Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography
Vol. 33, No. 3, 145-153, 2015
http://dx.doi.org/10.7848/ksgpc.2015.33.3.145

ISSN 1598-4850(Print)
ISSN 2288-260X(Online)
Original article

A Comparative Study on the Spatial Statistical Models for the Estimation of Population Distribution

Oh, Doo-Ri[1] · Hwang, Chul Sue[2]

Abstract

This study aims to accurately estimate population distribution more specifically than administrative unites using a RK (Regression-Kriging) model. The RK model is the areal interpolation technique that involves linear regression and the Kriging model. In order to estimate a population's distribution using a sample region, four different models were used, namely; a regression model, RK model, OK (Ordinary Kriging) model and CK (Co-Kriging) model. The results were then compared with each other. Evaluation of the accuracy and validity of evaluation analysis results were the basis RMSE (Root Mean Square Error), MAE (Mean Absolute Error), G statistic and correlation coefficient (ρ). In the sample regions, every statistic value of the RK model showed better results than other models. The results of this comparative study will be useful to estimate a population distribution of the metropolitan areas with high population density

Keywords : Areal Interpolation, Dasymetric Mapping, Kriging, Regression-Kriging, GIS, Population Density

1. Introduction

Many complicated researches need the spatial data with a finer resolution than the enumeration unit such as administrative boundaries. Aggregated statistical data such as census data can provide a summarized description, but it can also imply some interpretative misinformation. Openshaw (1984) claims that statistical sources can be biased because results are the value of summary in arbitrary areal units, which are called Modifiable Areal Unit Problem (MAUP). Another problem of census data is that it assumes a homogeneous population in area (Monmonier and Schnell, 1984; Langford and Unwin, 1994; Eicher and Brewer, 2001). Dasymetric mapping method can be one of the solutions to interpolate the census data because it allows realistic presentation of population (Eicher and Brewer, 2001; Mennis, 2003; Mennis and Hultgren, 2006).

Dasymetric mapping method, the theoretical foundation for the study, has been studied extensively. Many researchers generally consider dasymetric mapping method as one of the areal interpolation methods (Lee and Kim, 2007). In mathematical perspectives, it has been mainly addressed as a weight method for population interpolation for the target area. On the other hand, the use of higher resolution ancillary data is of interest in data aspects. Several ancillary data were used for dasymetric interpolation mapping such as residential area that were extracted from high resolution satellite image (Ku, 2008; Wu and Murray, 2005), cadastral map data (Maantay *et al*., 2007; Tapp, 2010), streets and roads data (Reibel and Bufalino, 2005), building data in downtown (Lwin and Murayama, 2009), and the point data of individual residence (Zandbergen, 2011). However,

Received 2015. 04. 09, Revised 2015. 05. 22, Accepted 2015. 06. 24

[1] Member, Dept. of Geography, Kyung Hee University (E-mail: droh@khu.ac.kr)
[2] Corresponding Author, Member, Dept. of Geography, Kyung Hee University (E-mail: hcs@khu.ac.kr)

these high resolution data is limited in a certain area. It is impossible to conduct a study if there is no data in the study area. Weighting method such as areal weighting (Mennis and Hultgren, 2006), population proportion, regression analysis method (Flowerdew and Green, 1992; Reibel and Agrawal, 2007) in dasymetric mapping has also limitations since it may not reflect the characteristics of spatial data or make higher estimation errors of the results. In order to reduce the error estimation value in this study, different statistical models were used to calculate a weight to estimate the population's distribution. This study has its own originality in that both RK model and Kriging models are used and compared for population estimation in study area. Also, because previous works (Wu and Murray, 2005) used CK model for population estimation in urban area, we used the same model and we used OK and regression models to compare the differences among the other models.

This study aims to estimate the population distribution with more precise spatial resolution using geo-statistical methods. In order to do this, we applied four geo-statistical models and compared the results with each other. Four models are (1) regression model, (2) Regression-Kriging (RK) model, (3) Ordinary Kriging (OK) model and (4) Co-Kriging (CK) model. We investigate the applicability of RK model to estimate the population of the metropolitan area with the high population density because RK model is suitable to predict the subtle changes. In this respect, we considered the properties and the relations between the size of residential area and the population size for the study area. We chose the study area where it is suitable to use RK model to estimate population through testing samples several times. Both of the study area, Dongdaemun-gu and Jungnang-gu (Case 1), and Mapo-gu and Seodaemun-gu (Case 2), have relatively high population density in Seoul, Korea.

