Available online at www.sciencedirect.com
SCIENCE DIRECT®
Ecological Modelling 189 (2005) 72–88
ECOLOGICAL MODELLING
www.elsevier.com/locate/ecolmodel

Modeling population density using land cover data

Yongzhong Tian [a,b,*], Tianxiang Yue [a], Lifen Zhu [c], Nicholas Clinton [d]

[a] *State Key Laboratory of Resources and Environment Information System, Chinese Academy of Sciences, Beijing 100101, PR China*
[b] *School of Resources and Environment Sciences, Southwest China Normal University, Chongqing 400715, PR China*
[c] *Center for Chinese Agriculture Policy, Chinese Academy of Sciences, Beijing 100101, PR China*
[d] *International Institute for Earth System Sciences, Nanjing University, Nanjing 210093, PR China*

Received 18 May 2004; received in revised form 16 February 2005; accepted 24 March 2005
Available online 16 June 2005

Abstract

This study investigates the correlation between land cover data and other factors that affect population distribution. The results show that land cover data contain sufficient information to infer population distribution and can be used independently to model the spatial pattern of population density in China. China's population distribution model (CPDM) was developed based on land cover data to calculate population density in China at 1 km resolution. For cells in rural areas, population probability coefficients were calculated based on weighted linear models, the weights of land cover types being derived from multivariate regression models and on a qualitative order of land types in 12 agro-ecological zones. For cells in urban areas, a power exponential decay model based on city size and the distance from urban center was employed to calculate population probability coefficients. The models were validated in sampled cells using ancillary population data. The validation shows the mean relative error of estimated population to be 3.13 and 5.26% in rural and urban areas, respectively. Compared to existing models, the accuracy of CPDM is much higher at cell, county and province scales.

Keywords: Population density; Land cover; Distance decay model; Grid

1. Introduction

Population density could be distinguished into human population density (Yue et al., 2003, 2005a) and wildlife population density (Miller et al., 2002; Sekimura et al., 2000; Kokko and Lindstroem, 1998). For both of them, land cover is a control variable (Yue et al., 2005b; Alexande and Shields, 2003; McCarthy and Lindenmayer, 1998). This paper focuses on modeling human population density on the basis of land cover data.

The distribution of human population has been identified as one of the key datasets required for improving understanding of human impacts on land and water resources. Human population distribution data will also improve projections of the environmental consequences that may be expected under varying models of economical growth (Clarke and Rhind, 1992; Elvidge et al., 1997). These predicted environmental

* Corresponding author. Tel.: +86 1064889847.
E-mail address: tianyz@lreis.ac.cn (Y. Tian).

0304-3800/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2005.03.012

and economic data provide valuable information for decision-makers such as governments, enterprises and individuals. Human population density is one of the major indicators used to describe human population distribution. However, it is a general indicator, and consequently hides the internal variability of choropleth units (Zhang, 1997). The larger the size of the choropleth unit, the more generalized the data are (Demers, 1997). So human population density in an administrative region does not provide the spatially explicit details (Elvidge et al., 1997) necessary to describe the actual distribution of human population in the region (Demers, 1997). Usually, two mapping methods, dot distribution mapping and dasymetric mapping, are employed to improve the detail of human population distribution mapping. Dot mapping records the amount and location of human population by points, but the exact geographical location is not precise (Demers, 1997). Dasymetric mapping based on the idea of choropleth maps, invented by Wright in 1936, improves the quality of the original choropleth maps by dissolving the boundaries imposed by some smaller sub-areas (Wright, 1936; Demers, 1997). With the rapid increase in demand for high resolution population data and the introduction of new technology, such as geographic information systems (GIS) and remote sensing (RS), many recent studies use the digital simulation technology of Dasymetric mapping to estimate raster based human population distributions.

Gridded Population of the World (GPW) and Landscan are frequently used global population datasets. GPW proportionally allocated total population to grid cells based on the assumption that population is distributed evenly over administrative units (Tobler et al., 1997). LandScan distributed census counts to 30- by 30-s grid-cells based on probability coefficients calculated from road proximity, slope, land cover, and nighttime lights (Dobson et al., 2000, 2003; Dobson, 2003). In addition, Lo (2001) developed allometric growth models and linear regression models to model the non-agricultural population of China in 1997 using the nighttime lights data from the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS). Sutton (1997) indicated the limitation of nighttime data in rural regions and estimated only the urban population of North America. Wang and Michel (1996) took advantage of a gravity model to simulate the urban population density. Liao and Sun (2003) spatialized census data in Qinghai–Tibet plateau based on five factors. Yang et al. (2002) studied the method of population specializations and used this effect to create a simulated population grid of the Shandong province. Sutton et al. (2003) explored some theoretical and empirical efforts at estimating ambient population density and proposed a quantitative means for evaluating their validity. Yue et al. (2005a,b) used surface modelling of population distribution (SMPD) based on grid generation method to simulate the population density in 1 km cells in 1930, 1949, 2000 and 2015.

Although these studies focused on the factors that affect population distribution, the correlations and relative influence of those factors are rarely investigated. The inclusion of too many factors makes modeling more complicated, causes information redundancy, and magnifies the effect of redundant information in population density simulation (Zhang and Yang, 1992). Some studies did not fully consider the difference in model parameters between regions. Others ignored the difference of distribution pattern between rural and urban populations and used the same models in both rural and urban regions. In this study, we used China as a case study to address three issues:

1. Can land cover data be used to model population density on grid-cells independently?
2. Using this technique, what is the most effective method for creating a raster based population density surface?
3. What is the accuracy of estimate?

2. Data sources

2.1. Data of population

The original census data comes from Chinese population by county in 2000 (Chinese Ministry of Public Security, 2001). These data are available as an attribute of the administrative polygons at the county level.

2.2. Data of land cover

The original land cover data comes from Data Center of Resources and Environment, Chinese Academy of Sciences (CAS). It is derived from Landsat Thematic Mapping (TM) images in 2000 according to the

Table 1
Land Cover classification system and residential classification

Primary types		Secondary types		Residential types
Code	Name	Code	Name	
1	Farmland	11	Paddy field	A
		12	Non-irrigated field	A
2	Woodland	21	Forest	A
		22	Shrub	A
		23	Sparse woodland	A
		24	Other woodland	A
3	Grassland	31	Dense grassland	A
		32	Moderate dense grassland	A
		33	Sparse grassland	A
4	Waters	41	River	N
		42	Lake	N
		43	Reservoir and pond	N
		44	Glacier and snow	N
		45	Beach	N
		46	Bottomland	N
5	Build up area	51	Urban area	R
		52	Rural Residential area	R
		53	Other built-up area	R
6	Unused land	61	Sandy land	N
		62	Gobi	N
		63	Saline-alkali land	N
		64	Marsh	N
		65	Bare land	N
		66	Bare rock and gravel land	N
		67	Other unused land	N

Note 1: N means exclusive area, R means residential areas, A means non-residential areas, R and A are habitable areas. Note 2: in this paper, the land cover types will be replaced by a symbol composed of "land" and their codes. For example, land11 will represent paddy field.

