International Journal of Geographical Information Science
Vol. 22, No. 4, April 2008, 431–447

Research Article

Population-density estimation using regression and area-to-point residual kriging

X. H. LIU*†, P. C. KYRIAKIDIS‡ and M. F. GOODCHILD‡

†Department of Geography & Human Environmental Studies, San Francisco State University, 1600 Holloway Avenue, HSS 279, San Francisco, CA 94132, USA
‡Department of Geography, Ellison Hall, University of California, Santa Barbara, Santa Barbara, CA 94106-4060, USA

(*Received 14 March 2005; in final form 31 January 2007*)

Census population data are associated with several analytical and cartographic problems. Regression models using remote-sensing covariates have been examined to estimate urban population density, but the performance may not be satisfactory. This paper describes a kriging-based areal interpolation method, namely area-to-point residual kriging, which can be used to disaggregate the residuals remaining from regression. Compared with conventional cokriging, the area-to-point residual kriging is much simpler in that only a semivariogram model for the point residuals is required, as opposed to a set of auto- and cross-semivariogram models involving the dependent variable and all the covariates. In addition, area-to-point residual kriging explicitly accounts for any scale differences between source data and target values. The method is illustrated by disaggregating population from census units to the land-use zones within them. Comparative results for regression with and without area-to-point residual kriging show that area-to-point residual kriging can substantially improve interpolation accuracy.

Keywords: Areal interpolation; Dasymetric mapping; Kriging; Geostatistics; Population surface

1. Introduction

Knowledge of the size and spatial distribution of human population in an urban area is essential for social, economic, and environmental applications. Traditionally, census data are the primary source of information on population distribution. Census data are usually reported as spatial aggregates for census zones, such as census tracts or census block groups. The problems of applying zone-based census data to geographical analysis are well documented in the literature. One is the modifiable areal unit problem (MAUP), which refers to the situation where the selection of the areal units or scales can significantly change the results (Openshaw 1984). Another problem is that a population may not be distributed uniformly within a census unit if the land use is heterogeneous; hence the spatial pattern provided by conventional choropleth mapping may not be accurate (Monmonier and Schnell 1984). Additionally, the boundaries of census units rarely coincide with those of other data-collecting units (e.g. school districts or watersheds), thus creating difficulty in spatial data integration (Goodchild *et al.* 1993).

*Corresponding author. Email: xhliu@sfsu.edu

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2008 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810701492225

To tackle these problems, areal interpolation is often utilized. Areal interpolation is designed to transform data from source zones to target zones. In the context of population distribution, census units such as census block groups or census tracts usually serve as the source zones; the target zones are typically grid cells or land-use zones. When ancillary information is used, areal interpolation is also referred to as dasymetric mapping (Wright 1936). Many areal interpolation methods have been developed and can be categorized as simple or intelligent depending on whether ancillary information is used (Okabe and Sadahiro 1997). Simple interpolation does not use any data other than the source-zone population. An example is areal weighting which allocates population according to the areal proportion of a target zone in the host source zone. Goodchild *et al.* (1993) discussed its implementation for various socio-economic variables. Another example is Tobler's pycnophylactic (mass-preserving) interpolation and its modification which creates a smooth population-density surface (Tobler 1979, Rase 2001). Bracken and Martin (1989) also described a centroid-based method which uses a kernel-based technique to disaggregate census data to grid cells. Kyriakidis (2004) proved that both these approaches can be viewed from a geostatistical perspective, and correspond to particular choices of a semivariogram model for an underlying (latent or unobserved) density surface.

Human population distribution is closely related to other information on the Earth such as land use and transportation facilities. These data can therefore be used as ancillary information to assist population interpolation. Wright (1936) used a land-use map to identify likely areas of denser or sparser population, and then allocated the population from towns to the land-use zones within them. This idea is still used in areal interpolation, except that today the ancillary information is mainly derived from remotely sensed images. Langford and Unwin (1994) and Mennis (2003) both used the land-use classification information from Thematic Mapper (TM) images. Harvey (2002) established a direct association between population and the spectral reflectance values of TM pixels. Wu and Murray (2005) used an ETM+ image for estimating urban population density, but the information used was the fraction of impervious surface in residential areas. Although TM and ETM+ images provide valuable information to improve population estimation, their 30-m spatial resolution limits their utility in urban applications. For the purpose of detailed population-density estimation, a spatial resolution of 0.5–5 m is recommended (Jensen and Cowen 1999). In the past, such images could only be obtained through low-altitude aerial photography. The advent of very-high-spatial-resolution satellite sensors such as IKONOS has recently offered new opportunities. However, to date, little research has explored these opportunities in the context of population interpolation.

The direct output of many methods reviewed above is a raster surface; others use zones like those in Wright's (1936) dasymetric mapping. The advantage and shortcomings of handling population data as a raster surface versus vector zones have been discussed by Martin (1996). Surface output is preferred to facilitate spatial data integration, since a 'cookie-cutter' can be used to aggregate cell-level population to any desired areal unit. However, vector zones in some interpolation methods can also be easily converted to a grid using GIS as long as the population within a target zone is uniformly distributed. Zonal-output methods are widely used when remote-sensing information is available. They are relatively easier to understand and less computationally intensive compared with surface-based interpolation. The method to be discussed in this paper belongs to this group.

The purpose of this paper is to report an experiment on using a zone-based interpolation method to disaggregate census population so that more detailed population-density estimates can be obtained. In particular, the paper demonstrates the utility of residual modelling to improve interpolation accuracy, thus enhancing population-density estimation obtained through popular regression models. More precisely, the residual population density values obtained from a regression model are disaggregated using area-to-point kriging (Kyriakidis 2004). This approach is simpler than conventional cokriging and does not require any additional data beyond those used in the regression model. To date, many zone-based methods available for estimating population density with remote-sensing data have focused on finding more powerful remote-sensing covariates. This research suggests that area-to-point residual kriging may be a worthwhile supplement.

The paper is organized as follows. Section 2 describes the study area and data. The initial efforts to estimate population density using data extracted from an IKONOS image are documented in section 3. Section 4 describes area-to-point residual kriging, which makes full use of the location and spatial autocorrelation information embedded in the source zone residuals. The interpolation results and their accuracy assessment are presented in section 5. Results for population-density estimation with and without area-to-point residual kriging are compared to draw the conclusions in section 6.