We used the land use data of 2000s from the NGII (National Geographic Information Institute) as the ancillary data or the target data in the dasymetric mapping. Two different types of residential area from the land use data were used as independent parameters in the regression analysis. We assumed that the population distribute only in the residential area using categories in land use data. Residential zones include binary area which are high population density area (buildings above 5th floor and subsidiary facilities) and low population density area (buildings below 5th floor and subsidiary facilities). And the smallest administrative areal unit (dong) from the census data of 2000 was used for the source data. Statistical analyses for the study were performed in the R 2.15.2 and then the ArcGIS 10 was used for the spatial analysis and mapping.

2. Approach

2.1 Dasymatric interpolation mapping with RK model

The RK model involves the linear regression model that analysis the whole drift/trend between the variables and the Kriging model that interpolates the residuals by regression (Hengl *et al*., 2004; Hengl *et al*., 2007). For example, the RK model can be written as Eq. (1).

$$\widehat{z_{RK}}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) \tag{1}$$

Suppose that $\widehat{z_{RK}}(s_0)$ is the predicted population density at the location, s_0, $\hat{m}(s_0)$ is the predicted value from the regression model, and $\hat{e}(s_0)$ is the predicted value using the Kriging model. This equation can also be expressed as Eq. (2).

$$\widehat{z_{RK}}(s_0) = \textstyle\sum_{k=0}^{p} \widehat{\beta_k} \cdot q_k(s_0) + \sum_{i=1}^{n} \omega_i\,(s_0) \cdot e(s_i), \tag{2}$$

where $\widehat{\beta_k}$ is the estimated regression coefficient, q_k is the independent variable, ω_i is the Kriging weight and $e(s_i)$ is the residual by the regression model at the location, s_i.

The RK model goes through the following process: Step 1. Correlation coefficients were used in order to determine the relationship of the combination of land uses and population density so the combination with the highest correlation coefficients is selected. Step 2. Ordinary least squares (OLS) multiple regression was used to generate the surface of the population density. In performing a regression analysis, if a spatial autocorrelation is shown in the residuals, a generalized least squares (GLS) regression analysis can supplement OLS multiple regression. Step

3. The values from the regression analysis and the kriged residuals are added, the result will be the estimated population distribution

2.2 Dasymetric interpolation mapping with Kriging model

In order to perform the OK and CK model, the covariance matrix and the variogram are calculated with a primary variable (census data) for the OK model, and a primary variable and a secondary variable (the information of residential area) for the CK model. The primary variable and the secondary variable used in the CK model are interrelated spatially. Among the theoretical semi-variogram, the spherical model shows the best fit using the factors of the semi-variogram in every case in this study. Population weight used on the OK and CK model can be extracted from the semi-variogram model. Refer to Choe (2007) and Knotters *et al*. (1995) for more model explanation. In this study, we performed Kriging for population distribution with continuous variable and re-aggregated values of population in each residential zones. In other words, the arithmetic mean of all the points in the grid cell of the estimated population density was calculated for the predictions.