Class System of Land use/Land cover in Remote Sense Mapping at 1:100,000 scale (Table 1, Liu, 1996; Liu and Buhe, 2000; Liu and Zhuang, 2003). In its original data format, the land cover data is an ArcInfo coverage. For this study, it was converted into 25 raster files in Environment Systems Research Institute (ESRI) Grid format at 1 km resolution using the cell-based encoding method of percentage breakdown. Each raster file represents a land cover type; the value of each grid-cell in the raster file is the percentage of the type of land cover in the grid-cell. Prior to this conversion, all the towns which have lower administrative grades than county seats were changed from 'rural residential' land cover type to 'urban' type. This step was necessary to standardize the labeling of towns and cities (as 'urban') for the purposes of population distribution modeling. In the original land cover data, areas of dense population ('urban' areas from the modeling perspective) are only designated as 'urban' if they have an administrative grade of county seat or higher.

2.3. Digital elevation model (DEM)

The original DEM is the Chinese part of the GTOPO30 (global topography at 30-arcsecond resolution) dataset derived from U.S. Geological Survey's (USGS) Earth Resources Observation System (EROS) Data Center. After reprojecting and resampling, it was converted into a 1 km resolution DEM of China. The slope data were derived from the DEM.

2.4. Temperatures

The temperature data originated from the observations of 636 climate stations of the National Weather Bureau. To get the ground temperatures of grid-cells, the geo-located temperatures were first converted to sea level equivalent values according to the altitude of observatory stations and the temperature lapse rate (6.4 °C 1000 m). Then they were interpolated to 1 km resolution cells using Ordinary Kriging. Finally, the DEM was used to convert the interpolated temperature of raster cells at sea level back into that at ground level according to the temperature lapse rate and the altitude of cells.

Ancillary data: Other data include railways, highways, rivers and cities. They were derived from the database of Chinese resources and environment at 1:1,000,000 and 1:4,000,000 scales in ArcInfo coverage format.

All the data were integrated into ESRI ArcGIS in an Albers Equal Area Map Projection.

3. Can land cover be used independently to model population density in grid-cells?

Population geographers divide the factors that affect population distribution into two types (Zhang, 1997;

Table 2
Principal component analyses of factors affecting population distribution in China

Principal component	1	2	3	4	5
Eigenvalues	5190630	1079054	125633	14	7
Contribution percent	81.16	16.87	1.96	0	0
Eigenvectors					
DEM	−0.3076	0.9512	0.0258	0.0032	0.0018
Slope	−0.0004	0.0007	−0.0004	0.2878	−0.9577
Temperature	0.0013	−0.0034	−0.0002	0.9577	0.2878
Cropland	0.9497	0.3086	−0.0538	−0.0001	−0.0002
Rural residential area	0.0591	−0.008	0.9982	0.0002	−0.0004

Hu, 1983). One is natural factors, such as climate, elevation and slope, which are the basic factors of population distribution. The second type, which also plays a decisive role in population distribution, is socioeconomic factors, such as land cover, railway, road, and city location (Zhang, 1997). The most often used factors for modeling population distribution include elevation, slope, temperature, transportation line proximity, cities and land cover (Yue et al., 2003). However, most of these factors are closely related to land cover.

3.1. Correlations between land cover and other factors affecting Chinese population distribution

3.1.1. Land cover and natural factors

To examine the redundancy of information in the factors affecting population distribution, a principal component analysis based on 1 km^2 grid-cells was conducted on two types of land cover (farmland, rural residential area) and three main natural factors (elevation, slope and temperature). The analysis results (Table 2) show that the first principal component includes 81% of the information about population distribution. According to the load matrix, the factor highly correlated with the first principal component is farmland, the coefficient of which reaches 0.9497. It means that land cover data, especially farmland, includes most of the population distribution information. Comparatively, the other factors are far less important than land cover. Although the DEM is also highly correlated with the second principal component, which contains nearly 17% of the population distribution information, it may not be used as a variable to model population density because of its correlation with other factors (Table 3).

3.1.2. Buffer analyses of land cover and other factors

Buffer analyses of railways, highways, rivers and cities were conducted to detect the relationships between land cover and these factors. Because those factors are linear or point features, they cannot be directly used to analyze their correlation with land cover. For this study, buffer zones of main highways and rivers (10 zones), as well as railways and cities (20 zones), were built at 10 km intervals (Fig. 1), and average rural population density, farmland area, and rural residential area were summed in each zone. This established the relationships between population density and buffer distance, farmland and rural residential area (Table 4). The rural population density used here is derived from the choropleth map of rural population density by county unit, converted to a 1 km resolution grid.

Table 3
Pearson correlation coefficients between the DEM and other factors affecting population distribution at three scales

Scale	Degrees of freedom	Pearson correlation coefficients between the DEM and				Critical correlation coefficients at the 0.01 confidence level
		Slope	Temperature	Cropland	Residential area	
Cell	9503010	0.3709	−0.7722	−0.4528	−0.2164	<0.1
County	2354	0.6546	−0.6542	−0.5617	−0.4363	<0.1
Province	31	0.5925	−0.68851	−0.6712	−0.5362	0.4426

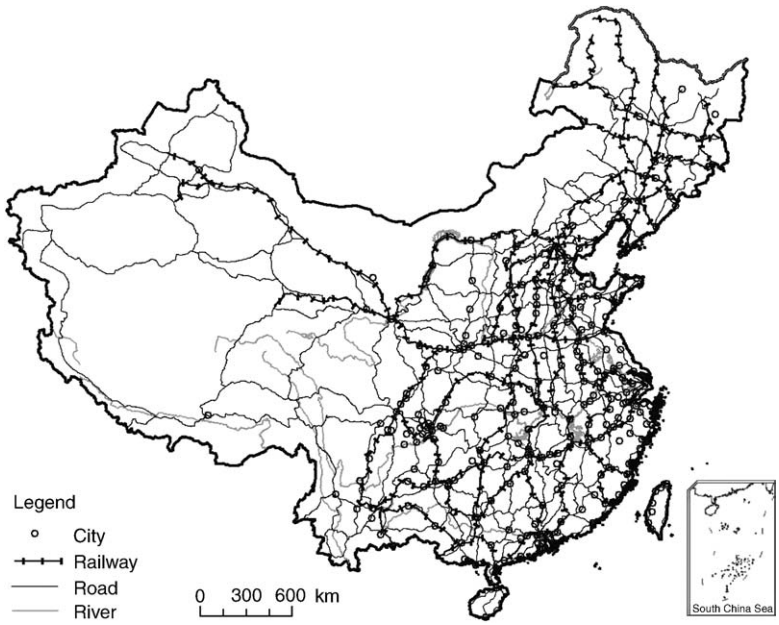

Fig. 1. Main railways, highways, rivers and cities in China.

The results listed in Table 4 show that rural population density is significantly correlated with the distance from railways, highways, rivers and cities, indicating that these factors have important influence on population distribution. The results in Table 4 also show that both farmland and rural residential areas have high correlation with the average rural population density in the buffer zones of railways, highways, rivers and cities, suggesting that population density change with the distance from these factors can be derived from land cover data.

