대한지리학회지 제45권 제6호 2010(806~819)

Regression-Kriging 모형을 이용한 인구분포 추정에 관한 연구

김병선* · 구자용** · 최진무***

Population Distribution Estimation Using Regression-Kriging Model

Byeong-Sun Kim* · Cha Yong Ku** · Jinmu Choi***

요약 : 센서스 단위의 인구자료는 기초적인 인문사회 자료로 행정구역 단위로 요약되어 공간분석에 사용된다. 정밀한 인구 분포를 추정하기 위해 기존의 연구에서는 위성영상과 회귀분석 모형을 이용하였다. 하지만 회귀식에 의한 추정치는 공간자료의 공간적자기상관과 잔차 때문에 정확도에 있어 한계가 있었다. 본 연구는 회귀모형과 회귀모형에서 추출된 잔차에 대해 공간적자기상관을 고려하도록 크리깅 보간하는 RK모형(Regression Kriging Model)을 이용하여 인구분포의 추정 정확도를 향상하였다. RK모형을 적용하여 서울시의 4개구를 대상으로 사례분석을 하였으며, 모형의 효율성을 검증하기 위해 회귀분석만을 이용한 예측 결과와 RK모형을 이용한 예측 결과를 서로 비교하였다. 비교한 결과로 상관관계 계수, 평균제곱근 오차, G 통계량 수치에서 RK모형의 추정 정확도가 기존의 회귀모형에 비해 높게 나온 것을 확인할 수 있었다. 향후 정확한 인구추정을 위해 RK모형이 많이 활용될 수 있을 것이다.
주요어 : RK모형, 크리깅, 회귀모형, 보간, 인구추정기법

Abstract : Population data has been essential and fundamental in spatial analysis and commonly aggregated into political boundaries. A conventional method for population distribution estimation was a regression model with land use data, but the estimation process has limitation because of spatial autocorrelation of the population data. This study aimed to improve the accuracy of population distribution estimation by adopting a Regression-Kriging method, namely RK Model, which combines a regression model with Kriging for the residuals. RK Model was applied to a part of Seoul metropolitan area to estimate population distribution based on the residential zones. Comparative results of regression model and RK model using RMSE, MAE, and G statistics revealed that RK model could substantially improve the accuracy of population distribution. It is expected that RK model could be adopted actively for further population distribution estimation.
Key Words : regression kriging model, kriging, regression model, interpolation, population estimation technique

1. 서론

인구가 어디에 분포하고 이들이 어떤 지역에 집중적으로 분포하고 있는가에 대한 자료는 사회, 경제, 환경 등 다양한 분야에서 요구하는 필수자료이다. 그러나 통계청을 비롯한 각종 정부기관에서 제공하는 인구자료는 보통 행정동이나 법정동과 같은 센서스 단위로 집계된 자료를 제공하고 있다. 이는 개인정보를 보호한다는 차원에서 불가피한 선택이라 할 수 있으나, 이

* 상명대학교 지리학과 박사과정(Ph.D. Candidate, Department of Geography, Sangmyung University), gisguy@paran.com
** 상명대학교 지리학과 부교수(Associate Professor, Department of Geography, Sangmyung University), koostar@smu.ac.kr
*** 상명대학교 지리학과 조교수(Assistant Professor, Department of Geography, Sangmyung University), jmchoi@smu.ac.kr

러한 형태의 인구정보는 공간에서 발생하는 다양한 현상을 연구하는데 다음과 같은 세 가지 문제점이 있다. 첫째, 응용 분야에 따라 다양한 공간경계로의 변환이 필요한데, 이는 MAUP(modifiable areal unit problem)[1]로부터 자유로울 수 없다. 둘째, 센서스 구역 내의 토지이용이 서로 상이할 경우 일반적으로 인구는 토지이용에 따라 불균등하게 분포하지만, 연구자는 집계 자료의 특성상 불가피하게 센서스구역 내의 인구가 균등하게 분포한다는 가정 하에 분석을 수행해야만 한다. 셋째, 센서스 단위로 집계한 인구자료는 교통존, 유역도 등 다양한 형태의 인문 · 자연 주제도와 그 기본 단위가 거의 일치하지 않는다. 따라서 이와 같은 자료들과 함께 분석을 수행하기 위해서는 기본 공간 단위를 일치시켜야 하는 어려움이 존재한다(Goodchild *et al.*, 1993).