2. Study area and data sources

The goal of this research is to obtain more detailed population-density estimates than those obtained using only census data. The city of Santa Barbara in California and its vicinity is used as the study area. The region is located 170 km north-west of Los Angeles in the foothills of the California Coast Range. It is about 300 km^2 in area and includes a total population of about 100 000. The region is characterized by various types of land use, including residential areas with variable housing density and socio-economic structure, commercial and industrial districts, and open spaces (e.g. farm land and wetland). Population data at the census block level were acquired from the 2000 census. Errors and positional inaccuracy in the census data were rectified, based on a parcel-level dataset from Santa Barbara County, which had parcel boundaries and building footprints (shape and position) information.

To assist population interpolation, remote sensing was used to obtain ancillary information. Seven multispectral IKONOS images acquired in 2001 between March and July were mosaicked to cover the entire study area. Because the image acquisition dates had varying atmospheric and illumination conditions, geometric and atmospheric corrections were conducted to create a geometrically rectified and normalized image. Details on the preprocessing of the IKONOS image are described by Herold *et al.* (2002). The mosaicked IKONOS image was visually interpreted into land-use regions by an experienced image analyst who was familiar with the study area. Because the spatial resolution of IKONOS images is comparable with that of low-altitude aerial photos, the principles of aerial photo interpretation were applied to the delineation of land-use regions. Each land-use zone is equivalent to a 'photomorphic region' in visual aerial-photo interpretation (Peplies 1974, Barnsley and Barr 1997), which is an image segment with homogenous size, shape, tone/colour, texture, pattern, etc. Each land-use region has a single *land use* (e.g. residential, commercial, agriculture, etc.) but the pixels it hosts can be of three *land-cover* types: built-up, vegetation, and others. The zone-level land-use and pixel-level land-cover information was evaluated and found to be highly accurate (Herold *et al.* 2003). Figure 1 is an example of the land-use regions found in a residential area. Note that the high spatial resolution of the IKONOS image enabled not only the recognition of residential land use but also the differentiation of size, architecture, and age of houses. With no additional information, it is reasonable to assume that the population density of each land-use region is uniform.

3. Regression-based interpolation using remote-sensing information

The goal is to disaggregate the population of a census unit to the land-use zones within it, or equivalently estimate their population densities. One method is to use equation (1) given below, where population is a weighted sum of density values within each land-use category:

$$P_i = \sum_j d_j A_{ij} \tag{1}$$

where P_i is the population of census unit i, d_j is the population density of land-use type j, and A_{ij} is the area of land-use type j in census unit i. d_j can be obtained using different methods, such as empirical sampling (Kraus and Senger 1974) or predefined population-density statistics (Mennis 2003). Another often-used approach is to conduct linear regression based on equation (1) (Langford *et al.* 1991). Though simple, the success of this method depends on the accuracy and degree of detail of the land-use classification. More detailed classification usually helps to improve the accuracy (Donnay and Unwin 2001). In this study, seven residential land-use classes were identified: high-density single-unit residential (HSU), medium-density single-unit residential (MSU), low-density single-unit residential (LSU), multiple-unit residential (MU), mobile homes (mobile), mixed single-and-multiple unit residential (S&M), and mixed residential-commercial (mixed) land use. The definition and illustration of these land-use types are described by Herold *et al.* (2003). Linear regression based on equation (1) was applied, and the following model was obtained:

$$\begin{aligned} P_i = {} & 1024.48A_{\mathrm{LSU}} + 3325.16A_{\mathrm{MSU}} + 5973.49A_{\mathrm{HSU}} + 10180.09A_{\mathrm{MU}} \\ & + 5063.15A_{\mathrm{mobile}} + 8826.56A_{\mathrm{S\&M}} + 7593.45A_{\mathrm{mixed}} \end{aligned} \quad R^2 = 0.44 \tag{2}$$

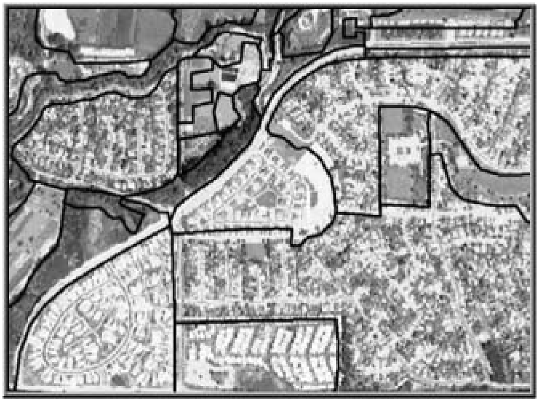

Figure 1. Example of land-use regions in the IKONOS image. Note that the very high spatial resolution of IKONOS enables differentiation in the residential area.

where the densities are in units of persons per square kilometre. The regression line was forced to pass through the origin so that the population was allocated to residential land-use categories only.

The poor proportion of variance explained by equation (2) suggests that the linear-regression approach did not perform well despite the highly detailed land-use classification. One major reason for this is that the population density of a land-use category is not the same across the entire study area. Considering that the remotely sensed image has a much richer biophysical information than that captured by land-use classification, other remote-sensing surrogates of human population have been examined in previous research. Harvey (2002) found a high correlation between census district population and the multi-band spectral reflectance data of a TM images. Wu and Murray (2005) found the fraction of impervious surface extracted from an ETM+ images useful to infer urban population density. In this study, the land-use zones were delineated from the IKONOS image according to the homogeneity of the tone, shape, colour, size, pattern, etc. This information should help explain the variation in population density. Previous research has identified landscape metrics as an efficient tool to quantitatively characterize a land-use zone (Liu and Herold 2007). Landscape metrics were developed in landscape ecology to describe the composition and pattern of a landscape. A detailed discussion on landscape metrics and their implementation in GIS are described by McGarigal *et al.* (2002). For the purpose of population-density estimation, various landscape metrics have been examined in terms of their association with population density. The results are documented by Liu *et al.* (2006) and can be summarized in the following model for this region:

$$\ln\left(\widehat{d}\right) = 8.819 + 1.772 p_1 - 2.612 p_2 + 0.0632 p_3 \quad R^2 = 0.55 \tag{3}$$

where $\widehat{d}$ is the estimated population density of a land-use zone, p_1 and p_2 are the areal proportions of the built-up and vegetation area within it, and p_3 is the patch density of the built-up area. Patch density is a landscape metric computed from patches that are formed by a set of contiguous built-up pixels. The number of built-up patches divided by the total area of a land-use zone is the patch density of the built-up area, i.e. p_3 in equation (3). More details on equations (2) and (3) are provided by Liu and Herold (2007).