3.2. Influence of China's population system and land system upon population distribution

Systems of land use and population location in China are the origin of relationships between population and land cover distribution evident in modern times. China was an agricultural country historically, and owning a piece of land has been the dream of Chinese peasants for generations. Land Revolution in the early 1950s allocated an average amount of land to peasants. Although collective farming was conducted

Table 4
Relationships of average rural population density with buffer distance, farmland area and rural residential area in buffer zones of some affecting factors

Factors	Items	Buffer distance (km)	Farmland (100 m^2)	Rural residential area (100 m^2)
Railway	Model	$y = -124.57\ln(x_1) + 704.77$	$y = 0.0845x_2 - 36.353$	$y = 0.8151x_3 + 33.62$
	R^2	0.9955	0.9781	0.984
Highway[a]	Model	$y = -126.88\ln(x_1) + 587.94$	$y = 0.0034x_2^{1.3845}$	$y = 0.9244x_3 + 9.9662$
	R^2	0.99	0.9998	0.9914
River[a]	Model	$y = 342.96x_1^{-0.1648}$	$y = 0.1132x_2 - 80.349$	$y = 0.6852x_3 + 54.116$
	R^2	0.8902	0.961	0.9509
City	Model	$y = 811.77\,e^{-0.017x_1}$	$y = 2e - 7x_2^{2.5898}$	$y = 0.0889x_3^{1.4396}$
	R^2	0.9928	0.9544	0.9463

[a] Buffer analyses of highways and rivers employed ten buffer zones of 10 km intervals and railways and cities employed 20 zones.

in 1960s and 1970s, it had little influence on the clustering of population because the collective farming was employed in the smallest administrative unit (the farming team, Shengchandui, is usually less than 1 km^2), had no more than 100 persons, and peasants participating in collective farming still resided in their original homes. From 1980s, the Household Responsibility System of land use was implemented across the country. Peasants contracted and managed dispersed lands according to their familial networks. The government of China recently declared that the household responsibility system would continue and the land tenure would extend to 30 years. This shifted the contract term of land from short to long and further enhanced the spatial attachment between land and peasant. However, another Chinese system, the Household Registration System (hukou), divided the population into "agricultural" and "non-agricultural" sectors, with different privileges according to their residential status, and further reinforced the land-peasant bond. As noted by Zhong (2001), the system almost binds peasants to their lands.

Generally, the rural–urban migration now is much easier than in earlier times, and in some regions, peasants are encouraged to move into cities and towns. These people are registered as urban citizens and get the corresponding legal statuses and privileges, but at the same time, they lose their lands in rural regions and are not counted in the rural population any more. For most of the peasants who rush into cities, their intentions are to hunt for jobs, not to settle down. Before they can obtain the legal status and corresponding privileges as urban citizens, they will not give up their rights to contract rural land and, in any case, they would like to retain their rights to houses and lands in rural regions (Yang and Wang, 2002; Zhong, 2001). When farming is busy, they will go back to their villages. For this reason, the relaxation on reform of migration laws is unlikely to alter the close relationship between rural population and land in the near future.

In addition, it should also be noted that the progress of agricultural industrialization in China is slow and the agro-production mode is still predominantly at the pre-industrial stage in which labor is the most important productive factor. The distribution of population is in large part a relic of a more agrarian time (Zhong, 2001). Before the rise of large-scale mechanized agriculture, especially in the south, the proximity of labor to the farm is constrained to about 1 km, meaning that peasants are not able to live far from their lands.

3.3. Two problems caused by scale issues

3.3.1. Are the transportation lines, rivers and cities always necessary for modeling?

Population distribution is influenced by factors at all scales simultaneously, however, the intensity and manifestations of the influences from different scales are very different. The factors at a national scale, such as main roads, railways, rivers and cities, control the basic pattern of national population distribution. But at a county level, the roads between county sites, or towns, or even villages may be much more important for rural population distribution. It follows that the entrances of larger roads and the stations of railways are better represented as attractions for population by point features rather than line features. Lower–grade roads are very easy to access owing to their numerous entrances, making their "linear attractiveness" more obvious. Fig. 2 illustrates the difference of attraction, from a residential perspective, between an expressway and an ordinary highway in Chongqing. Clearly, if administrative units at the county scale were used as control areas in population simulation, the factors affecting population at the county scale would be more important and the factors at a national scale would be almost unnecessary. However, it is very difficult to get road network data in all counties. Furthermore, a series of tasks, such as determination of appropriate road types to be examined, establishment of function indices, function distances and distance decay manners of these roads for the purposes of population estimation, are problematic due to high place to place variability at smaller scales. The same questions would apply to railways, rivers and cities. However, land cover data, which has close correlation with those factors as discussed in the foregoing sections, can be easily obtained via remote sensing imagery.

3.3.2. Is the residential area of land cover sufficient for modeling?

The loss of information in each land type at different scales is rather different. Li and Zhuang (2002) studied land cover data at three scales and found that the area error of all land cover types became larger when the scale became smaller. For example, in Shandong

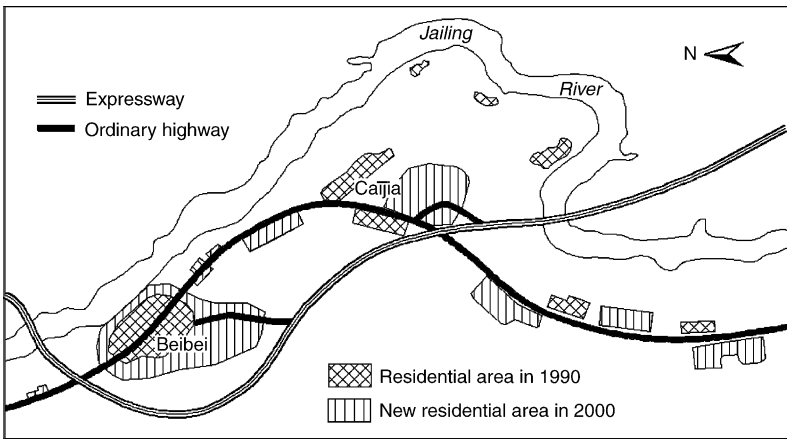

Fig. 2. Comparison of the point attraction of expressway and the linear attraction of ordinary highway for population in Beibei, Chongqing.

province of China, the area of urban and rural residential land at 1:100,000, compared with that at 1:10,000, decreased 10.4 and 15.1%, respectively. At the same time, paddy field and non-irrigated field increased 18.2 and 24.8%, respectively, while the difference of woodland and grassland between the two scales is very small. This indicates that small residential areas are "absorbed" by larger patches of farmland. In addition,

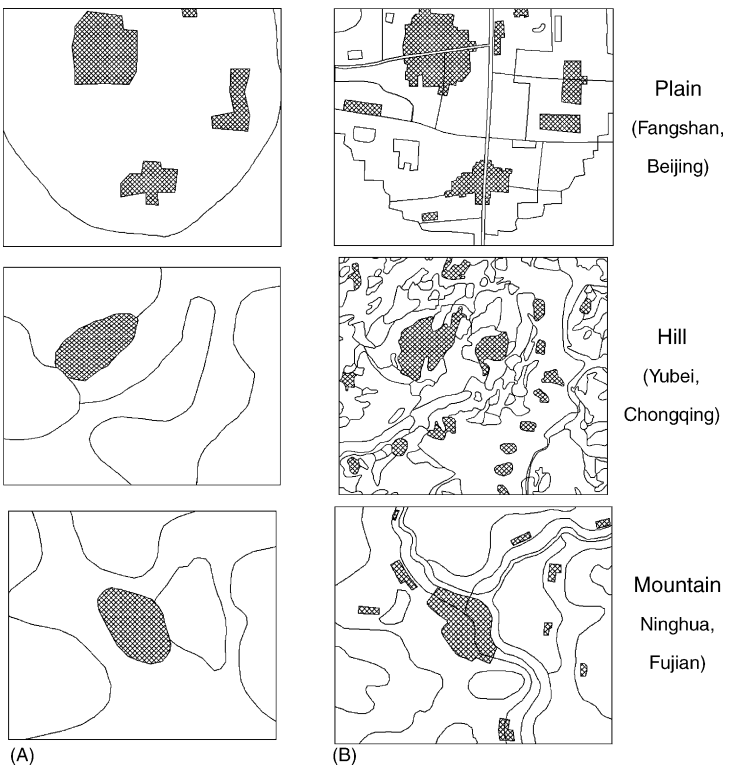

Fig. 3. Contrast of residential areas in three typical topographical regions at two scales.