3. Case Studies

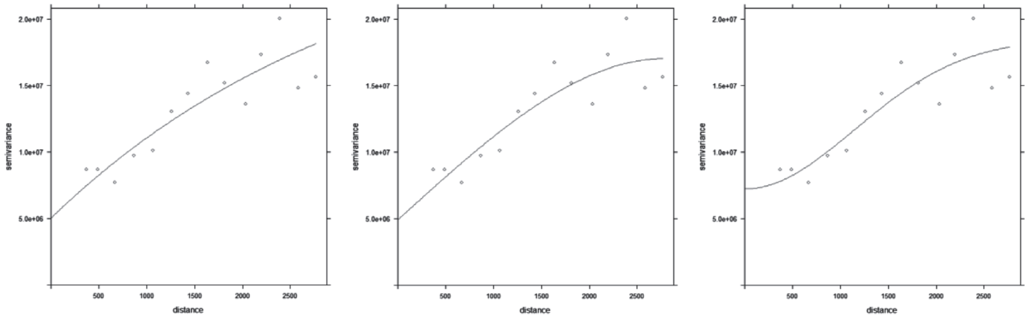
3.1 A case study with the RK model

We used the census data and binary residential area of the land use data to facilitate the dasymetric mapping with the RK model. After a correlation analysis between the census data and the binary residential area to determine the regression parameters, regression analysis was performed to generate the trend surface of the population density. GLS regression analysis was used instead of OLS regression based on the Durbin-Watson test. Minimize AIC (Akaike Information Criterion) can be used as criteria for model selection and we applied the AIC to the spherical, exponential, and Gaussian model. Refer to Akaike (1977) and Eldeiry and Garcia (2010) for more AIC explanation. Fig. 1 and Fig. 2 show the theoretical semi-variogram of the residuals and it was fit with an exponential model in Case 1 and a spherical model in Case 2. Table 1 provides the estimated coefficients by the type of regression models and the values of AIC for model selection in Case 1. The GLS residuals were interpolated using Kriging model to make the minimum and unbiased variance in locality. Estimated population density is the adding value of the results of regression model and the Kriging model.

Table 1. Regression coefficients and AIC of the Case 1

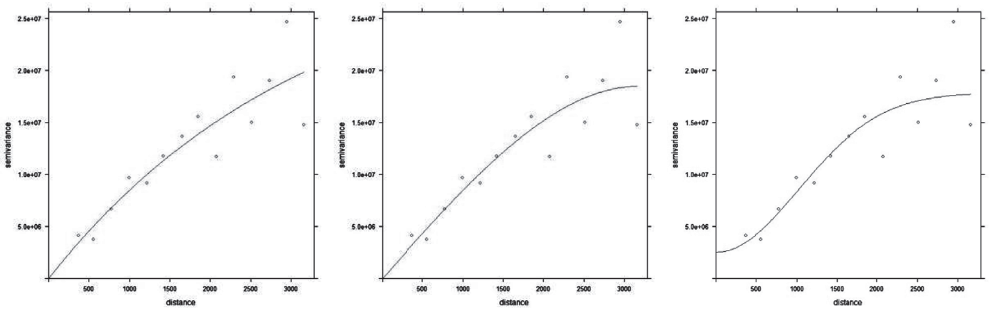
Regression Model		Regression Coefficients		AIC
		Low-density Residential Area	High-density Residential Area	
OLS		47.06*** (0.000)	54.11*** (0.000)	894.23
GLS	Spherical Model	47.06*** (0.000)	54.11*** (0.000)	898.23
	Exponential model	38.41*** (0.000)	46.47*** (0.000)	882.63
	Gaussian Model	35.62*** (0.000)	42.59*** (0.000)	883.31

Significance level: ***: 0, **: 0.05, *:0.1



(a) (b) (c)

Fig. 1. The theoretical semi-variogram, (a) exponential model, (b) spherical model, and (c) Gaussian model of the regression residuals of the Case 1



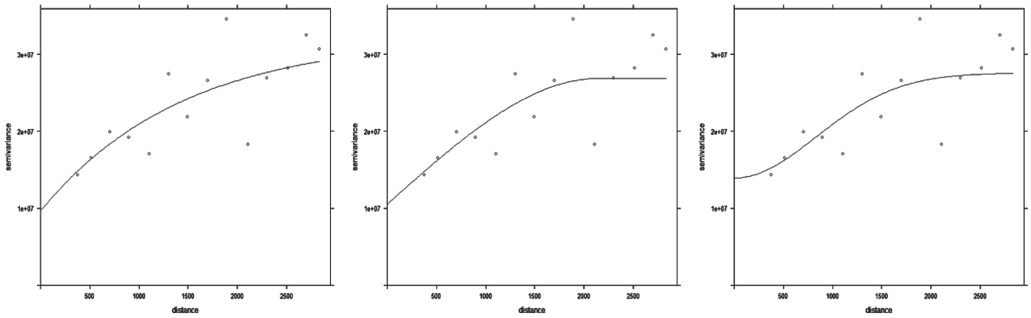
(a) (b) (c)