Although equation (3) achieved a higher regression coefficient than equation (2), it seems only marginally more promising as a basis for estimating population density. In fact, it was found through additional experiments that only 40–60% of the variance in population density can be explained by regression based on covariates derived from remote sensing data. Clearly, regression modelling alone is not sufficient. Other methods must be explored to account for the substantial residuals that remain after regression.

4. Residual interpolation using area-to-point kriging

The information used by the regression models in equations (2) and (3) concerns a single land-use zone only. Other information embedded in the source data is not utilized. One example of such information is the location of a land-use zone—a residential area surrounded by high-population-density land-use is likely to have a high density as well. Another is the relationship between a census unit and the land-use zones within it—the population of a census unit is the sum of that of the land-use zones within it. This information is not utilized by regression modelling, which probably explains the significant variance of the resulting residuals. In search of a method to account for this information, kriging was selected. Kriging is a geostatistical method for spatial prediction utilizing information on the spatial auto-correlation (and cross-correlation in the case of cokriging) of different attributes, i.e. the correlation of a variable with itself (or other variables in the case of cokriging) through space (Isaaks and Srivastava 1989). Cokriging is especially useful when the variable to be estimated (e.g. the population density of the land-use zones) is under-sampled but when abundant ancillary information is available (e.g. the population density of a census unit). Geostatistical methods have been widely used in natural-resource management and remote sensing (Goovaerts 1997, Curran and Atkinson 1998), but their application in areal interpolation is rather limited. The studies by Lam (1983) and Wu and Murray (2005) are probably the only examples in the context of population-density estimation.

4.1 *Area-to-point kriging*

The kriging method used in this study was proposed by Journel (1999) under the name cokriging and discussed in detail by Kyriakidis (2004) in the context of areal interpolation under the name area-to-point kriging. We prefer the latter term, because we reserve the term cokriging for cases whereby the source data and sought-after target values pertain to different attributes. Kriging and its multivariate extension cokriging were originally developed for mining applications. They are based on the regionalized variable theory where the value of a variable $d(u)$ at point with coordinate vector u is considered a realization of a random variable $D(u)$. The collection of (infinitely many) spatially correlated random variables $\{D(u), u \in A\}$, where A denotes the study region, is termed a random function (RF). $D(u)$ consists of two components: a deterministic component $m(u)$, which indicates the geographical trend or drift, and a stochastic zero-mean residual component $R(u)$, which is auto-correlated in space:

$$D(u) = m(u) + R(u) \tag{4}$$

Given a lag vector h, i.e. a vector with specified distance and direction, the expected value of the difference between $R(u)$ and $R(u+h)$ is 0. The variance of the difference is given by

$$Var\{[R(u+h) - R(u)]\} = E\left\{[R(u+h) - R(u)]^2\right\} = 2\gamma_R(h) \tag{5}$$

where $\gamma_R(h)$ is called the semivariogram of residuals.

Let $C_R(0)$ denote the variance of $R(u)$, i.e. $C_R(0)=\text{var}\{R(u)\}$. The covariance between $R(u)$ and $R(u+h)$ is obtained as

$$C_R(h) = E\{R(u+h) \cdot R(u)\} = C_R(0) - \gamma_R(h) \tag{6}$$

In this paper, we perform kriging with two types of data pertaining to the same attribute: the target residual $r(u)$, which is of point support and partially sampled, and the source residual $r(v_u)$, which provides ancillary information and is defined on an areal support. We call this kriging variant area-to-point residual kriging. In conventional kriging, both the source data and target values are of point support and pertain to the same attribute, whereas in conventional cokriging, the source data and target values are also of point support but pertain to different attributes. The kriging variant used in this research can also be regarded as a cokriging variant, whereby the primary variable represents the unknown residuals at target zones, and the secondary attribute represents known residuals at source zones. As noted before, we do not use the term cokriging, because we are dealing with target values and source data of the same (residual) attribute.

In area-to-point residual kriging, any source residual $r(v_u)$ is functionally linked to the point residuals within it as:

$$r(v_u) \simeq \sum_{i=1}^{N_u} \omega_{u_i} r(u_i), \quad u_i \in v_u \text{ and } \sum_{i} \omega_{u_i} = 1 \tag{7}$$

where N_u is the number of points within an areal unit v_u, and ω_{u_i} is the weight associated with $r(u_i)$ and is assumed to be known; it is the areal proportion of a land-use zone within a census tract (see below). The above equation implies the following assumptions: (1) point residuals are defined at the centroids of land-use zones within a census tract, and (2) a point residual is representative of the entire land-use zone within which it is located. These two assumptions imply that we ignore the scale difference between a point and a target land-use zone. In other words, we assume that the population within such zones is homogeneous.

Under this framework, the variance of the source residual random variable $R(v_u)$ is given by

$$C_R(v_u, v_u) = Cov\{R(v_u), R(v_u)\} \simeq \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} \omega_{u_i} \omega_{u_j} C_R\left(u_i - u_j\right), u_i \in v_u, u_j \in v_u \tag{8}$$

and the covariance between a point residual random variable $R(u)$ and an area residual random variable $R(v_u)$ is given by

$$C_R(u, v_u) = C_R(v_u, u) = Cov\{R(u), R(v_u)\} \simeq \sum_{i=1}^{N_v} \omega_{u_i} C_R(u - u_i), u_i \in v_u \tag{9}$$

In area-to-point residual kriging, the unknown residual value $r(u)$ within a source zone v_u is estimated using the corresponding source residual $r(v_u)$ and n nearby point residuals $\{r(u_j), j=1, \ldots, n\}$ at other land-use zones within or outside v_u as

$$r^*(u) = \sum_{j=1}^{n} \lambda_j(u) r\left(u_j\right) + \lambda_0(u) r(v_u) \tag{10}$$

where $\lambda_j(u)$ is the weight assigned to $r(u_j)$ and $\lambda_0(u)$ is the weight assigned to $r(v_u)$. These weights can be calculated by solving the following system of equations:

$$\begin{cases} \sum\limits_{j=1}^{n} \lambda_j(u) C_R\left(u_j - u_i\right) + \lambda_0(u) C_R(v_u, u_i) = C_R(u - u_i), \quad i = 1, \ldots, n \\ \sum\limits_{j=1}^{n} \lambda_j(u) C_R\left(u_j, v_u\right) + \lambda_0(u) C_R(v_u, v_u) = C_R(u, v_u) \end{cases} \tag{11}$$

The weights estimated by this method are optimal in the sense that the variance of the error between the true and the estimated value is minimized. Note that equations (10) and (11) constitute the simple cokriging estimate and system, respectively; there are no mean terms in equation (10) because the residuals are assumed to have zero mean (per regression). For a detailed discussion on area-to-point kriging, please refer to Kyriakidis (2004).

