Computers, Environment and Urban Systems 41 (2013) 167–176

Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

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

Estimating the spatial distribution of the population of Riyadh, Saudi Arabia using remotely sensed built land cover and height data

Mohammed Alahmadi [a,b,*], Peter Atkinson [a], David Martin [a]

[a] Geography and Environment, University of Southampton, Southampton SO17 1BJ, UK
[b] Space Research Institute, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

ARTICLE INFO

Article history:
Received 18 November 2012
Received in revised form 12 June 2013
Accepted 12 June 2013
Available online 8 July 2013

Keywords:
Population
Dwelling
Land cover
Estimation
Riyadh

ABSTRACT

This paper investigates the use of Landsat ETM+, remotely sensed height data, ward-level census population, and dwelling units to downscale population in Riyadh, Saudi Arabia. Regression analysis is used to model the relationship between density of dwelling units and built area proportion at the block level and the coefficients used to downscale density of dwelling units to the parcel level. The population distribution is estimated based on average population per dwelling unit. Seven models were fitted and compared. First, a conventional approach, using ISODATA-classified built land cover alone as a covariate, is used as a benchmark against which to evaluate six more sophisticated models. The conventional model results in low accuracy measured by overall relative error (ORE) (+116%). Approaches for potentially increasing accuracy are explored, incorporating above-surface height data into the downscaling process. These include masking out zero and near-zero height areas when estimating built area; using height to estimate the number of floors; replacing the ISODATA model with a support vector machine; estimating volume-adjusted habitable space; stratifying the study area into different building categories; and preservation of the dependent variable for the best model. These approaches result in large increases in accuracy in the density of dwelling unit estimates. However, while the height data accounts for the vertical dimension (primarily through the number of floors), it is not possible to account for variation in dwelling density which arises due to other factors such as living standards, affluence and other spatially varying factors, without further data.

1. Introduction

Growth in the human population is exerting pressure on global resources and the environment, creating an urgent need for sustainable development (Li & Weng, 2005). Timely and accurate information about the size and spatial distribution of population in urban areas has become important for understanding and responding to social, economic and environmental problems. Population data, from national to census block level, are essential for purposes such as urban and regional planning, resource management, emergency planning and utility and service allocation (Dobson, Bright, Coleman, Durfee, & Worley, 2000; Sutton, Roberts, Elvidge, & Baugh, 2001). Unfortunately, population data in Saudi Arabia, including Riyadh, are only available at a coarse spatial resolution (ward level, mean population 30,000).

Traditionally, data on demographic characteristics and socio-economic status are obtained from censuses, field surveys and administrative registers (Rhind, 1991). These data sources, however, are not usually recorded annually and in developing countries, may not be available at all. The challenges of regular updates, long census intervals, spatial aggregation and boundary designation problems connected have encouraged researchers to develop alternative techniques for estimating population (Li & Weng, 2005; Liu & Herold, 2007).

Since the 1950s, many methods of population estimation have been developed using remotely sensed satellite sensor data and aerial photography. According to Wu, Qiu, and Wang (2005) there are two categories of population estimation: areal interpolation and statistical modelling. The former uses census population as the input to the estimation process (Holt, Lo, & Hodler, 2004; Langford & Unwin, 1994), whereas the latter uses socio-economic covariates and a model of the relation between census population and those covariates. Sutton (1997) examines theoretical models of urban population density decay functions based on a circular city to irregularly shaped clusters identified from the Defense Meteorological Satellite Program (DMSP) night-time satellite sensor imagery. He uses distance from the edge of a cluster instead of an urban

* Corresponding author at: Geography and Environment, University of Southampton, Southampton SO17 1BJ, UK. Tel.: +44 (0) 23 8059 4617; fax: +44 (0) 23 8059 3295.

E-mail addresses: mh_alahmadi@yahoo.com (M. Alahmadi), P.M.Atkinson@soton.ac.uk (P. Atkinson), D.J.Martin@soton.ac.uk (D. Martin).

0198-9715/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2013.06.002

centre. Sutton (1997) reports that these simple theoretical models need very little input information and could be a starting point for development of a more accurate method of modelling population density.

This paper concentrates on statistical methods (regression) to predict population counts. Specifically, information on urban extent derived from satellite data is aggregated to the zone level of the reference population and provides explanatory variables to estimate the population.

Studies such as Wellar (1969), Lo and Welch (1977) and Lo (2003) use the allometric growth method. Lo (2003) applies the ISODATA algorithm to Landsat TM imagery to classify low-density urban land use. In that study, population counts and numbers of dwelling units are estimated with overall relative errors of 0.23% and 2.4% respectively.

Langford (2006) provides a detailed examination of the relationship between population and two different land cover schemes in Leicestershire. The analysis is conducted at ward level, population being areally interpolated to enumeration districts (EDs). Global and regional regression models are compared over the two schemes. He concludes that the regional model produce more accurate results. He also explores 3-class dasymetric mapping, determining density ratios through selective sampling, global and regional regression. It is concluded that density calibration through regression are more accurate than from selective sampling. 3-class dasymetric mapping provides more accurate results than the regional model, being more highly parameterised and locally fitted. However, it is less accurate than binary dasymetric mapping.