after studying a single property using land cover data at 1 km resolution, Liu et al. (2001) found that more than 60% of residential areas were lost. Fig. 3 compares the residential areas in land cover maps at 1:100,000 used in this study with that at 1:10,000 in three typical topographical regions of China. The comparison shows that little of the residential area is lost in plains, while in mountains and hills, a greater number of residential areas are lost due to their scattered distribution. So, for China, a country mostly composed of mountains and hills, it is not enough to take into account only residential areas when modeling the population in grid-cells using land cover data.

4. Methods

4.1. Research approach

4.1.1. Modeling based on land cover

Land is a synthesis of many natural and social factors which have acted in concert for long periods of time (Zhang, 1997). As discussed above, because land cover is highly correlated with many factors affecting population distribution, it is a good proxy for estimating the characteristics of population distribution. This study determines the feasibility of modeling population density by means of land cover data.

4.1.2. Controlling total population at the county scale by rural area and urban area, respectively

Administrative polygons at the county scale are the minimum mapping units available at a countrywide scope at present, and therefore represent the best country level population data. Quantification of total population by rural area and urban area at the county scale therefore avoids the reallocation of population between different counties and between rural and urban areas. Population probability coefficients were normalized in rural and urban areas of counties in order to make use of the best national level population data.

4.1.3. Modeling rural area and urban area separately

Urban and rural areas were treated separately due to the difference of affecting factors between rural and urban areas on population distribution. This separation was necessary to avoid mistakes such as over or under estimation of population density within rural or urban grid-cells. Integrated models run this risk due to the over generalization of variables controlling population distribution when applied to disparate land types.

4.1.4. Zonal modelling

Significant geographic differences of natural, social, economic and historical factors have resulted in different characters of land use in different regions. Modeling in relatively homogeneous zones can reduce effects of these differences on population distribution. The model was implemented in homogenous zones in an effort to minimize population estimate error resulting from the geographic disparity in the factors affecting population.

4.2. Model

Consistent with the above analyses and research approach, the Chinese Population Distribution Model (CPDM) was developed for land cover data at 1 km resolution according to the following:

$$\mathrm{POP}_{ij} = P_{i\mathrm{r}} \times \frac{V_{j\mathrm{r}}}{\sum_{j=1}^{k} V_{j\mathrm{r}}} + P_{i\mathrm{u}} \times \frac{V_{j\mathrm{u}}}{\sum_{j=1}^{k} V_{j\mathrm{u}}} \quad (1)$$

Where POP_{ij} is the population of the jth cell in ith county; $P_{i\mathrm{r}}$ is the rural population in ith county; $P_{i\mathrm{u}}$ is the urban population in ith county; k is the number of cells in ith county; $V_{j\mathrm{r}}$ is the rural population probability coefficient of the jth cell in ith county; $V_{j\mathrm{u}}$ is the urban population probability coefficient of the jth cell in ith county.

The first term of Eq. (1) is used to calculate the rural population in cells, and the second term is used to calculate urban population. From Eq. (1), it can be seen that a key point of the model is the determination of population probability coefficients of rural and urban cells, which are normalized by county. Thus the rural and urban population in each county will be distributed to the cells in the county based on their normalized rural and urban coefficients.

4.2.1. Probability coefficients of rural population

Land cover types are divided into two kinds in this study: exclusion areas and habitable areas (Table 1). The former, including all the secondary types of waters and unused land, are excluded as input variables for

modeling because they are not fit for habitation at least at present, which means their probability coefficients of population ought to be assigned zero. The latter, including all the other land cover types, also can be subclassified into residential and non-residential areas. Residential areas, including urban areas, rural residential areas and other built-up areas, are the essential variables for modeling because they are directly related to population distribution. Although non-residential areas, including farmland, woodland and grassland, are not areas used primarily for dwellings, they are never the less habitable and contain some scattered residences. They are likely designated "non-residential" due to the resolution of remotely sensed data, map accuracy, and the scale issues of the land cover data (Fig. 3). The loss of residence information differs significantly between southern and northern regions, plains and mountains and hills. For example, population is more centralized in the north and plains areas, whereas it is more scattered in the south, mountainous areas and hills. To address this issue and calculate accurate probability coefficients of cells, it is necessary to examine the residential information inherent to each type of land. For this purpose, the following process was adopted:

(1) Ecological zoning: Based on the mode of agricultural production, the productivity of farmland, heat, water, and landform, China can be classified into 12 agro-ecological zones (Fig. 4, Chen, 2001).
(2) Selecting variables: Univariate linear regression models were built between area of habitable land types and rural population by county in each ecological zone. If the model was significant, then the land type in the model was used as one of the variables to determine probability coefficients.
(3) Modeling: The following multivariate regression model was developed to calculate the weight of input variables in each ecological zone:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (2)$$

where y is rural population; x is the area of land cover type chosen in step 2; β is the parameter.

Although the chosen variables are significant in univariate regression, some of them might not be significant in multivariate regression. However, for a reasonable regression equation, only those significant variables should be used. Moreover, the collinearity between variables should be as small as possible (William, 2000; Zhang and

Fig. 4. Agro-ecological zones of China (from Chen, 2001).

Table 5
Stepwise regression coefficients of land cover types with population by county in each agro-ecological zone

Ecological zones	Land cover types									
	Land 11	Land 12	Land 21	Land 22	Land 23	Land 24	Land 31	Land 32	Land 33	Land52
1	0.00104	0	0	0	0	0	0	0	0	0.22361
2	*0.02784*	0.02784	0	0	0	0.00755	0	0	0	0.21002
3	0.03869	0.03112	0	0	0.00424	0	0	0	0	0.08826
4	0.07662	0.02550	0.00314	0	0.00266	0.00471	0	0	0	0.10240
5	0.01259	0.00448	*0.00049*	0	0	0	0.00053	0.00087	0	0.05307
6	0.01596	0.00776	0	0.00056	0.00022	0.00061	0	0	0	0.01610
7	0.02590	0.00418	*0.00139*	0	0	0	0	0	0	0.48514
8	0.06216	0.04468	0	0	0	0	0	0	0	0.17316
9	0.03566	0.03227	0	*0.00236*	0	0	0	0.00175	0.00262	0.22797
10	0.04991	0.02052	0	0	0	0	0	0	0	0.14774
11	0.01044	0.01044	0	0	0	*0.00095*	0	0	0	0.06393
12	0.01250	0.01021	0	0	0	0.00041	0	0	0	0.05540

Note: Italic numbers have been modified according to the qualitative order and the food productivity of each land type.