이러한 문제점을 해결하기 위해 원구역(source zone)의 자료를 대상구역(target zone)으로 보간할 수 있는 다양한 보간기법을 사용하여 센서스 단위의 인구자료를 토지이용단위 또는 그리드 단위로 인구를 추정하는 연구가 활발히 진행되어 왔다. Okabe and Sadahiro(1997)는 이러한 보간법을 단순 보간법과 지능형 보간법으로 분류하였는데, 단순 보간법은 추정하고자 하는 자료의 특성을 설명할 수 있는 보조자료가 없이 비정형적인 공간 자료를 그리드와 같은 정형화된 공간으로 보간하는 방법을 말하며, 지능형 보간법은 단순 보간법과 유사하나 보간의 정확성을 높이기 위해 보조자료를 사용한다는 점에서 차이가 있다. 비록 지능형 보간법이 단순 보간법에 비해 더 많은 연산 시간이 소요된다는 단점이 있으나, 분석 정확도가 상대적으로 높게 나오기 때문에 지능형 보간법을 활용한 다양한 연구가 진행되어 왔다(Fisher and Langford, 1995).

이러한 지능형 보간법을 사용하여 인구를 추정할 경우 토지피복도 또는 용도지역도와 같은 토지이용도를 보조자료로 하여 회귀모형을 통해 대상구역의 인구를 추정한다. 하지만 토지이용도를 이용한 회귀모형 방법 역시 여러 가지 문제점이 제기되고 있으며, 이는 크게 회귀모형에서 사용하는 분석자료의 문제점과 회귀모형 자체가 내포하는 분석방법에서의 문제점으로 요약할 수 있다. 우선 분석자료의 문제점에 대해 살펴보면, 위성영상을 이용하여 회귀모형에 필요한 토지이용도나 토지피복도를 구축할 경우, 구축 과정에서 위성영상의 상세한 생물 · 물리적인 정보가 손실된다는 점이다(Jensen, 1983). 따라서 정형화된 구조의 토지이용자료를 이용하여 자세한 인구분포를 추정하기에는 어려움이 있다. 하지만 이러한 문제점은 고 해상도 위성영상 자료를 사용하여 상세한 토지이용자료를 구축함으로써 다소 해결할 수 있었다. 그리고 분석방법 차원과 관련하여 공간자료를 회귀모형에 사용할 경우, 공간자료가 가지는 공간적 자기상관성 때문에 회귀식의 독립변수에 사용되는 표본들의 임의성이 위배되어 회귀식의 설명력에 영향을 줄 수 있다(Griffith and Can, 1996). 이러한 공간자료가 가지는 공간적 자기상관성은 회귀식에 의해 추정된 값과 실측값과의 차이 즉 잔차 값의 분포 패턴을 통해 확인할 수 있다. 따라서 회귀식에 의해 추정된 값과 더불어 예상되는 잔차를 고려할 수 있다면 좀 더 실측치에 근접한 값을 추정할 수 있을 것이다.

본 연구는 이러한 회귀모형 기반의 인구 추정 모형이 갖는 방법론에 있어서 공간적 자기상관을 갖는 잔차의 분포를 고려하여 보간 결과를 향상하고자 하였다. 이를 위해 회귀모형과 회귀모형에서 추출된 잔차를 크리깅을 통해 보간하는 하이브리드 보간법[2]인 Regression-Kriging 모형(이하 RK모형)을 이용하여 인구를 추정하였다. 그리고 이 연구에서 제안하는 RK모형의 타당성을 검증하기 위해 기존의 단순 회귀식을 이용한 방법과 비교 · 분석하였다. 이를 위해 다음과 같이 네 가지 단계로 연구를 진행한다. 첫째, 인구추정과 관련된 연구 동향에 대해 살펴본다. 둘째, 이 연구에서 사용하는 RK모형에 대하여 논한다. 셋째, RK모형의 타당성을 검증하기 위해 연구지역을 대상으로 사례분석을 수행하고 정확도를 평가한다. 넷째, 이 연구가 갖는 시사점과 향후 발전 방향에 대해 제시한다.