4.2 *Residual semivariogram modelling*

Equation (11) shows clearly that solving for the weights $\lambda_0(u)$ and $\lambda_j(u)$ requires the values of the residual covariance terms. In conventional cokriging, their computation requires the semivariograms of the primary and secondary variable and the cross-semivariogram between them (Goovaerts 1997). However, in the case of two variables functionally related according to equation (7), the variance of a source residual random variable (e.g. $C_R(v_u, v_u)$), as well as the covariance between a target and a source residual random variable (e.g. $C_R(u, v_u)$) are completely specified in terms of the covariance model of the regression residuals (equations (8) and (9), i.e. in terms of the point covariance model $C_R(h)$. Per equation (6), the covariance model $C_R(h)$ of such residuals can be obtained from their semivariogram model $\gamma_R(h)$. An empirical semivariogram can be computed from the residual population density data to be explained in section 4.3 as:

$$\widehat{\gamma}_R(h) = \frac{1}{2N(h)} \sum_{j=1}^{N(h)} \left(r(u_j) - r(u_j + h)\right)^2 \tag{12}$$

where $N(h)$ is the number of samples separated by lag h. The empirical semivariogram in equation (7) can be fitted by a theoretical model, which is usually a nugget effect model, a spherical model, an exponential model, a Gaussian model, or a combination of them. Although the shapes of these theoretical models differ, they can be described by two parameters: sill which is the variance, and range which is the distance at which the semivariogram reaches the sill; the nugget effect which represents purely random spatial variation has only a sill parameter. An excellent introduction to semivariogram modelling can be found in Isaaks and Srivastava (1989).

4.3 *Interpolating population density using area-to-point residual kriging*

Suppose a land-use zone is referred to by its centroid u. Its population density can be written as

$$d(u) = \widehat{d}(u) + r(u) \tag{13}$$

where $d(u)$ is the unknown population density of land-use zone u, $\widehat{d}(u) = m(u)$ is the value estimated by the regression model in equation (2) or (3), and $r(u)$ is the residual population density. The task is to interpolate $r(u)$ so that $d(u)$ can be estimated. Let v_u be the census unit containing u. v_u also has a residual population density $r(v_u)$ due to the difference between the census-reported population and that obtained by aggregating the regression estimates within it. $r(u)$ and $r(v_u)$ are related by:

$$\begin{aligned} r(v_u) &= \frac{P_{v_u} - \sum_j \widehat{d}(u_j) A_{u_j}}{A_{v_u}} = \frac{\sum_j d(u_j) A_{u_j} - \sum_j \widehat{d}(u_j) A_{u_j}}{A_{v_u}} \\ &= \sum_j \frac{\left\{ d(u_j) - \widehat{d}(u_j) \right\} A_{u_j}}{A_{v_u}} = \sum_j \frac{A_{u_j}}{A_{v_u}} r(u_j) \end{aligned} \tag{14}$$

where A_{u_j} is the area of land-use zone u_j; P_{v_u} and A_{v_u} are the population and area of the host census unit, respectively. The values of P_{v_u}, A_{u_j}, A_{v_u} are known, and $\widehat{d}(u_j)$ is estimated from regression; thus $r(v_u)$ can be estimated, too. $r(u_j)$ is unknown except when a census unit has one and only one land-use zone, i.e. the census unit has homogenous land use, and hence no need for disaggregation.

In this study, the regression model in equation (3) was applied to obtain the value of $\widehat{d}(u)$ in equation (13). Area-to-point residual kriging was then applied to estimated residual population density $r(u)$. Equation (14) shows that $r(v_u)$ is an area-weighted linear average of $r(u)$ values. To apply the area-to-point residual kriging method discussed in section 4.1, two approaches were investigated in this research, both of which have been used in other interpolation studies (e.g. Bracken and Martin 1989, Mennis 2003). The first was to use the population-weighted centroid to represent a land-use zone. Because the population density of a land-use zone is assumed uniform, the population-weighted centroid is equivalent to the geometric centroid. Area-to-point kriging was conducted for each centroid; the residual value obtained was interpreted as being representative of the corresponding land-use zone as a whole. The other approach was to rasterize the land-use zones into a grid and then treat each cell as a point; the interpolated values of the cells were aggregated to the land-use zone level by averaging the estimates of the cells in the same land-use zone.

The semivariogram of the residuals was computed from those land-use zones whose population density is known. These land-use zones are either the only land-use zones in their host census units, or zero-population zones within a host census unit with zero population. The residual population density of these land-use zones was calculated using equation (4). The centroid and grid approach were then applied to compute the residual semivariogram. In using the centroid approach, it is found that the empirical semivariogram was difficult to be fitted by a theoretical model, possibly due to the sensitivity of semivariogram to abnormal values. As an alternative, the correlogram was computed instead of the semivariogram (figure 2(*a*)). A correlogram is similar to a semivariogram except that it represents the correlation coefficient, instead of semivariance, between attribute values separated by lag h. The semivariogram based on the grid approach is shown in figure 2(*b*). Ideally, the resolution of the grid should be fine enough so that each cell can be treated as a point. In reality, computational cost increases as the resolution gets finer. To balance the two factors, an experiment was conducted by varying the cell size from 70 to 10 m in order to identify an optimal resolution. It was found that the semivariogram based on the grid approach was fairly robust in the sense that different resolutions resulted in similar semivariograms. The 30-m resolution was selected to perform interpolation.