Yang, 1992). The best solution for these problems is to calculate the parameters using multivariate stepwise regression. To ensure the credibility of the parameters, the critical significance level in the calculation was 0.10. Table 5 shows the regression coefficients of the chosen land cover types.

(4) Modifying coefficients: The outcome of this stepwise regression is derived from the statistical correlations between population and land, but it must also obey certain geographic rules. Usually, land with higher food productivity can support a larger population so it is logical to assign a higher coefficient to this type of land. According to the degree of correlation between agricultural land and rural population, the following qualitative order of weight for land types was built as expression (3), and those coefficients that did not conform to expression (3) were modified (Table 5) according to the food productivity of each land type (Cao et al., 1995; Chen, 2001).

$$\begin{aligned}&\text{rural residential areas} \geq \text{paddy field} \geq \\ &\quad \text{non-irrigated field} > \text{woodland and} \\ &\quad \text{grassland} > \text{other land} \geq 0\end{aligned} \tag{3}$$

(5) Calculating probability coefficients: The following weighted linear model was used to calculate the probability coefficients of rural population for each grid-cell:

$$V_{jr} = \sum_{n=1}^{10} A_{jn} W_{mn} \tag{4}$$

where A_{jn} is the area of the nth land type in the jth grid cell; W_{mn} is the modified stepwise regression coefficient of nth land type in mth ecological zone shown in Table 5.

4.2.2. Probability coefficients of urban population

Generally, urban average population density is directly proportional to urban size. The larger the size, the larger the population density is (Ye, 2001). However, inner differences of urban population density also exist. Usually, population density decreases from the center toward the outside of town. Although the factors affecting urban population distribution are too many to be formulated by a simple mathematical equation (Zhang, 1997), as a general rule, the density is correlated with the size of the town and the distance from the center. That general rule can be expressed as the following equation:

$$V_L = f(S, D) \tag{5}$$

Where V_L is the urban population density coefficients at site L, S is the urban size, and D is the distance of L from the urban center.

As a model of spatial distribution of urban population, the exponential decay model (Clark, 1951) was

typical in early research. Thereafter, there appeared many other relative models such as the Gauss model (Sherratt, 1960; Tanner, 1961), and the negative power-exponential model (Smeed, 1961). In recent years, fractal models for the decay of urban population density were proposed (Chen, 1999; Feng, 2002,). As research progressed, some "unified" models were introduced such as the Newling model, a unification of the Clark model and Sherratt model (Newling, 1969), and the Gamma model, which unifies the Clark and Smeed models (Batty and Longley, 1994). The former models are special cases of the unified models (Shen, 2002).

The following equation is an often-used power-exponential model of spatial urban population distribution:

$$\rho(r) = \rho_0 \exp\left[-\left(\frac{r}{r_0}\right)^{\sigma}\right] \tag{6}$$

where $\rho(r)$ is the population density at distance r from the urban center; ρ_0 is the population density of the urban center; r_0 is the functional radius of the town, σ is a restriction parameter reflecting the tendency for spatial changes of information entropy in the urban geographic system. Although the model has been widely criticized, urban economists have demonstrated its theoretical justification (Wang and Michel, 1996).

It is very difficult to get the central population density of all towns in China. However, as discussed above, the population density has positive correlation with the size of the town, and the functional radius is also related to urban size. Therefore, by integrating Eqs. (5) and (6), a model based on the size of town and the distance from urban center was constructed. This model was used to calculate probability coefficients of urban population as follows:

$$\begin{aligned} V_{mn} &= A_n \ln A_m \exp\left[-\left(\frac{r_n}{\sqrt{A_m/\pi}}\right)^{\sigma}\right] \\ &= A_n \ln A_m \exp[-(r_n\sqrt{\pi/A_m})^{\sigma}] \\ &= A_n \ln A_m \exp(-1.9874 r_n^{\sigma} A_m^{-0.5\sigma}) \end{aligned} \tag{7}$$

where V_{mn} is the urban population probability coefficients of the nth grid-cell in the mth town, A_n is the area of urban land in the nth grid cell, A_m is the area of the mth town, r_n is the distance from the nth cell to the urban center, σ is a parameter corresponding to the stage of development of the town.

In Eq. (7), the functional radius is derived from the radius of a circle with same area as the town. In regard to Eq. (7), two other questions deserve mention. One is how to determine the centers of towns. If the polygons of towns are irregular, their centroids may fall outside of them. To solve this problem, the label points of the urban polygons, which always locate inside the polygons, were used as their centers. Another question relates to the value of σ. According to the theory of urban development, a city will experience 'developing', 'developed' and 'old' stages; each stage has different characteristics of spatial distribution of population density. For example, in conjunction with suburbanization, the population density of urban center will decrease, which may cause a crateriform distribution of population density. Most Chinese cities, despite many years of building since the reform and opening, are still in the early developed stage; many middle and small towns are still in the developing stage. This means the value of σ will not be high. To distinguish the difference of spatial structure caused by the different development stages, all the cities and towns in China were classified into three types (main cities, middle cities and small towns) according to their size and other socio-economical indicators such as gross domestic production (GDP), and population (Zhang, 1997; Zhou and Xu, 1997; Feng, 2002). The values of σ for the three types are 1.43, 1.26 and 1.14, respectively, which are derived from the population density sampling of 112 urban cells.

After calculating the probability coefficients of rural and urban populations, Eq. (1) was implemented in ArcGIS software to create the simulated Chinese population density map (Fig. 5).

5. Validation and comparison

5.1. Validation

Samplings of population density were conducted in rural and urban cells to validate the simulation outcome. For rural population density, pure random sampling was applied. According to the formula for necessary sample size (Tian et al., 2003), when the sampling precision is more than 95%, allowable error is less than 7% and the sampling ratio is 0.5, the necessary sample size is 196. A random function was

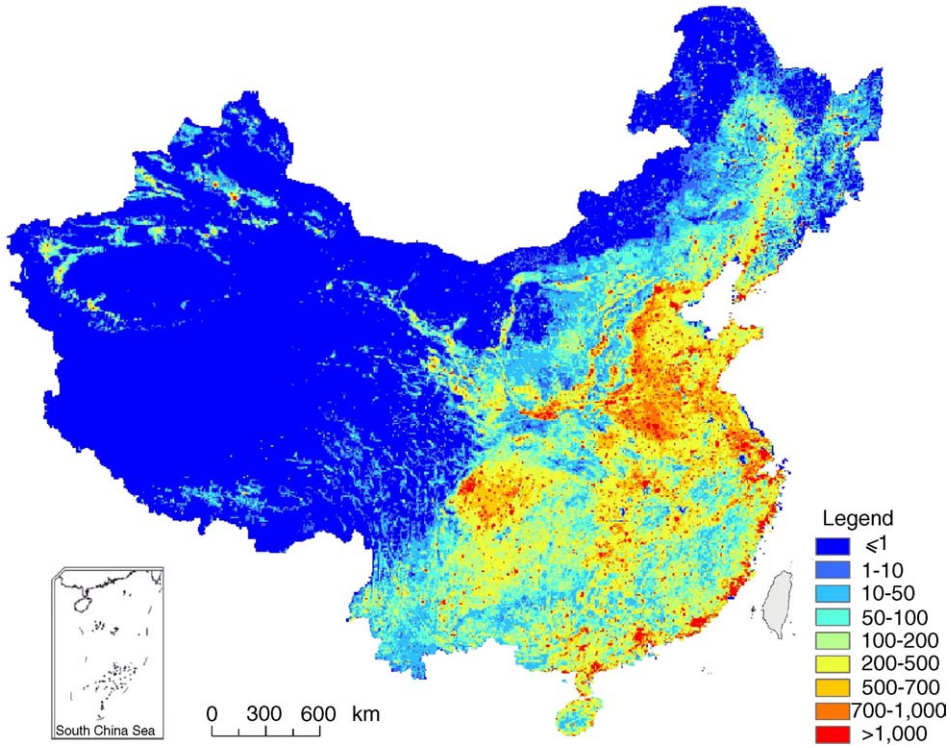

Fig. 5. Simulated Chinese populations of 1 km^2 grid-cells in 2000.