2. 연구동향

도시지역의 인구를 추정하는 대부분의 연구는 도시의 토지피복과 인구와의 상관관계를 통해 회귀모형을 이용하여 도시지역의 인구를 추정한다. 특히 인구밀도의 공간 분포 추정을 위해 토지이용과 같은 보조정보를 적극적으로 활용하는 방법으로 대시메트릭(dasymetric mapping)방법이 있는데, 소스 구역 체계를 보다 더 작은 공간 단위로 인구밀도를 분할하여 나타낼 수 있도록 한다(Holt *et al.*, 2004). 인구밀도의 분할을 위한 가중치의 도출을 위해 토지이용을 보조정보로 사용하는데, 토지이용과 인구밀도를 변수로 OLS 회귀식을 적용해 토지이용별 인구밀도 계수를 가중치로 사용하는 방법이 있다(Reibel and Agrawal, 2007). 이러한 방법을 통해 Lee and Kim(2007)은 서울시 코로플러스 매핑보다 밀도면을 보다 잘 반영하는 인구밀도 분포도를 작성하였다. 이와 같이 토지이용은 인구밀도 추정에 있어서 가장 중요한 보조자료이며 공간적인 분포의 정확도를 향상할 수 있는 자료이다.

토지이용을 바탕으로 인구분포 추정을 위하여 회귀모형을 이용하는 방법이 일반적으로 사용되어 왔다(Lo, 1995; Kim, 2006; Ku, 2008; Liu *et al.*, 2008). 이 회귀모형에서 설명력을 높이는 중요한 관건은 위성영상이나 각종 공간데이터를 사용하여 인구가 거주하는 주거지역을 최대한 세분화하고 이들이 거주하는 주거지역을 정확하게 추출해내는 것이다(Donnay and Unwin, 2001). 이를 위해 초기에는 Landsat TM 영상을 이용하여 밴드별 분광특성과 센서스 단위의 인구분포와의 상관관계를 분석하여 회귀모형을 통해 인구를 추정하였다(Harvey, 2002; Chen, 2002). 그러나 Landsat 영상의 경우 공간해상도의 한계 때문에 주거지역을 세분화하는데 많은 어려움이 있었다. 또한 상업지역이나 농업지역 등 인간이 거주하지 않는 지역을 회귀모형에 포함시키면서 분석 과정에서 인구 가중치가 정확하게 계산되지 않는 단점이 있었다.

이러한 점을 보완하기 위해 기존의 중 · 저해상도 위성영상을 이용하여 화소기반분류법으로 토지피복을 분석하는 방법 대신 고해상도 위성영상을 이용하여 객체지향분류법으로 인구가 거주하는 지역을 좀 더 정확히 추출하고 그 정보를 회귀모형에 반영하여 인구를 추정하는 연구가 고해상도 위성영상이 보급된 이후 활발히 진행되어 왔다. Ku(2008)는 IKONOS영상을 이용하여 객체지향분류법으로 토지이용자료를 구축한 후 인구를 추정하는 네 가지 모형인 shotgun 모형, focused 모형, simple 모형, 로그변환모형의 적합성을 검증하였다. 그 결과 로그변환모형이 다른 모형에 비해 예측 정확도가 높게 나온다는 사실을 확인하였고, 이와 함께 인구추정에 있어서 고해상도 위성영상의 활용 가능성을 제시하였다. Liu *et al.*(2006) 역시 고해상도 위성영상인 IKONOS 영상을 이용하여 도시지역을 추출하고 이를 대상으로 경관매트릭스를 방법을 통해 건조지역의 패치 밀도를 고려한 인구추정 회귀모형을 구축하였다. 이 연구 역시 선형모형에 비해 로그모형을 사용할 경우 예측 정확도가 높게 나온 다고 제시하였다. 즉 인구와 도시지역 또는 주거지역과의 관계는 선형 관계 보다는 로그함수 관계에 더 가깝다는 것을 의미하며, 이상의 연구 외에도 다수의 연구에서 이러한 로그함수 관계를 이용하여 인구추정 모형을 구축하였다(Lo, 1995; Kim, 2006).

