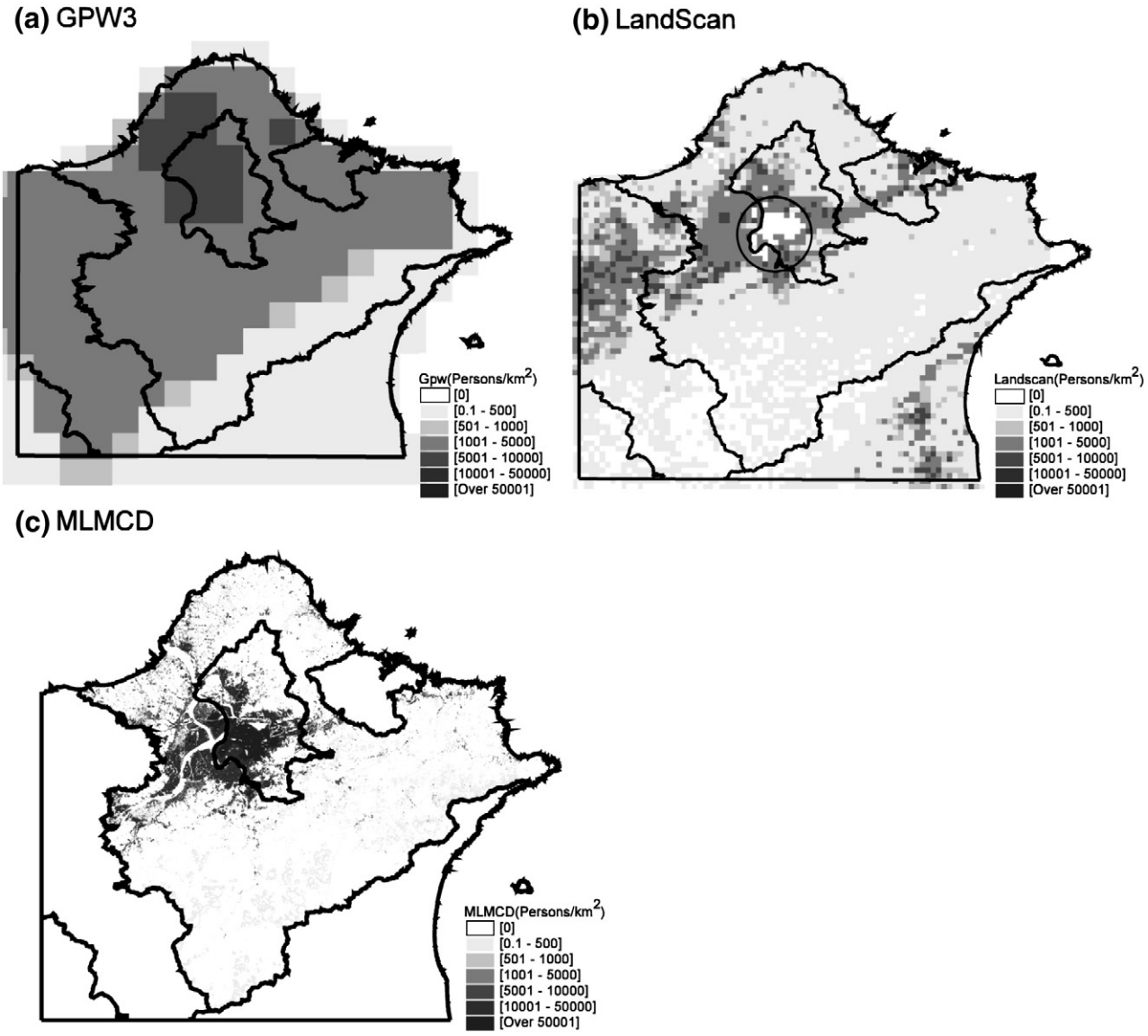

Fig. 7. Population distribution pattern from GPW3, LandScan and the MLMCD model.

2000 population census (DGBAS, 2000, 2007), and so the population redistribution weighting factors for agricultural and non-agricultural zones were set as 0.01 and 0.99 accordingly. The percentage of population living in the agricultural zone was found to be related to the regional mean population density as shown in Fig. 8. This relationship may be used to determine the weighting factors for other region in Taiwan, but similar algorithm may still need to be established for other locations.

Table 3
Error comparisons among GPW, LandScan and MLMCD.