The parameters of the semivariograms generated by the centroid and grid approach were fitted by the theoretical models in table 1. Although the two semivariograms were not exactly the same, the values of their parameters were quite similar: both had a nugget of 0.01, a range around 1100 meters, and a sill between 1.01 and 1.09.

The centroid- and grid-based semivariograms were applied to estimate the residual population density of each land-use zone using the area-to-point simple kriging system in equation (11). The kriging-estimated residual population density was then combined with the regression-estimated population density to obtain the overall population density of each land-use zone. Since population density cannot be less than 0, negative estimates were adjusted to 0. The proportion of the negative estimates was relatively low—18% in the centroid approach and 10% in the grid approach. All of the negative estimates were found to lie in non-residential areas

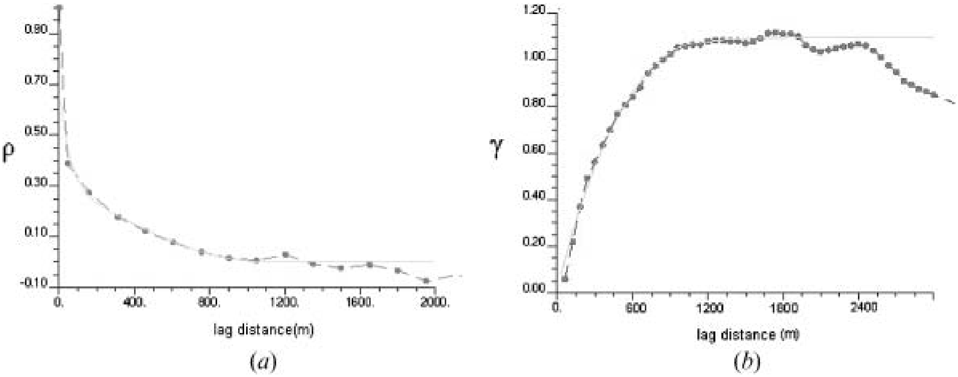

Figure 2. Modelling the spatial autocorrelation of the residual population density: (*a*) correlogram estimated from the centroid approach; (*b*) semivariogram estimated from the grid approach using 30-m resolution.

except 12 in the centroid approach and eight in the grid approach. Examination of their locations revealed that most of them were located in the boundaries between residential and non-residential land use, and their sizes were small compared with other land-use zones in the same census unit. This was no surprise as area-to-point kriging, like most interpolation methods, is known to become less accurate in areas with fewer samples nearby or if the study area boundary is reached. For a recently developed approach accounting for inequality constraints, such as non-negativity, in area-to-point kriging, the reader is referred to Yoo and Kyriakidis (2006).

The population density patterns obtained from the centroid and grid approaches were very similar; that pertaining to the grid approach is shown in figure 3. High population density occurs in the residential uses in downtown area, and low density is associated with commercial land use or less inhabited places. This pattern agrees with field knowledge about the study area. Figures 4 and 5 contrast the population density reported by census and that estimated by regression supplemented by area-to-point residual kriging. It can be seen that by using the information from the IKONOS image, a more detailed population distribution than that reported by the census is obtained.

To understand the difference between the centroid and grid estimates, their relative discrepancy was calculated:

$$\mathrm{err} = \frac{d_{\mathrm{centroid}} - d_{\mathrm{grid}}}{d_{\mathrm{grid}} + 1} \tag{15}$$

Table 1. Theoretical models for the centroid-based correlogram and grid-based semivariogram.

Centroid				Grid			
Type	Nugget	Range (m)	Sill	Type	Nugget	Range (m)	Sill
S[a]	0.01	0.1	0.30	S[a]	0.01	1080	0.54
S[a]		150	0.40				
E[a]		1100	0.31	E[a]		1080	0.55

[a] S: spherical model; E: exponential model.

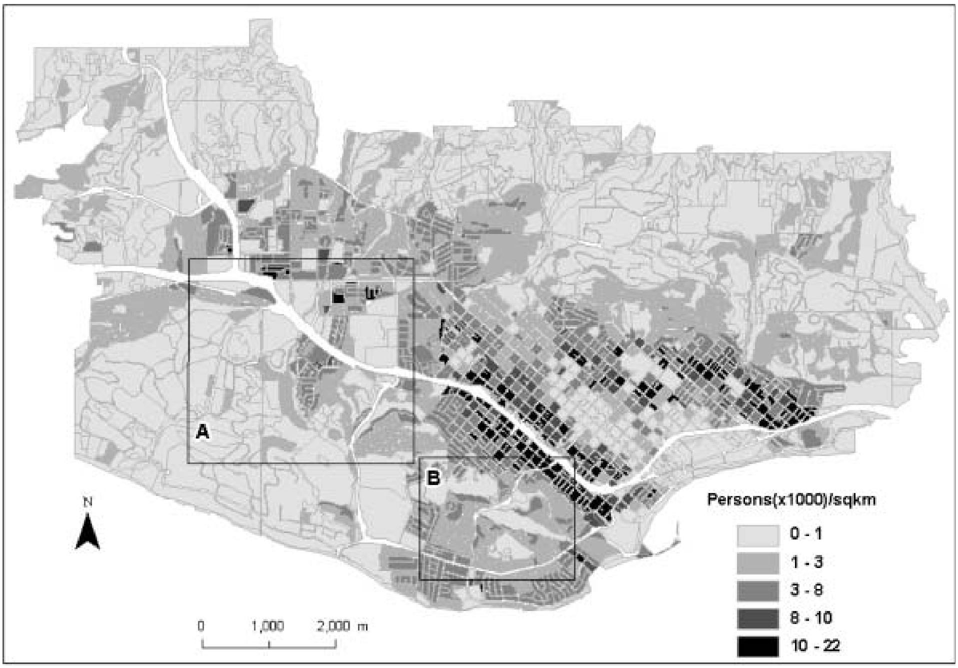

Figure 3. Population density estimated by regression supplemented by area-to-point residual kriging using the grid approach.

where err is the relative discrepancy; d_{centroid} and d_{grid} are the population densities estimated by the centroid and grid approach respectively. Equation (15) uses $d_{\text{grid}}+1$ as the denominator instead of d_{grid} to avoid division by 0. The magnitude of non-zero values of d_{grid} is usually much higher than $100\,\text{km}^{-2}$, so the difference between d_{grid} and $d_{\text{grid}}+1$ is negligible. Figure 6 shows that in most areas, the discrepancy between the centroid- and grid- based estimates was less than 10%. Severe discrepancy occurs in boundary areas or areas of very irregular shapes.