하지만 회귀모형 기반의 인구추정 모형은 전술한 바와 같이 토지이용도나 토지피복도를 구축하는 과정에서 발생하는 자료의 문제점과 더불어 회귀모형 자체가 가지는 방법론의 문제점이 꾸준히 제기되어 왔다. 즉 회귀모형을 공간자료에 적용할 경우 자료의 공간적자기상관으로 설명계수가 편향 될 수 있으므로 공간적자기상관을 고려할 수 있는 방법을 적용하여야 한다. Wu and Murray(2005)는 이러한 회귀모형의 문제점을 극복하기 위해 Landsat ETM+ 영상에서 추출된 건조지역과 인구자료 간의 상관성을 이용하여 건조지역의 면적과 인구, 두 개의 변수를 공동크리깅(co-kriging) 방법을 통해 인구 분포를 추정하였고, 부분적으로 추정의 정확도를 향상시켰다.

그리고 이 연구에서 사용하는 RK모형과 관련된 연구를 살펴보면 Triantafilis *et al.*(2000)은 회귀모형, 공동크리깅, 3차원 크리깅(three-dimension kriging), 정규크리깅(ordinary kriging), RK모형 등 5가지 모형을 사용하여 토양의 염류도를 추정하였으며, 그 결과 RK

모형의 예측 정확도가 다른 모형에 비해 가장 우수하게 나온 것을 확인할 수 있었다. Eldeiry and Garcia (2009) 역시 Landsat영상을 사용하여 토양의 염류도를 공동크리깅 모형과 RK모형을 사용하여 추정하였으며, 그 결과 RK모형이 공동크리깅 모형에 비해 정확도가 높게 나왔다. 이와 같은 결과가 나타난 원인으로 공동크리깅 모형은 상관관계가 존재하는 모든 밴드들이 보조변수로 사용되었으나, RK모형은 회귀분석을 통해 밴드 간의 상관관계가 존재하는 부분만을 추출하여 이를 보간하는데 사용하기 때문에 예측 정확도가 높게 나왔다고 해석하였다. 즉 단일 방법론을 사용하는 것보다 RK모형과 같이 하나 이상의 방법론을 조합하는 하이브리드 방법이 예측의 정확도를 향상시킬 수 있다는 점을 시사한다.

이상의 선행연구를 종합해보면 인구추정모형의 정확도를 향상시키기 위해서 사용하는 보조자료를 개선하는 연구와 모형에 적용되는 방법론을 발전시키는 연구로 구분할 수 있는데, 전자의 경우 고해상도 위성영상이나 항공사진 등이 보편화 되면서 보조자료의 정확도가 크게 향상되었음을 확인할 수 있었다. 하지만 이에 비해서 분석방법론을 개선하는 연구는 미진한 실정이다. 특히 이 연구와 관련하여서는 기존 연구에서 회귀모형이나 공동크리깅 등 단일 모형만을 사용하였고, 하나 이상의 방법론을 조합하는 하이브리드적 접근에 대한 고려가 부족하였다. 또한 RK모형의 예측 정확성이 높게 나온다는 기존의 연구결과에도 불구하고 RK모형은 토양의 염도나 지질과 같은 자연과학 분야에 한정적으로 사용되어 왔다. 즉 이 연구에서처럼 인구분포를 추정하는 인문사회과학 분야에서의 활용성이 제시되지 못하고 있다는 점을 지적할 수 있다.

3. RK모형 고찰

공간상에서 발생하는 다양한 현상의 패턴을 탐색하는 지리통계(geostatistics)는 지역변수이론(Regionalized Variable Theory)에 기초를 둔다. 지역변수이론에서 특정 변수가 갖는 공간적 변화는 크게 전역적인 변화의 경향을 반영하는 결정론적인 요소($m(x)$)와 결정론적 요소로 설명할 수 없는 확률적인 요소($e'(x)$) 그리고 불규칙적인 변동(e'') 요소로 구성되어있다고 가정한다(Burrough and McDonnell, 1998).