Fig. 2. The theoretical semi-variogram, (a) exponential model, (b) spherical model, and (c) Gaussian model of the regression residuals of the Case 2

3.2 A case study with the Kriging model

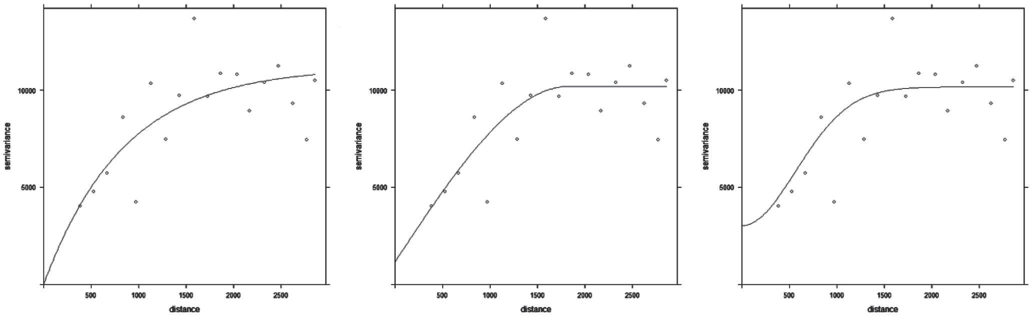
We applied the OK model and the CK model to estimate the population's distribution using Kriging in the study area. The OK model needs data on population as primary variable, whereas the CK model requires additional secondary variable that has spatial correlation between variables. Grid was created using midpoints of the smallest administrative areal unit to calculate the Kriging weights.

In order to model the population density using Kriging, covariance model has to be employed with primary variable, secondary variable, and cross variable. Utilizing the covariance model, we were able to perform the semi-variogram experiments several times and came up with factors of the semi-variogram. Semi-variogram provides information



(a) (b) (c)

Fig. 3. The theoretical semi-variogram, (a) exponential model, (b) spherical model, and (c) Gaussian model of the census data as primary variable of the Case 1



(a) (b) (c)

Fig. 4. The theoretical semi-variogram, (a) exponential model, (b) spherical model, and (c) Gaussian model of the residential area as secondary variable of the Case 1

on the nugget, sill, and range to perform the Kriging. Using these factors, we selected the best model of theoretical semi-variogram. Fig. 3, Fig. 4, and Fig. 5 show the theoretical semi-variogram of the census data, residential area data, and cross-variogram between census and residential area, respectively for the Case 1. We also performed in the same procedure for the Case 2. Population weights used on an OK model and a CK model can be extracted from semi-variogram.