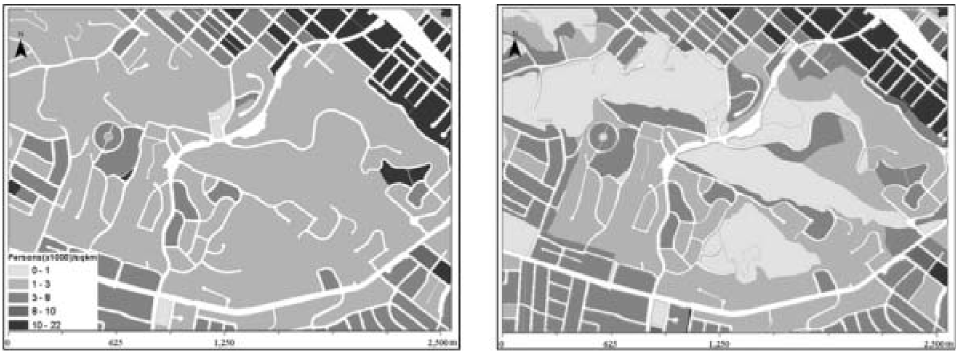

Figure 4. Details of the low-density area, specified as rectangle A in figure 3, showing the difference between the population density reported by the census (left) and that estimated by areal interpolation (right).

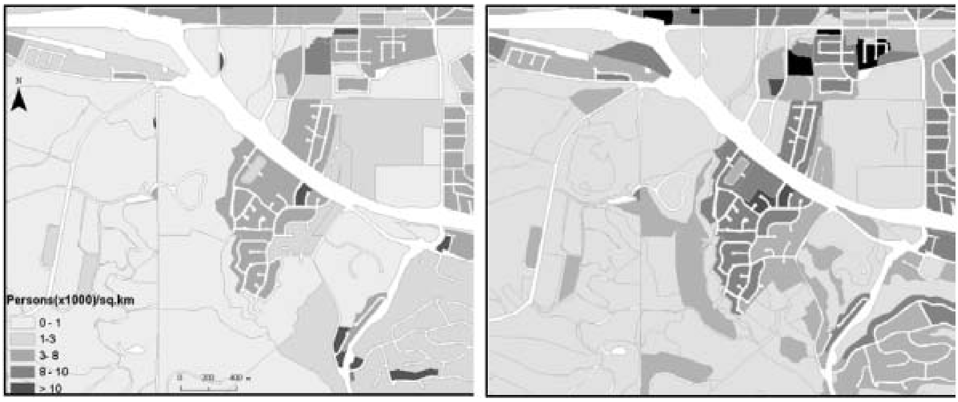

Figure 5. Details of a study area, specified in figure 3 as rectangle B, showing the difference between the population density reported by the census (left) and that estimated by areal interpolation (right).

5. Accuracy assessment

The overall accuracy of population-density estimation was evaluated by aggregating the population of the land-use zones to the census blocks. In theory, regression supplemented by area-to-point residual kriging reproduces the population of the census blocks. However, this mass-preserving property was lost in this study because the negative estimates from kriging were reset to 0. The correlation coefficient ρ between the estimated population and that reported by the census was thus

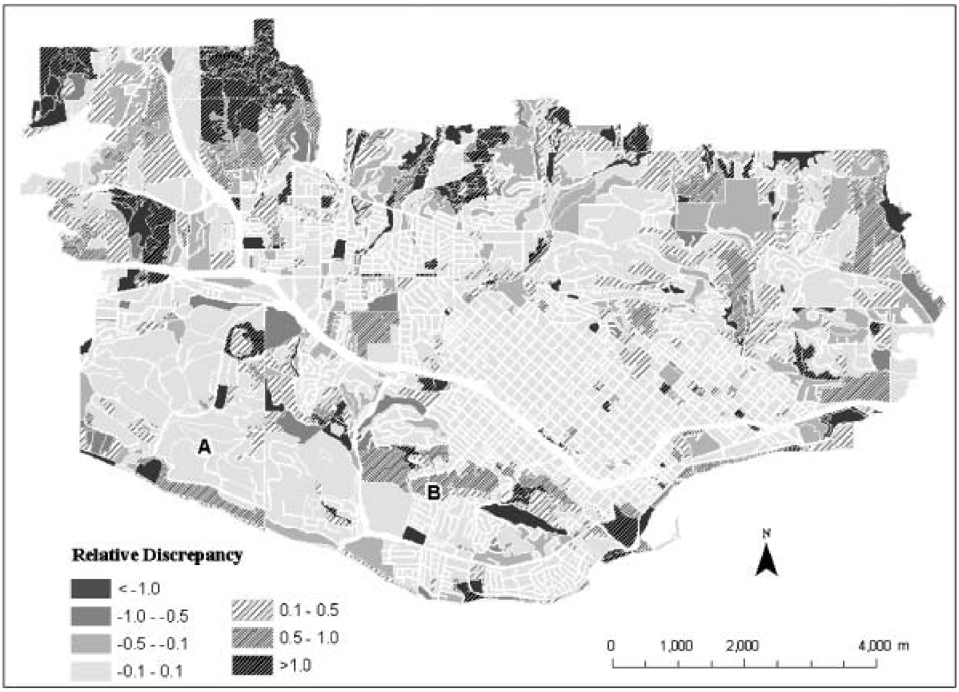

Figure 6. Relative discrepancy between population-density estimates obtained by the centroid and grid approaches.

calculated to examine how well the source data were preserved (Martin 1996). Both the centroid and grid approaches yielded a correlation coefficient close to 1.0. To further differentiate them, the root mean squared error (RMSE) and the mean absolute error (MAE), which were applied by Fisher and Langford (1996) and Wu and Murray (2005) to assess the accuracy of their interpolation methods, are calculated:

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{N}\left(\widehat{P}_i-P_i\right)^2}{N}} \quad \mathrm{MAE}=\frac{1}{P}\sum_{i=1}^{N}\left|P_i-\widehat{P}_i\right|$$

where N is the number of census blocks, P_i is the population of block i, $\hat{P}_i$ is the estimated population. P is the total population in the study area; the accuracy to reproduce it can be measured by the relative error:

$$E=\frac{\widehat{P}-P}{P}$$

where $\widehat{P}$ is the estimated total population.