Item	GPW	LandScan	MLMCD
Average error	−192.3	−3253.6	0.01
MAD	5799.4	4360.4	2520.4
RMSE	11020.7	7057.3	4211.0

7. Summary and conclusion

Population redistribution models are frequently needed, as most population data are published as aggregated statistics based on some spatial areal units. A multi-layer multi-class dasymetric framework was proposed in this study to better redistribute the regionally aggregated population into smaller areal units and reveal the actual spatial population distribution pattern.

As the traditional binary or multi-class dasymetric methods, ancillary data were used to better capture the population distribution

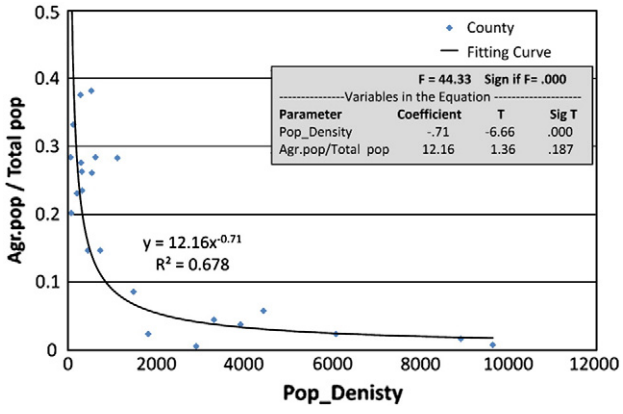

Fig. 8. Relationship between population density and the agricultural population rate.

patter. But the ancillary data used in the proposed MLMCD were applied in a progressive order according to their relationship or correlation to the population distribution characteristics. For example, Populated/Unpopulated information is applied before the land use. And the infrastructure density/accessibility will be applied after land use layer because the former can more discriminate the population distribution than the later.

Some benefits may be recognized as ancillary data are applied layer by layer in the proposed MLMCD model. For example, if detail information such as the street network density is used as a single surrogate to the population distribution, it may encounter problem in areas (e.g. commercial or industrial) with more developed but non-residential transportation infrastructure (Reibel and Bufalino, 2005). This problem can be solved by first eliminate the non-populated area (Layer 0 in MLMCD) and separate the industrial land use from the residential one (Layer 1 in MLMCD) before the street-weighting is applied in layer 2.

Metropolitan Taipei in Taiwan was used as the study area to demonstrate the efficacy of the proposed framework and the improvements of this model as compared to the traditional binary dasymetric method. Assorted data, including remote sensing images, land use zoning, topography, transportation and accessibility to facilities were incorporated in different layers of the model to improve the redistribution of the aggregated regional population. The concept of multi-layer multi-class dasymetric modeling was both useful and flexible in this case study, but cautions must be exercised in generalizing the result since it is only based on a single empirical study. More tests may be needed using data from other locations to make more firm conclusion.

This paper demonstrates that population redistribution errors can be reduced by introducing more ancillary information for population disaggregation. Different levels of accuracy in this population redistribution process can be achieved and depend on the availability and cost of data through the proposed multi-layer multi-class dasymetric framework.

The dasymetric model, whether binary, multi-class, or the proposed MLMCD, is effective and useful in capturing the spatial heterogeneity of population distribution when only aggregated population is available. The multi-class dasymetric and MLMCD improve the redistribution performance by adding more spatial discrimination into the model using further ancillary data. But the improvement may diminish if the population is originally aggregated or available at a finer spatial scale and the binary dasymetric model is applied accordingly. Practical applications of the dasymetric model should be constructed using the finest level of census data available to maximize its precision when interpolating onto other non-census areas of interest.

References

Bracken I, Martin D. The generation of spatial population distributions from Census centroid data. Environ Plann A 1989;21:537–43.

Center for International Earth Science Information Network [Internet]. Columbia University: gridded population of the world [cited: 2009 Mar 26]. Available from: http://sedac.ciesin.columbia.edu/gpw/index.jsp.

Department of Civil Affairs, Taipei City Government [Internet]. Statistic of registrated population [cited: 2007 Oct 26]. Available from: http://sedac.ciesin.columbia.edu/gpw/index.jsp.

Department of Urban Development, Taipei City Government [Intenet]. Current development of Taipei [cited 2007 Oct 26]. Available from: http://www.udd.taipei.gov.tw/planweb/FiRoger/Whole_City.htm.

DGBAS (Directorate General of Budget, Accounting & Statistics). Population and housing census; 2000. Taiwan.

DGBAS (Directorate General of Budget, Accounting & Statistics). 2005 Agricultural, forestry, fishery and husbandry census; 2007. Taiwan.

Directorate General of Budget, Accounting & Statistics, R.O.C. Population and Hosing Census 2000.