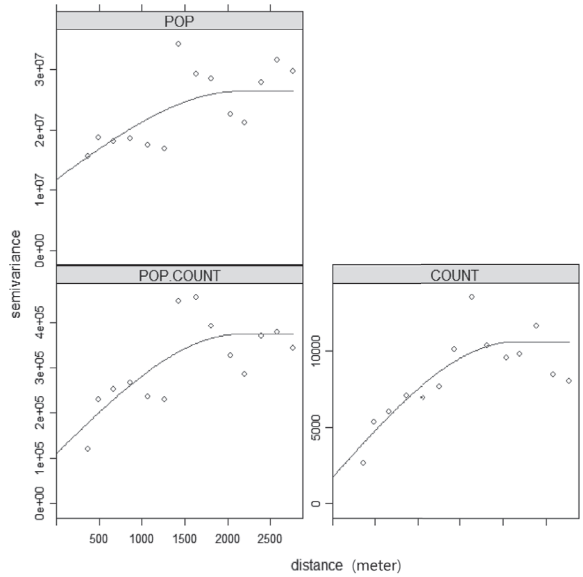
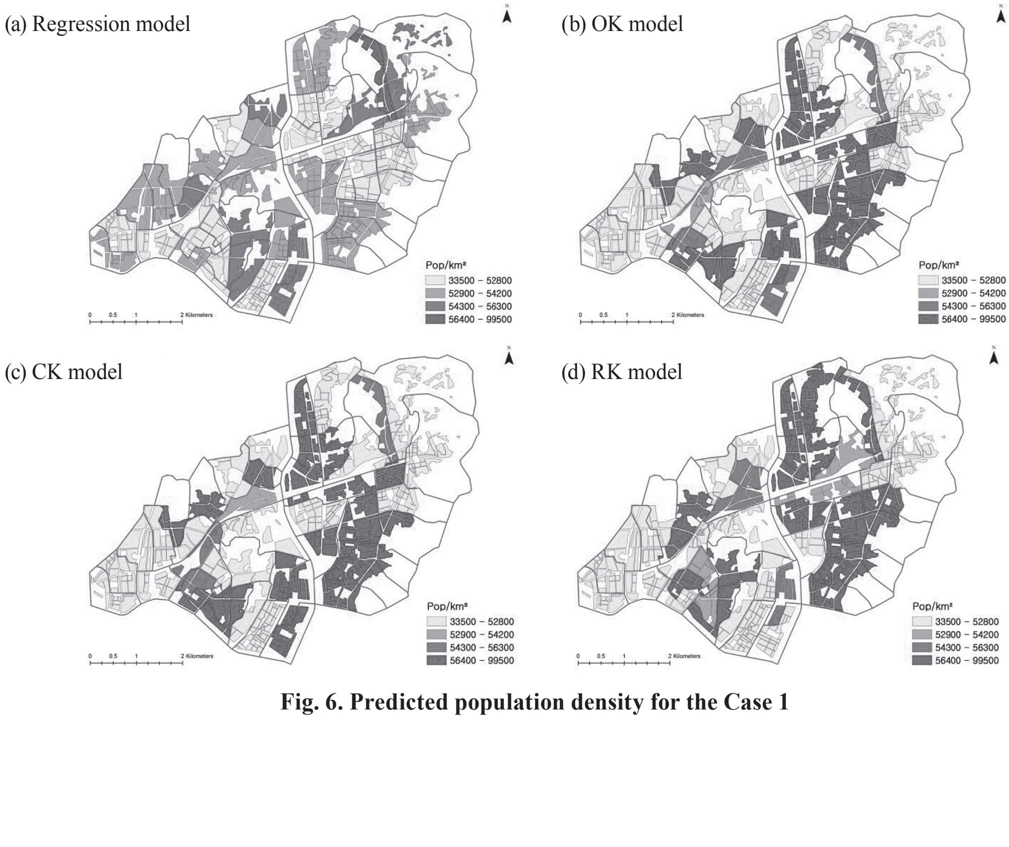


Fig. 5. The theoretical semi-variogram of cross-variogram between the primary and the secondary variable of the Case 1

4. Results

4.1 Results of the dasymetric interpolation mapping

The results of population distribution using the different models were counted by the census units for comparison each other. All models have a similar average value of the estimated population compared to the ordinary/original data. However, the maximum value of the population decreased while its minimum value increased in all models. The degree of change is bigger in the Kriging model than in the RK model because the Kriging model uses distance functions when it estimates new variables (Fig. 7). Fig. 6 reveals the predicted population distribution in the study area. The range of y-values are different because the two figures have different calculation method: Fig. 6 is about population density per residential units and Fig. 7 is about population for comparison with census data.