Recall that area-to-point residual kriging was used to enhance regression-based population density estimation. To assess the utility of residual modelling in improving population density estimation, the summary statistics ρ, RMSE, MAE, and E were calculated for the regression models in equations (2) and (3). The results are shown in table 2. It can be seen that the two regression models yield very similar accuracies, and their values of the mean absolute error are fairly high (around 45.6–48.5%). When supplemented with area-to-point residual kriging, the accuracy is substantially improved, as indicated by the much lower values of MAE and RMSE. All of the methods seem to perform well in terms of relative error of the total population in the study area, with overestimations or underestimations within 2–3%. The accuracies of the centroid and grid approaches are similar. The grid approach may be slightly better, considering that its MAE value is lower, and the corresponding residual semivariogram modelling procedure is less difficult to implement and more robust than the centroid approach.

It has to be pointed out that the accuracy assessment method used in this paper is conducted at the source zone level, i.e. using census blocks. A better approach would be to validate the method using a secondary study area with similar characteristics, as done by Harvey (2002). Furthermore, the assessment should be conducted at the target-zone level, i.e. using land-use zones. After all, the point of areal interpolation is to obtain accurate estimates for the target zones instead of reproducing the

Table 2. Accuracy assessment of regression-based predictions with and without the supplement of area-to-point kriging.

	Regression		Regression supplemented by cokriging	
	Land use (equation (2))	Spatial metrics (equation (3))	Centroid	Grid
ρ	0.70	0.72	0.98	0.98
RMSE	104.8	107.6	21.1	22.4
MAE	48.5%	45.6%	2.9%	4.3%
E	2.1%	−7.8%	−2.6%	2.2%

source-zone information. In practice, it is not always feasible to find a second area and obtain its data at the target-zone level. An alternative is to use the Monte Carlo simulation method discussed by Fisher and Langford (1996), which creates artificial target zones by merging source zones and then conducts interpolation and validation. This method is currently under examination, and the preliminary results show that the grid approach seems to perform better.

6. Discussion and conclusion

Population density data are important for various applications. Data obtained from the census are often refined using ancillary information in order to obtain more detailed estimates. One simple method is to use regression based on remote-sensing covariates such as land-use categories or their characteristics. However, this method may not perform well in some areas, probably due to its limited ability to account for locational dependence and spatial correlation in the residuals. In this paper, an area-to-point residual kriging method is presented, which can be used to interpolate the residuals resulting from regression. Comparative results show that area-to-point residual kriging substantially improved estimation accuracy. In particular, the value of RMSE at the census-block level was lowered from 107 to 22, and the mean absolute error was lowered from 48–49% to 2.8–6.3%. These results suggest that area-to-point residual kriging is a strategy worthy considering for enhancing regression-based population-density estimation.

An interesting aspect of the residual modelling method presented in this study is area-to-point kriging. One reason explaining the improvement brought by area-to-point residual kriging is that it accounted for the location dependence and spatial correlation aspects of residual population density. Conventional cokriging requires semivariogram models for the primary and secondary variables plus the cross-semivariogram model between them to conduct interpolation. The area-to-point residual kriging method presented in this paper capitalizes on the fact that the residual population density of a source zone is a weighted linear average of the residual population density of the target zones, and simplified the semivariogram modelling procedure by requiring the semivariogram model of the point residuals. Two approaches to semivariogram computation were explored in this paper, and the grid approach is recommended. Note that several researchers (e.g. Chiles and Delfiner 1999) have pointed out that the semivariogram of regression-based residuals is a biased estimate of the true semivariogram of such residuals and recommended maximum-likelihood approaches for inference. In this paper, we opted for the simpler and less computationally intensive approach, which does not yield significantly biased results for short lag distances; we will be reporting improvements on the residual semivariogram inference in the near future.

The population-density estimation method presented in this paper is a two-step procedure: first, regression-based estimates are derived along with the corresponding residuals, and then these residuals are interpolated in space via area-to-point kriging. An alternative would be to integrate these two steps into a single area-to-point cokriging system by using population density as the primary variable and the information from the land-use/land-cover map or remotely sensed image as the secondary variable. Wu and Murray (2005) presented such a cokriging method (ignoring scale effects, however) and found it was also superior to regression-based interpolation. Interestingly, the secondary variable they used was the fraction of imperious surface, which is related to the two variables in equation (3) for built-up areas—the percentage of built-up area (P_1) and built-up patch density (P_3). Interested researchers may want to compare the two approaches to find out whether cokriging is better used to interpolate population density directly or reserved for the residual population density only. Because the two studies are not conducted in the same area (Columbus, OH vs. Santa Barbara, CA), their results are not comparable, and so no conclusion can be drawn. Only a speculation can be presented based on a loose comparison. The mean population per census block is 41 in Columbus and 91 in Santa Barbara, yet the RMSE values of the two studies are about 45 and 22, respectively. Also, in terms of the absolute mean error (MAE), their values at the census block level are 34.7% and 2.8%, respectively. Although these statistics may suggest the possibility that regression supplemented by area-to-point residual kriging is better than cokriging, it has to be stressed that unless the two methods are tested using the same set of source data in the same study area, no conclusion should be drawn. One advantage of regression supplemented by area-to-point residual kriging is that it is relatively easier to implement than cokriging because of the simplification in semivariogram modelling discussed previously.

Both the residual modelling approach and cokriging might be vulnerable to errors in the source data. Errors in image classification and census data can both affect interpolation results. In this research, the remote-sensing image was visually interpreted by an expert to minimize the error in the resulting land-use map. However, the census data were not error-free. The population density of some blocks was actually abnormally high. Image analysis and fieldwork show that these blocks are characterized by large apartment complexes and thus understandably exhibit a high population density. However, there is no way to determine whether the population density is as high as that reported by the census, since population counting is not feasible. Additionally, the mixed land use in downtown Santa Barbara also presented a challenge, although the semivariogram and correlogram in figure 2 confirmed the existence of spatial auto-correlation over short distances. The robustness of interpolation algorithms to these errors is yet to be studied.