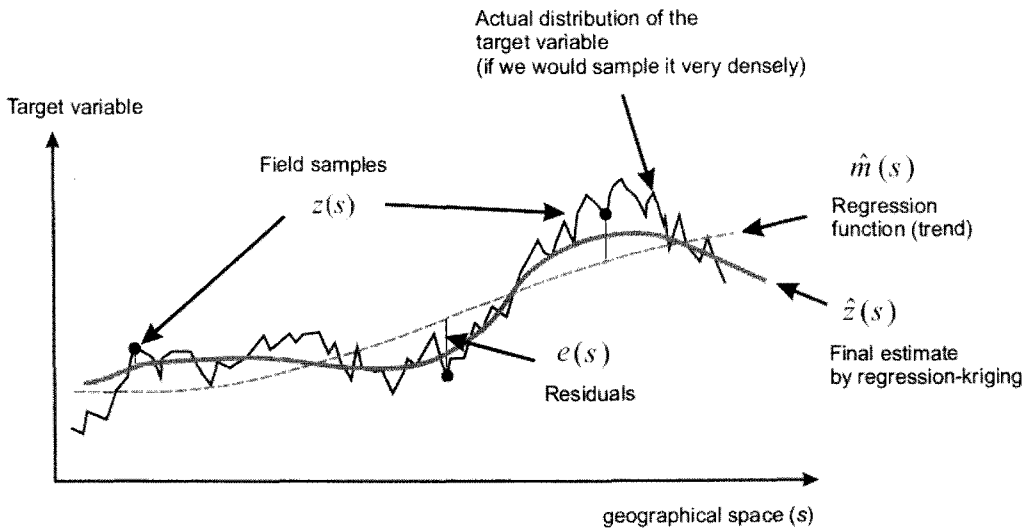

Figure 1. A schematic concept of the regression-kriging model (Source: Hengl, 2009). RK모형 개념도.

$$z(x) = m(x) + e'(x) + e'' \quad (1)$$

일반적으로 이러한 패턴을 분석한 대다수의 연구에서는 점 형태의 관측 자료를 바탕으로 크리깅이나 역거리가중법(inverse distance weighting)과 같은 보간기법을 통해 대상변인(target variable)의 값을 추정하거나 위성영상, 수치고도모델 또는 기타 보조자료를 활용하여 회귀모형으로 대상변인의 값을 추정하였다. 그러나 이러한 개별적인 접근법만으로는 지역변수이론에서 정의하는 각각의 요소를 포괄적으로 추정하는데 한계가 있다. 이러한 배경에서 최근에는 두 개 이상의 보간법을 서로 결합한 하이브리드 모형을 이용하는 다수의 연구가 진행되고 있으며, 하이브리드 모형을 이용할 경우 기존에 단일모형을 사용해서 분석하는 것보다 예측 정확성과 정밀성을 크게 향상시킬 수 있다(Bishop and McBratney, 2001). 하이브리드 모형 중 회귀모형과 크리깅모형을 결합하여 추정치의 정확도를 향상 시키는 모형으로 RK모형[3]이 있다.

이러한 RK모형은 결정론적인 부분과 확률론적인 요소를 개별적으로 분석하고 이를 종합하는 구조다. Figure 1에서처럼 회귀모형을 이용해 결정론적인 요소를 분석하고 회귀모형의 잔차 즉 확률론적인 요소를 크리깅을 이용하여 추정한다. 이와 같은 지역변수이론에 RK모형의 기본구조를 접목시키면 식2와 같다(Hengl *et al.*, 2003).

$$\hat{z}(s_0) = \hat{m}(s_0) + \hat{e}(s_0)$$
$$= \sum_{k=0}^{p} \hat{\beta}_k q_k(s_0) + \sum_{i=1}^{n} \lambda_i e(s_i) \quad (2)$$