(a) Regression model

(b) OK model

(c) CK model

(d) RK model

Fig. 6. Predicted population density for the Case 1

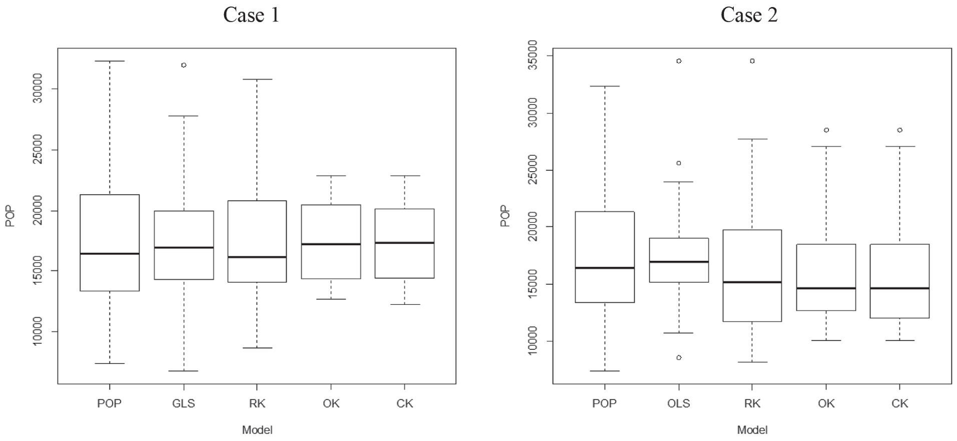


Fig. 7. Box-Whisker plots of the predicted population density

4.2 Model evaluation and validation

Evaluation of the accuracy and validation were the basis on the root mean square error (RMSE), mean absolute error (MAE), goodness of prediction statistic (G statistic) (Kravchenko and Bullock, 1999; Guisan and Zimmermann, 2000; Eldeiry and Garcia, 2010; Kim *et al*., 2010) and correlation coefficient (ρ). The RMSE can be defined as Eq. (3) and the MAE can be calculated as Eq. (4).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{p_i} - p_i)^2}{N}} \tag{3}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N}|p_i - \widehat{p_i}|}{N}, \tag{4}$$

where $\widehat{p_i}$ is the predicted value of population at location i, p_i is the ordinary/original value of the population at location i, where i=1, 2, 3, ⋯ ⋯ . , N.

The effectiveness of the models was measured by the G statistic that can be written as Eq. (5).

$$G = \left[1 - \left\{\sum_{i=1}^{N}\frac{(p_i - \widehat{p_i})^2}{\sum_{i=1}^{N}(p_i - \bar{p})^2}\right\}\right], \tag{5}$$

where p_i is the ordinary/original value of the population at location i, $\widehat{p_i}$ is the predicted value of population at location i, and $\bar{p}$ is the mean of the population in the sample area. The model is more efficient when the G statistic has a positive value close to 1. The model is not very efficient when the G statistic has a negative value.

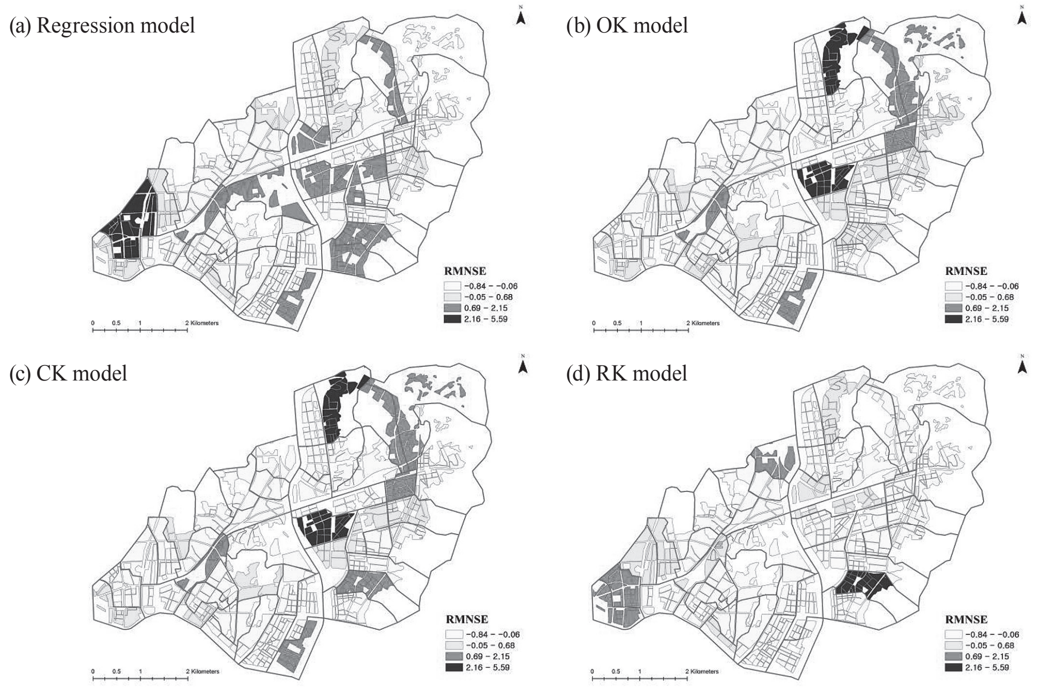
Correlation coefficient (ρ) measures the pattern between the ordinary/original value and the predicted value. Every statistical value of the RK model, the Kriging model, and the regression model, following this order, shows better results. As for the OK model and CK model, both show similar results in this study. Tables 2 and 3 show the evaluation and validation values on the models.