The research in this paper used the regression model in equation (3) to obtain initial estimates of population density. This is not a necessity. Other regression models such as that in equation (2) can also be applied. In fact, area-to-point residual kriging can be used to supplement any other method as long as the residual density is found to be spatially varying and auto-correlated. In the future, other methods found in the literature may be supplemented by area-to-point residual kriging and tested to further evaluate the utility of residual modelling. Clearly, more research needs to be conducted to identify a simple yet accurate method to estimate population density in urban areas.

Acknowledgements

The authors thank Melissa Kelly at University of California, Santa Barbara for sharing the land-use map created from visual interpretation of aerial photos. This research is supported by the National Geospatial-Intelligence Agency under the award 'Strategic Enhancement of NGA's Geographic Information Science Infrastructure'.

References

Barnsley, M.J. and Barr, S.L., 1997, A graph based structural pattern recognition system to infer urban land-use from fine spatial resolution land-cover data. *Computers, Environment and Urban Systems*, **21**, pp. 209–225.

BRACKEN, I. and MARTIN, D., 1989, The generation of spatial population distributions from census centroid data. *Environment and Planning A*, **21**, pp. 537–543.

CHILES, J. and DELFINER, P., 1999, *Geostatistics: Modeling Spatial Uncertainty* (New York: Wiley).

CURRAN, P. and ATKINSON, P.M., 1998, Geostatistics and remote sensing. *Progress in Physical Geography*, **22**, pp. 61–78.

DONNAY, J.-P. and UNWIN, D., 2001, Modelling geographical distributions in urban areas. In *Remote Sensing and Urban Analysis*, J.-P. Donnay, M.J. Barnsley and P.A. Longley (Eds), pp. 205–224 (London: Taylor & Francis).

FISHER, P. and LANGFORD, M., 1996, Modeling sensitivity to accuracy in classified imagery: A case study of areal interpolation by dasymetric mapping. *The Professional Geographer*, **48**, pp. 299–309.

GOODCHILD, M., ANSELIN, L. and DEICHMANN, U., 1993, A framework for the areal interpolation of socioeconomic data. *Environment and Planning A*, **25**, pp. 383–397.

GOOVAERTS, P., 1997, *Geostatistics for Natural Resources Evaluation* (Oxford: Oxford University Press).

HARVEY, J.T., 2002, Population estimation models based on individual TM pixels. *Photogrammetric Engineering and Remote Sensing*, **68**, pp. 1181–1192.

HEROLD, M., LIU, X. and CLARKE, K., 2003, Spatial metrics and local texture for mapping urban land use. *Photogrammetric Engineering and Remote Sensing*, **69**, pp. 991–1002.

HEROLD, M., MULLER, A., GUENTER, S. and SCEPAN, J., 2002, Object-oriented mapping and analysis of urban land use/cover using IKONOS data. In *The 22nd EARSEL Symposium, 4–6 June 2002*, Prague, Czech Republic, pp. 531–538.

ISAAKS, E. and SRIVASTAVA, M., 1989, *An Introduction to Applied Geostatistics* (London: Oxford University Press).

JENSEN, J.R. and COWEN, D.C., 1999, Remote sensing of urban/suburban infrastructure and socio-economic attributes. *Photogrammetric Engineering and Remote Sensing*, **65**, pp. 611–622.

JOURNEL, A.G., 1999, Conditional geostatistical operations to nonlinear volume averages. *Mathematical Geology*, **31**, pp. 931–953.

KRAUS, S. and SENGER, L., 1974, Estimating population from photographically determined residential land use types. *Remote Sensing of Environment*, **3**, pp. 35–42.

KYRIAKIDIS, P., 2004, A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, **36**, pp. 259–289.

LAM, N.-N., 1983, Spatial interpolation methods: A review. *The American Cartograher*, **10**, pp. 129–149.

LANGFORD, M., MAGUIRE, D. and UNWIN, D., 1991, The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In I. Masser and M. Blakemore (Eds). *Handling Geographical Information: Methodology and Potential Applications*, pp. 55–77 (London: Longman).

LANGFORD, M. and UNWIN, D.J., 1994, Generating and mapping population density surfaces with a geographical information system. *The Cartographic Journal*, **31**, pp. 21–26.

LIU, X., CLARKE, K. and HEROLD, M., 2006, Population density and image texture: A comparison study. *Photogrammetric Engineering and Remote Sensing*, **72**, pp. 187–196.

LIU, X. and HEROLD, M., 2007, Estimating population distributions in urban areas. In *Urban Remote Sensing*, Q. Weng and D. Quattrochi (Eds), pp. 269–290 (London: CRC Press/Taylor & Francis).

MARTIN, D., 1996, An assessment of surface and zonal models of population. *International Journal of Geographical Information Systems*, **10**, pp. 973–989.

MCGARIGAL, K., CUSHMAN, S.A., NEEL, M.C. and ENE, E., 2002, FRAGSTATS: Spatial pattern analysis program for categorical maps. Available online at: http://www.umass.edu/landeco/research/fragstats/fragstats.html (accessed 8 April 2007).

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

MONMONIER, M.S. and SCHNELL, G.A., 1984, Land-use and land-cover data and the mapping of population density. *The International Yearbook of Cartography*, **24**, pp. 115–121.

OKABE, A. and SADAHIRO, Y., 1997, Variation in count data transferred from a set of irregular zones to a set of regular zones through the point-in-polygon method. *International Journal of Geographical Information Science*, **11**, pp. 93–106.

OPENSHAW, S., 1984, The modifiable areal unit problem. *Concepts and Techniques in Modern Geography*, 38 (Norwich, UK: Geobooks).

PEPLIES, R.W., 1974, Regional analysis and remote sensing: a methodological approach. *Remote Sensing: Techniques for Environmental Analysis*, J. Estes (Ed.), pp. 277–291 (Santa Barbara, CA: Hamilton).

RASE, W.-D., 2001, Volume-preserving interpolation of a smooth surface from polygon-related data. *Journal of Geographical Systems*, **3**, pp. 199–213.

TOBLER, W., 1979, Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, **74**, pp. 519–536.

WRIGHT, J., 1936, A method of mapping densities of population with Cape Cod as an example. *Geographical Review*, **26**, pp. 103–110.

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

YOO, E.-H. and KYRIAKIDIS, P.C., 2006, Area-to-point Kriging with inequality-type data. *Journal of Geographical Systems*, **8**, pp. 357–390.