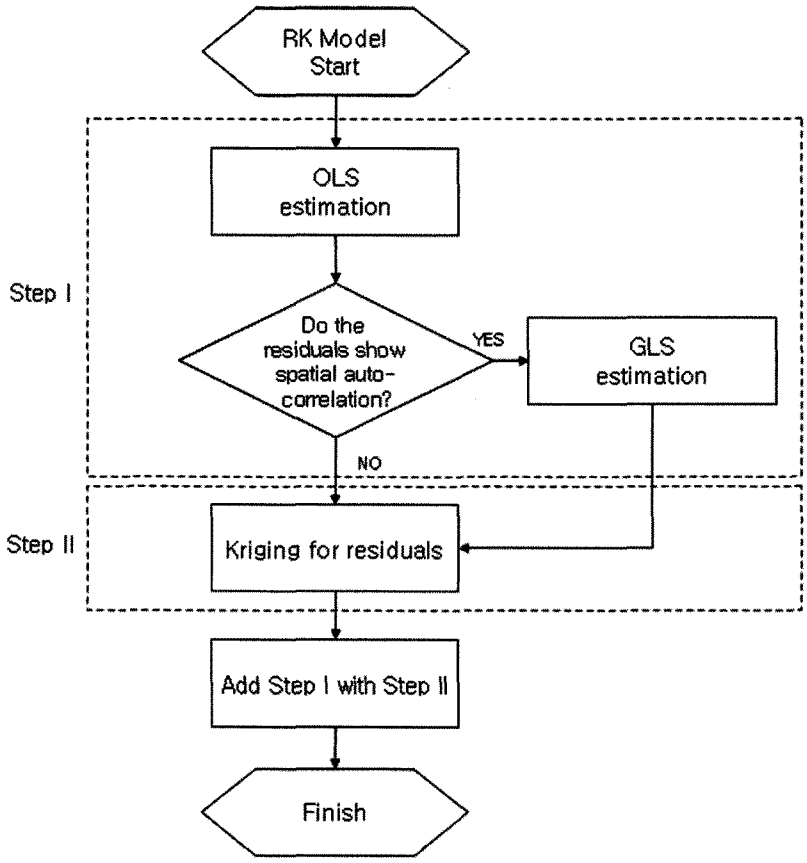

Figure 2. Process flow of RK model. RK모형의 분석 흐름도.

RK모형(식2)에서 $\sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0)$는 전체적인 변화 경향을 분석하는 부분으로 회귀모형 형태를 취한다. 여기서 $q_k(s_0)$는 s_0지점에 대한 독립변수의 값이고, $\hat{\beta}_k$는 회귀계수, p는 모형에서 사용하는 독립변수의 수를 가리킨다. 그리고 $\sum_{i=1}^{n} \lambda_i \cdot e(s_i)$는 회귀분석의 잔차를 단순크리깅을 이용하여 추정하는 부분이다. 여기서 λ_i는 예측오차의 분산을 최소화하는 크리깅 가중치이고 $e(s_i)$는 s_i지점의 잔차값이다.

일반적인 회귀모형에서 회귀계수 s_i는 최소자승법(OLS: Ordinary Least Squares)을 이용하여 추정한다. 이러한 최소자승법을 이용한 회귀모형에서 오차의 기본가정인 선형성, 등분산성, 독립성, 정규성 등 임의오차 성질을 충족시켜야만 회귀모형의 타당성을 보장받을 수 있다. 그러나 공간 자료와 같이 공간적 자기상관성이 존재할 경우 예측 값이 편향(biased)될 수 있으며, 이 경우 전술한 기본 가정을 수용할 수 없기 때문에 최소자승법은 효율적인 모형이라 할 수 없다. RK모형에서는 이러한 공간적 자기상관성 문제점을 고려한 회귀모형 가운데 하나로 일반 최소자승법(GLS: Generalized Least Squares)을 사용하며 그 구조는 아래와 같다.

$$\hat{\beta}_{GLS}=(q^T \cdot C^{-1} \cdot q)^{-1} q^T \cdot C^{-1} \cdot z \quad (3)$$

여기서 역행렬 C^{-1}는 잔차의 분산-공분산 행렬로서, 가중치 역할을 수행한다. 즉 잔차의 크기에 따라 가중치를 상이하게 적용하여 공간적 자기상관성에 따른 오차의 이분산성(heteroskedasticity)을 일정한 분