introduced to select the 196 sample cells from a sorted list of all the cells in the grid of simulated population. For urban population, stratified sampling was applied and 112 cells from towns with different landforms and sizes were used to do the validation. Because it is very difficult to get the population in the 1 km^2 cells selected for sampling, the average population densities of villages or blocks within sampled cells were used to assess the accuracy of the simulation (Table 6).

The assessment shows that the accuracy of the simulation in rural areas is much higher than that in urban areas. The percentage difference usually is less than 5 in populous regions such as Huanghuaihai plains, Sichuan basin, and the northeast plain. In some regions

Table 6
Statistics of percentage difference between simulated population and sampled population in Landscan and CPDM

Regions	Models	Cells		Percentage difference					Mean absolute relative difference (%)
				>3%	>5%	>7%	>10%	>15%	
Rural area	CPDM	Number	196	94	78	69	46	32	3.13
		%	100	47.96	39.80	35.20	23.47	16.33	
	Landscan	Number	196	111	102	89	67	55	5.49
		%	100	56.63	52.04	45.41	34.18	28.06	
Urban area	CPDM	Number	112	62	58	46	39	26	5.26
		%	100	55.36	51.79	41.07	34.82	23.21	
	Landscan	Number	112	76	68	59	50	36	8.17
		%	100	67.86	60.71	52.68	44.64	32.14	

Table 7
Comparison of population error between Landscan and CPDM at county and province scales

Level	Model	Administrative unit		Error						
				>3%	>5%	>7%	>10%	>15%	>20%	>50%
County	Landscan	Number	2317	1586	1199	852	502	236	138	22
		%	100	67.37	50.93	36.19	21.33	10.03	5.86	0.93
	CPDM	Number	2317	0	0	0	0	0	0	0
Province	Landscan	Number	31	17	9	3	2	1	1	0
		%	100	54.84	29.03	9.68	6.45	3.23	3.23	0.00
	CPDM	Number	31	0	0	0	0	0	0	0

with sparse population, the percentage difference is very large, even greater than 100 in a few cells, but the absolute error of population is low, often less than one person. In urban areas, more than half of the sampling cells have more than 3% difference. For some cities, especially for those located in hills and mountains, the average difference is larger than that in other towns because of their complicated spatial structure. In this study, it was assumed that the towns have only one center, which is not accurate for towns with multi-centers or sub-centers. This could explain the discrepancy of urban population density. However, for population simulation at 1 km resolution, the population distribution in towns is not as important as that in rural areas. Moreover, 1 km resolution is insufficient to show the inner details of urban population distribution. Another reason for the differences between the sample and the estimate from the simulation is that the population density used for validation was from villages, which does not correspond exactly to the location of sampling cells.

5.2. Comparisons

The available population datasets at 1 km resolution in China include the Landscan of Dobson et al. (2003), the results from Lo (2001) and the results form Yue et al. (2003). However, the outcome of Yue has not been validated and its digital version is unavailable.

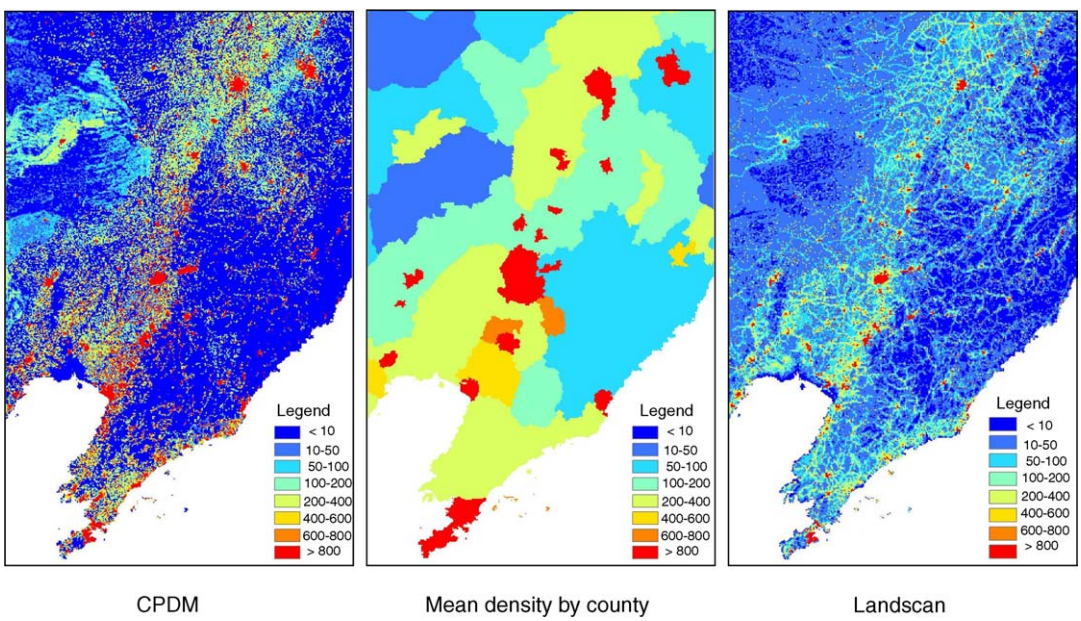

Fig. 6. Comparison of population density in Northeast of China.