Table 2. Evaluation and validation values using the regression, RK, OK, and CK models (Case 1)

	regression	RK	OK	CK
RMSE	4159.51	1465.19	3851.78	3866.07
MSE	3370.05	1001.21	3098.42	3096.70
G statistic	0.54	0.94	0.60	0.60
ρ	0.74	0.97	0.84	0.86

Table 3. Evaluation and validation values using the regression, RK, OK, and CK models (Case 2)

	regression	RK	OK	CK
RMSE	4489.87	2837.32	2948.81	2946.32
MSE	3638.93	1407.37	2199.93	2190.55
G statistic	0.53	0.81	0.80	0.80
ρ	0.76	0.92	0.84	0.86



(a) Regression model

(b) OK model

(c) CK model

(d) RK model

Fig. 8. NRMSE values for the Case 1

4.3 Zonal errors in population estimations

The normalized root mean square error (NRMSE) is utilized to investigate the prediction error of a specific unit area. As a result, the RK model's estimation error is large in a low population density zone and small in a high population density zone. Among them, the area with the highest estimation error is revealed to be in the bordering sample areas, even in low population density areas.

Results of estimating the population distribution using the OK and CK model produce high error in boundaries or urban areas with small residential districts in the sample region. Consequently, a very distinct area gap compared to surrounding areas created big errors.

5. Discussion and Conclusions

In the case of Dongdaemun-gu and Jungnang-gu (Case 1), every statistic value of the RK model, Kriging model and regression model showed better results than other models. The OK model and CK model always showed similar results. For Mapo-gu and Seodaemun-gu (Case 2), the statistical results of the RK model, CK model, OK model, and the regression model shows better results in the order named. However, the difference in statistical results between the Kriging models and the RK model was not statistically significant.

It is hard to estimate the population distribution using previous weighting methods for dasymetric interpolation mapping, such as the areal weighting method (Goodchild *et al*., 1993) or the population proportion method (Eicher and Brewer, 2001), of the area with land use pattern of the high complexity. The RK model has higher accuracy using the study area compared to the regression, OK, and CK models because the RK model has both advantages of the regression model and the Kriging model. And estimated population from the RK method has similar values of descriptive statistics as the ordinary/original data. However, the forms of the model in conjunction with different spatial statistical models including RK model involve complicated calculations.

This study provides that RK model can be an alternative method of estimating a population distribution although the

Kriging model is frequently used for interpolation. The RK model is suitable for areas with a high population density and a positively high correlation between target data and source data. Therefore, the RK model will be useful for metropolitan areas with a high population density.

References

Akaike, H. (1977), On entropy maximization principle, In: Krishnaiah, P. R. (ed.), *Proceedings of the Symposium on Applications of Statistics*, North-Holland, Amsterdam, pp. 27-41.

Choe, J. (2007), *Geostatistics*, Sigma-press, Seoul. (in Korean)

Eicher, C.L. and Brewer, C.A. (2001), Dasymetric mapping and areal interpolation implementation and evaluation, *Cartography and Geographic Information Science*, Vol. 28, No. 2, pp. 125-138.

Eldeiry, A.A. and Garcia, L.A. (2010), Comparison of ordinary kriging, regression kriging, and cokriging techniques to estimate soil salinity using Landsat images, *Journal of Irrigation and Drainage Engineering*, Vol. 136, No. 6, pp. 355-364.

Flowerdew, R. and Green, M. (1992), Developments in areal interpolation methods and GIS, *The Annals of Regional Science*, Vol. 26, No. 1, pp. 67-78.