Dobson JE, Bright EA, Coleman PR, Durfee RC, Worley BA. LandScan: a global population database for estimating populations at risk. Photogramm Eng Remote Sens 2000;66:849–57.

Eicher CL, Brewer CA. Dasymetric mapping and areal interpolation: implementation and evaluation. Cartogr Geogr Inf Sci 2001;28:125–38.

Fisher PF, Langford M. Modeling sensitivity to accuracy in classified imagery: a study of areal interpolation by dasymetric mapping. Prof Geogr 1996;48:299–309.

Flowerdew R, Green M. Statistical methods for inference between incompatible zonal systems. In: Goodchild M, Gopal S, editors. Accuracy of spatial databases. London, U.K: Taylor and Francis; 1989. p. 239–548.

Flowerdew R, Green M. Developments in areal interpolation methods and GIS. Ann Reg Sci 1992;30:67–78.

Fotheringham AS, Brunsdon C, Charlton ME. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Environ and Plann A 1998;30(11):1905–27.

Fotheringham AS, Brunsdon C, Charlton ME. Quantitative Geography. London: Sage; 2000.

Goodchild MF, Anselin L, Deichmann U. A framework for the areal interpolation of socioeconomic data. Environ Plann A 1993;25:383–97.

Holt JB, Lo CP, Hodler TW. Dasymetric estimation of population density and areal interpolation of census data. Cartogr Geogr Inf Sci 2004;31:103–21.

Keping C, John M, Russell B, Roy L, Laraine H, Christina M. Defining area at risk and its effect in catastrophe loss estimation: a dasymetric mapping approach. Appl Geogr 2004:97-117.

Langford M. Obtaining population estimates in non-census reporting zones: an evaluation of the 3-class dasymetric method. Comput Environ Urban Syst 2006;30:161–80.

Langford M, Higgs G. Measuring potential access to primary healthcare services: the influence of alternative spatial representations of population. Prof Geogr 2006;58:294–306.

Langford M, Unwin DJ. Generating and mapping population density surface within a grographical information system. Cartogr J 1994;31:21–6.

Liu X, Clarke K, Herold M. Population density and image texture: a comparison study. Photogramm Eng Remote Sens 2006;72:187–96.

Lo CP. Population estimation using geographically weighted regression. GISci Remote Sens 2008;45(2):131–48.

Maantay JA, Maroko AR, Herrm C. Mapping population distribution in the urban environment: the cadastral-based expert dasymetric system. Cartogr Geogr Inf Sci 2007;34:77-103.

Martin D. An assessment of surface and zonal models of population. Int J Geogr Inf Sci 2006;10:973–89.

Martin D, Tate NJ, Langford M. Refining population surface models: experiments with Northern Ireland census data. Trans GIS 2000;4:343–60.

Mennis J. Generating surface models of population using dasymetric mapping. Prof Geogr 2003;55:31–42.

Mennis J, Hultgren T. Intelligent dasymetric mapping and its application to areal interpolation. Cartogr Geogr Inf Sci 2006;33(3):179–94.

Ministry of the Interior, Taiwan. Land Use Investigation 1995.

Peters A, MacDonalds H. Unlocking the Census with GIS. California: ESRI; 2004.

Rase WD. Volume-preserving interpolation of a smooth surface from polygon-related data. J Geogr Syst 2001;3:199–213.

Reibel M, Bufalino ME. Street weighted interpolation techniques for demographic count estimation in incompatible zone systems. Environ Plann A 2005;37:127–39.

Relbel M, Agrawal A. Land use weighted areal interpolation. GIS Planet 2005 International Conference, Estoril, Portugal; 2005.

Statistics Bureau Japan [Internet]. Population census 2007 [cited 2007 Nov 12]. Available from: http://www.stat.go.jp/english/data/kokusei/e_cen_en.htm.

Unwin DJ. GIS, spatial analysis and spatial statistics. Prog Hum Geogr 1996;20:540–51.

USCB. Population and Household. http://www.census.gov/. US Census Bureau USA, 2005.

Weichselbaum J, Petrini-Monteferri F, Papathoma M, Wagner W, Hackner N. Sharpening census information in GIS to meet real-world conditions—the case for earth observation. In: Brebbia CA, Kungolos A, editors. Sustainable development and planning II. Greece: Wessex Institute of Technology; 2005. p. 143.

Wu SS. Incorporating GIS building data and census housing statistics for sub-block population estimation. 2006 Summer Assembly University Consortium For Geographic Information Science; 2006.

Wu SS, Qiu X, Wang L. Population estimation methods in GIS and remote sensing: a review. GISci Remote Sens 2005;42:58–74.