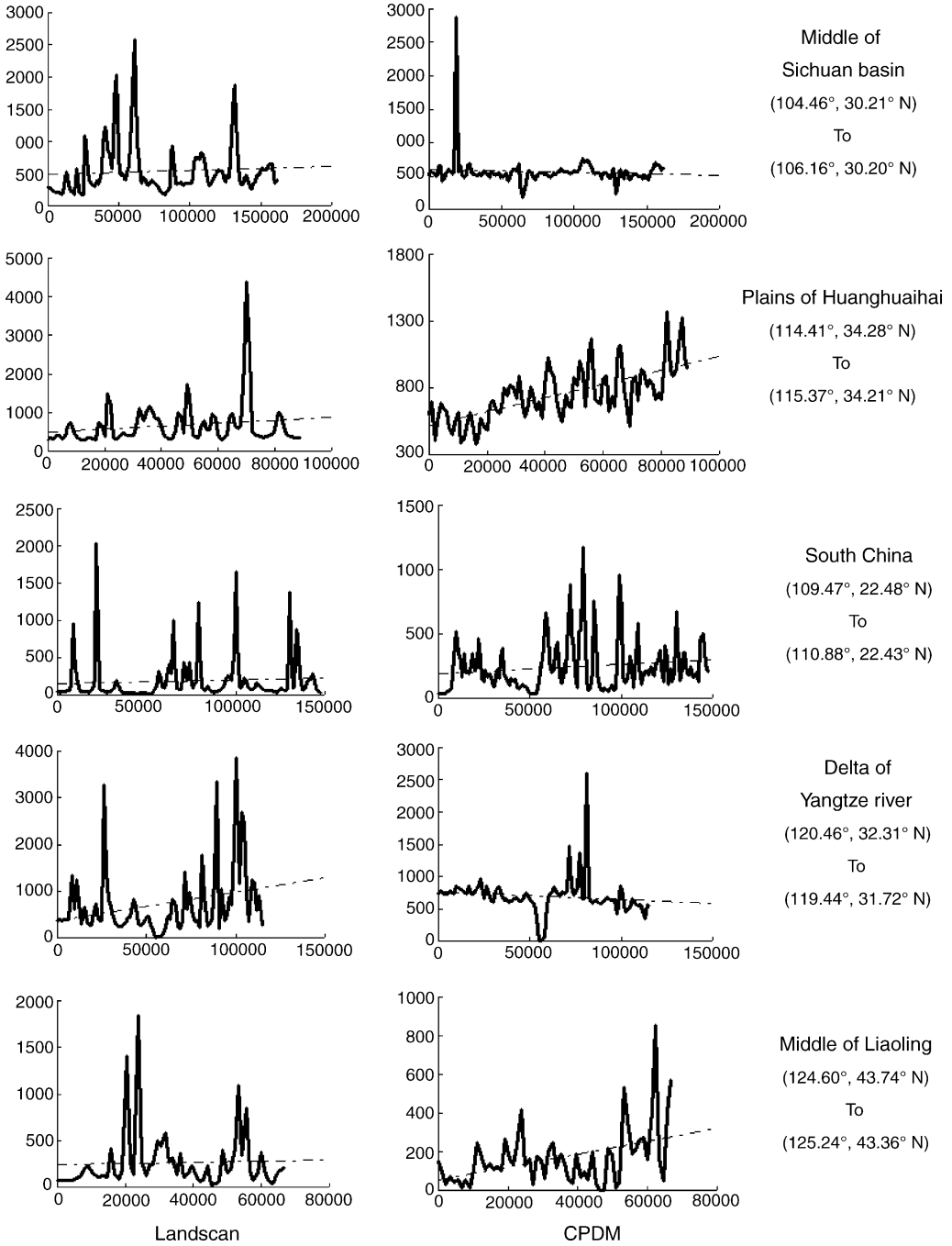

Fig. 7. Geographic profiles of population density in five populous regions of China.

Lo estimated only the non-agricultural population of China, with mean relative error of population density (in 17 cities) of 40.84% (Lo, 2001). Landscan is the best available population dataset for comparison with the outcome of this study (CPDM).

To compare the population of CPDM with Landscan, Landscan was re-projected to the Albers projection. To avoid "loosing people" when re-projecting Landscan, we first convert Landscan to points, re-project the points, rasterize the points to a high resolution, and lastly resample the points to the 1 km cell size.

The population of CPDM and Landscan are compared at three scales (sampling cells, county, and province, Tables 6 and 7). Table 6 shows that the percentage difference of CPDM in sampled cells is lower than that of Landscan in both rural and urban areas. A closer examination shows that the cells with higher differences are mainly located in the highway and the suburb areas of cites. The validation of Landscan also shows that the population of cells near transportation lines are overestimated and the population of suburbs are underestimated because the coverage of cities in Landscan is likely a large underestimate. Table 7 shows that the population errors of 1586 counties and 17 provinces in Landscan are more than 3%, which are respectively 67.37 and 54.84% of the counties and provinces of China. Additionally, 21.33% of counties and 6.45% of provinces have population errors are than 10% in Landscan. However, the errors in CPDM are zero at both county and province scales because of control of total population by county.

Fig. 6 demonstrates the difference between CPDM, Landscan and mean population density of counties in Northeast of China. Clearly, CPDM and Landscan contain a much more detailed population distribution than that of the mean density map. It also can be seen that the population of Landscan is more centralized in the transportation lines than that of CPDM, and the suburbs of cities have more population in CPDM than that in Landscan.

Six pairs of population profiles of CPDM and Landscan are plotted in Fig. 7. This figure shows that the ranges of population density in profiles of CPDM are less than that of Landscan, many of the cells with high values in Landscan are highways. It also can be noted from Fig. 7 that the rural population of CPDM is slightly higher than that of Landscan, especially in Sichuan basin.

6. Discussion and conclusions

CPDM is a dasymetric interpolation model. Its key function is to distribute census counts to cells based on population probability coefficients. Although many factors affect population distribution (elevation, slope, roads, may all be used as input variables to calculate probability coefficients), it was found that land cover is the best choice for population distribution modeling, because it is highly correlated with many other factors and includes most of the population distribution information.

The raster files of land cover types used in this study were converted from vector land cover polygons using the percentage breakdown encoding method, to avoid the loss of land information in land cover data with a single property (type). The weights of land types were determined by their stepwise regression coefficient, and also controlled by a qualitative geographical rule. The control of total population by rural and urban area at the county scale not only prevented population from being reallocated, but also rendered some national scale factors, such as main roads and cities, unnecessary for inclusion in the model. The difference between the factors influencing population distributions in urban and rural areas was addressed through the use of different modeling algorithms for urban and rural areas. Zonal modeling was used to efficiently reduce the geographic difference of land cover and customize model parameters for local regions. Based on a distance decay function, the model of urban population coefficients was built as a function of urban size and distance from urban center, which greatly simplifies the calculation. Although CPDM can elucidate basic patterns of urban population distribution, some details of the distribution caused by multi-centers or sub-centers are not taken into consideration.

The validation and comparison of the CPDM output shows that the simulation based on land cover is feasible and the outcome has high accuracy. Compared with other models, CPDM has only two input variables, census data and land cover data. It reduces the redundancy of information and thus avoids the overuse of information in the factors affecting population distribution. At the same time, it improves the simulation accuracy and the calculation efficiency. The method presented in this study is applicable to other countries, especially agricultural countries.

Further research is needed to assess whether improvements to the simulation are possible. For example, using a finer agro-ecological zoning may make the weights of land cover types more credible; more detailed information about urban spatial structure may improve the precision of urban population density estimates; and validating with the actual population in 1 km grid-cells can better depict the source of error in the model.

Acknowledgments

This work is supported by a Project of the National Natural Science Foundation of China (40371094) and the National Basic Research Priorities Program (2002CB412506) of the Ministry of Science and Technology of the People's Republic of China. The authors would like to thank Dr. Mingkui Cao for taking time to reivew this paper, the authors also thank the annoymous reviewers for theirs comments and suggestions on the earlier version of the manuscript.

References

Alexande, R.R., Shields, D.W., 2003. Using land as a control variable in density-dependent Bioeconomic models. Ecol. Model. 170, 193–201.

Batty, M., Longley, P.A., 1994. Fractal Cities: A Geometry of Form and Function. Academic Press, London.

Cao, M.K., Ma, S.J., Han, C.R., 1995. Potential productivity and human carrying capacity of agro-ecosystems: an analysis of food production potential of China. Agric. Syst. 47, 387–414.