Goodchild M.F., Anselin, L., and Deichmann, U. (1993), A framework for the areal interpolation of socioeconomic data, *Environment and Planning A*, Vol. 25, No. 3, pp. 383-397.

Guisan, A. and Zimmermann, N.E. (2000), Predictive habitat distribution models in ecology, *Ecological Modelling*, Vol. 135, No. 2, pp. 147-186.

Hengl, T., Heuvelink, G.B.M., and Stein, A. (2004), A generic framework for spatial prediction of soil variables based on regression-kriging, *Geoderma*, Vol. 120, No. 1, pp. 75-93.

Hengl, T., Heuvelink, G.B.M., and Rossiter, D.G. (2007), About regression-kriging from equations to case studies, *Computers & Geosciences*, Vol. 33, No. 10, pp. 1301-1315.

Kim, B., Ku, C., and Choi, J. (2010), Population distribution estimation using regression-kriging model, *Journal of the Korean Geographical Society*, Vol. 45, No. 6, pp. 806-819. (in Korean with English abstract)

Knotters, M., Brus, D.J., and Oude Voshaar, J.H. (1995), A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations, *Geoderma*, Vol. 67, No. 3, pp. 227-246.

Kravchenko, A. and Bullock, D.G. (1999), A comparative study of interpolation methods for mapping soil properties, *Agronomy Journal*, Vol. 91, No. 3, pp. 393-400.

Ku, C. (2008), A study on estimating the population in urban area with high resolution satellite image, *The Geographic Journal of Korea*, Vol. 42, No. 1, pp. 137-148. (in Korean with English abstract)

Langford, M. and Unwin, D.J. (1994), Generating and mapping population density surfaces within a geographical information system, *The Cartographic Journal*, Vol. 31, No. 1, pp. 21-26.

Lee, S. and Kim, K. (2007), Representing the population density distribution of Seoul using dasymetric mapping techniques in a GIS environment, *Journal of the Korean Cartographic Association*, Vol. 7, No. 2, pp. 53-67. (in Korean with English abstract)

Lwin, K. and Murayama, Y. (2009), A GIS approach to estimation of building population for micro-spatial analysis, *Transactions in GIS*, Vol. 13, No. 4, pp. 401-414.

Maantay, J.A., Maroko, A.R., and Herrmann, C. (2007), Mapping population distribution in the urban environment the cadastral-based expert dasymetric system (CEDS), *Cartography and Geographic Information Science*, Vol. 34, No. 2, pp. 77-102.

Mennis, J. (2003), Generating surface models of population using dasymetric mapping, *The Professional Geographer*, Vol. 55, No. 1, pp. 31-42.

Mennis, J. and Hultgren, T. (2006), Intelligent dasymetric mapping and its application to areal interpolation, *Cartography and Geographic Information Science*, Vol. 33, No. 3, pp. 179-194.

Monmonier, M. and Schnell, G. (1984), Land use and land cover data and the mapping of population density, *International Yearbook of Cartography*, Vol. 24, pp. 115-121.

Oh, D. (2013), *A Comparative Study on the Spatial Statistical Models for the Estimation of Population Distribution*, Master's thesis, Kyung Hee University, Seoul, Korea, 74p. (in Korean with English abstract)

Openshaw, S. (1984), *The Modifiable Areal Unit Problem*, Geo Books, Norwick Norfolk.

Reibel, M. and Agrawal, A. (2007), Areal interpolation of population counts using pre-classified land cover data, *Population Research and Policy Review*, Vol. 26, No. 5-6, pp. 619-633.

Reibel, M. and Bufalino, M.E. (2005), Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems, *Environment and Planning A*, Vol. 37, No. 1, pp. 127-139.

Tapp, A.F. (2010), Areal interpolation and dasymetric mapping methods using local ancillary data sources, *Cartography and Geographic Information Science*, Vol. 37, No. 3, pp. 215-228.

Wu, C. and Murray, A.T., (2005), A cokriging method for estimating population density in urban areas, *Computers, Environment and Urban Systems*, Vol. 29, No. 5, pp. 558-579.

Zandbergen, P.A. (2011), Dasymetric mapping using high resolution address point datasets, *Transactions in GIS*, Vol. 15, No. s1, pp. 5-27.