Three-dimensional information (e.g. height data) has been proposed as important to help distinguish between uses such as buildings and non-buildings within the urban land use class (e.g. villas, apartments). Studies such as Charaniya, Manduchi, and Lodha (2004); Secord and Zakhor (2007), Antonarakis, Richards, and Brasington (2008) use Light Detection and Ranging (LiDAR) data to refine the classification process. Remotely sensed height data from LiDAR data or stereo images (Binard, Devriendt, Cornet, & Donnay, 2006) can potentially increase the accuracy of population estimation. Lu, Im, Quackenbush, and Halligan (2010) use QuickBird and LiDAR data at census block level in Denver to estimate population using both area-based and volume-based approaches. Both produce promising results, but the area-based approach is slightly more accurate. Similar studies, including Dong, Ramesh, and Nepali (2010) and Ural, Hussain, and Shan (2011), demonstrate the benefits of utilizing remotely sensed height data.

Lo (2008), using four land covers derived from Landsat data for Atlanta, reports that regional and GWR models are more accurate than a global model. Dong et al. (2010) use Landsat and LiDAR data for Denton to provide small-area population estimates at the block level. Three explanatory variables are used: the number, area and volume of buildings over continuous and random samples. Ten OLS and GWR models are compared, showing that random samples provide more accurate results than continuous samples (Dong et al., 2010).

Hsu (1973) is the first use of Landsat multispectral scanning system (MSS) data (1×1 km grid) with a multiple regression model to estimate population distribution. Li and Weng (2005) examine a variety of different covariates such as principal components, vegetation indices, fraction images, texture surfaces and a temperature surface. Lo (1995) and Harvey (2002) report that most low density areas are overestimated and most high density areas underestimated. Lo (2003) and Li and Weng (2005) stratify the study area into different densities based on the areal units, which are insufficient to explain the real densities being mapped. For example, if a ward contains blocks of flats and this ward is not fully developed, the average density will not reflect the actual density.

The above studies demonstrate that remotely sensed information on built area distribution can be useful in predicting population distribution, and that building height data can add valuable explanatory information to the usual land cover/use class covariates, when using regression-type approaches. However, the optimal mode for introducing height information (e.g. height, volume, number of floors) is not yet established. Moreover, these studies tend to focus on the western world, and the value of such height information for a rapidly urbanizing city in a desert environment has not been explored. The city of Riyadh, Saudi Arabia is interesting because population data do not exist at a fine spatial resolution in the national census. Given the rapid rate of urbanization, there is an urgent need to find suitable methods for mapping and monitoring the population distribution.

This paper focuses on detailed population mapping in residential areas based on input data about residential buildings. Detailed population mapping is important in decisions about where to build public facilities such as schools, mosques, health centres and parks and in determining public transport routes and the locations of private businesses.

The aim of this paper is to provide small-area population estimates using remotely sensed data for a selected ward of Riyadh called Um Alhamam. First, a conventional model of population estimation is applied. The strategy is to analyse the density of dwelling units and then use a simple empirical factor to transform this into population estimates. The conventional model is developed by introducing height data in various forms and comparing the results. Of secondary interest is to examine how different classification algorithms affect the results.

The remainder of this paper is organized as follows: methods are provided in the next section, covering image classification, regression through the origin and accuracy assessment. The study area and data are described in the third section, while the production of the response and explanatory variables is explained in the fourth section. Analysis of the different models is reported in the penultimate section. Finally, a discussion and conclusion are presented.

2. Methods

2.1. Image classification

Two classification algorithms are used in this research: the Iterative Self-Organizing Data Analysis Technique (ISODATA) and support vector machine (SVM). ISODATA is a well-known unsupervised algorithm which does not use training samples as the basis of classification. Unknown pixels are classified into natural groupings or clusters. Detailed discussions may be found in Jensen (2005) and Campbell (2006).

A detailed explanation of the SVM algorithm is beyond the scope of this paper, but greater explanation can be found in Vapnik (1995) and Haikin (1998). SVM is a supervised classification method based on statistical learning theory. It distinguishes classes of interest with a decision surface (optimal hyperplane) that maximizes the margin between the classes. To understand SVM, four concepts are needed (Noble, 2006): (i) the hyperplane: SVM constructs a hyperplane which can be used for classification and regression, (ii) the maximum-margin hyperplane: the best separation is obtained by the hyperplane that has maximum distance from the nearest sampling points (known as support vectors), (iii) the soft margin: the sampling points are not always linearly separable so one approach is to allow misclassification by modifying the soft margin and (iv) the kernel function: when the sampling data are hard to separate linearly, one solution is to use kernels that allow data to be linearly separable by