Chen, B.M., 2001. Integrated Productivity and Carrying Capacity of Agricultural Resources in China. Weather Press, Beijing (in Chinese).

Chen, Y.G., 1999. The fractal model of urban population density. J. Xinyang Normal College 12, 60–64 (in Chinese).

Clark, C., 1951. Urban population densities. J. Roy. Stat. Soc. 114, 490–496.

Clarke, J.T., Rhind, D.W., 1992. Population data and global environmental change. Human Dimensions of Global Environmental Change Programme, Report No. 3. International Social Science Council, Paris.

Demers, M.N., 1997. Fundamentals of Geographic Information System. Wiley, New York.

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., Worley, B.A., 2000. LandScan: a global population database for estimating populations at risk. Photogrammet. Eng. Remote Sens. 66, 849–857.

Dobson, J.E., 2003. Estimating Populations at Risk. In: Geographical Dimensions of Terrorism. Routledge, New York and London, pp. 161–167.

Dobson, J.E., Bright, E.A., Coleman, P.R., Bhaduri, B.L., 2003. LandScan2000: a new global population geography. In: Remotely-Sensed Cities. Taylor and Francis, London, pp. 267–279.

Elvidge, C.D., Baugh, K.E., Kihn, E.A., Kroehl, H.W., Davis, E.R., 1997. Mapping city lights with nighttime data from the DMSP operational linescan system. Photogrammet. Eng. Remote Sens. 63, 727–734.

Feng, J., 2002. Modeling the spatial distribution of urban population density and its evolution in Hangzhou. Geogr. Res. 21, 635–646 (in Chinese).

Hu, H.Y., 1983. Study on the Distribution of Population in China. East China Normal University Press, Shanghai (in Chinese).

Kokko, H., Lindstroem, J., 1998. Seasonal density dependence, timing of mortality, and sustainable harvesting. Ecol. Model. 110, 293–304.

Li, J., Zhuang, D.F., 2002. Analysis of feasible scales of spatial data. Acta Geogr. Sinica 57 (Suppl.), 52–59 (in Chinese).

Liao, S.B., Sun, J.L., 2003. GIS-based spatialization of population census data in Qinghai–Tibet plateau. Acta Geogr. Sinica 58, 25–33 (in Chinese).

Liu, J.Y., 1996. Study on the Micro Survey of Chinese Resources and Environment by Remote Sensing and its Dynamic. Weather Press, Beijing (in Chinese).

Liu, J.Y., Buhe, A.S., 2000. Study on spatial-temporal feature of modern land-use change in China: using remote sensing technique. Quaternary Sci. 3, 229–239 (in Chinese).

Liu, J.Y., Zhuang, D.F., 2003. Land cover classification of China: integrated analysis of AVHRR imagery and geographical data. Int. J. Remote Sens. 24, 2485–2500.

Liu, M.L., Tang, X.M., Liu, J.Y., 2001. Research on scale effect of spatial data of 1 km grids. J. Remote Sens. 5, 183–189 (in Chinese).

Lo, C.P., 2001. Modeling the population of China using DMSP operational linescan system night-time data. Photogrammet. Eng. Remote Sens. 67, 1037–1047.

McCarthy, M.A., Lindenmayer, D.B., 1998. Population density and movement data for predicting mating systems of arboreal marsupials. Ecol. Model. 109, 193–202.

Miller, D.H., Jensen, A.L., Hammill, J.H., 2002. Density dependent matrix model for gray wolf population projection. Ecol. Model. 151, 271–278.

Newling, B.E., 1969. The spatial variation of urban population densities. Geogr. Rev. 59, 242–252.

Sekimura, T., Roose, T., Li, B., Maini, P.K., Suzuki, J., Hara, T., 2000. The effect of population density on shoot morphology of herbs in relation to light capture by leaves. Ecol. Model. 128, 51–62.

Sutton, P., 1997. Modeling population density with night-time satellite imagery and GIS, computers. Environ. Urban Syst. 21, 227–244.

Sutton, P.C., Elvidge, C., Obremski, T., 2003. Building and evaluating models to estimate ambient population density. Photogrammetr. Eng. Remote Sens. 69 (5), 545–553.

Shen, G.Q., 2002. Fractal dimension and fractal growth of urbanized areas. Int. J. Geogr. Information Sci. 16, 419–437.

Sherratt, G.G., 1960. A model for general urban growth. In: Management Sciences, Model and Techniques, Proceedings of the Sixth International Meeting of Institute of Management Sciences, vol. 2, Pergamon Press, Elmsford, NY, pp. 147–159.

Smeed, R.J., 1961. The Traffic Problem in Towns. Manchester Statistical Society, Manchester.

Tanner, J. C., 1961. Factors Affecting the Amount Travel. Road Research Technical Report No. 51. HMSO, London, pp. 15–220.

Tian, Y.Z., Qiu, D.C., Yang, Q.Y., Yin, W., 2003. An allocation model of urban land price monitoring sites. J. Southwest China Normal Univ. (Nat. Sci.) 28, 313–323 (in Chinese).

Tobler, W., Deichmann, U., Gottsegen, J., Maloy, K., 1997. World population in a grid of spherical quadrilaterals. Int. J. Population Geogr. 3, 203–225.

Wang, F.H., Michel, G.J., 1996. Simulating urban population density with a gravity-based model. Socio-Economic Plann. Sci. 30, 245–256.

William, H.G., 2000. Economics Analysis. Prentice-Hall, New Jersey.

Wright, J.K., 1936. A method of mapping densities of population with cape cod as an example. Geogr. Rev. 26, 103–110.

Yang, C.M., Wang, S., 2002. Study on the choice of path for agriculture system in China. Finance Res. 9, 12–22 (in Chinese).

Yang, X.H., Jiang, D., Wang, N.B., 2002. Method of pixelizing population data. Acta Geogr. Sinica 57 (Suppl.), 71–75 (in Chinese).

Ye, Y.X., 2001. City and optimization of land use in 21 century. China Land Sci. 15, 10–13 (in Chinese).

Yue, T.X., Wang, Y.A., Liu, J.Y., Chen, S.P., Qiu, D.S., Deng, X.Z., Liu, M.L., Tian, Y.Z., Su, B.P., 2005a. Surface modelling of human population distribution in china. Ecol. Model. 181 (4), 461–478.

Yue, T.X., Wang, Y.A., Liu, J.Y., Chen, S.P., Tian, Y.Z., Su, B.P., 2005b. MSPD scenarios of spatial distribution of human population in China. Population Environ. 26 (3), 207–228.

Yue, T.X., Wang, Y.A., Chen, S.P., Liu, J.Y., Qiu, D.S., Deng, X.Z., Liu, M.L., Tian, Y.Z., 2003. Numerical simulation of population distribution in China. Population Environ. 25 (3), 249–271.

Zhang, C., Yang, B.G., 1992. Basic of Quantitative Geography. Higher Education Press, Beijing (in Chinese).

Zhang, S.Y., 1997. China's Population Geography. East China Normal University Press, Shanghai (in Chinese).

Zhong, D.J., 2001. Misplay of China: influence of public domain system and residence-registered system on land resources. New Econ. 3, 18–22 (in Chinese).

Zhou, C.S., Xu, X.Q., 1997. Population distribution and its change in Guangzhou city. Trop. Geogr. 17, 53–60 (in Chinese).