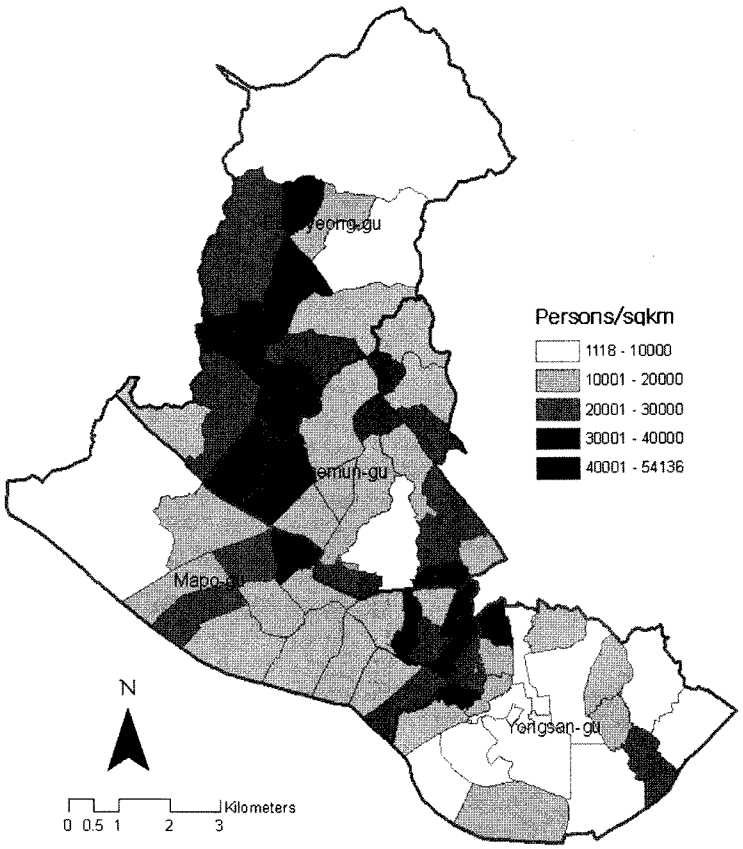

Figure 3. Population distribution for study area, 2005. 연구지역 인구분포도(2005년).

산을 갖도록 변환한다. 따라서 잔차에 공간적 자기상관성이 존재하지 않을 경우 일반 최소자승법은 최소자승법과 같다. 이러한 회귀모형 부분에서 나온 잔차는 다시 단순크리깅을 이용하여 보간하며 최종적으로 두 모형에서 나온 결과를 합산하여 대상변인의 값을 추정한다. 이러한 RK모형의 전체 구조를 행렬 형태로 나타내면 다음과 같다.

$$\hat{z}(s_0)=q_0^T\cdot\hat{\beta}_{GLS}+\lambda_0^T\cdot(z-q^T\hat{\beta}_{GLS}) \quad (4)$$

이상의 RK모형의 전체 흐름을 도식화하면 Figure 2와 같다.

4. 실증분석

1) 사례지역 분석

RK모형의 실효성을 분석하기 위해 서울시의 인접한 네 개의 자치구인 은평구, 서대문구, 마포구, 용산구를 대상으로 사례분석을 하였다. 사례분석에서 사용하는 분석 도구로는 회귀분석 및 공간통계는 R 2.11.1[4]을 사용하였고, 공간자료를 시각화하기 위해서 ArcGIS 9.1을 사용했다. 분석 자료는 한국토지정보시스템(KLIS: Korea Land Information Systems)의 2005년 용도지역지구자료와 환경부에서 제공하는 2000년대 중분류 토지피복분류도 그리고 서울시에서 제공하는 2005년 인구자료를 사용했다. 인구자료를 이용하여 연구지역의 동별 인구분포 현황을 도식화하면 Figure 3과 같다.

회귀모형을 이용하여 인구를 추정하는데 가장 중요한 것은 전술한 바와 같이 인구가 거주하는 지역을 정확하게 추출하는 것이다. 즉 용도지역도만을 이용하여 용도별 인구를 추정할 경우 계획 상 용도지역이 주거지역이라도 실제 토지피복은 초지, 농지와 같이 인간이 거주하지 않는 지역일 경우가 있기 때문에 단순히 용도지역도만을 이용하여 실제 인구가 거주하는 주거지역을 추출하는데 한계가 있다. 이러한 점을 극복하기 위해 토지피복의 건조지역과 용도지역의 1종 · 2종 · 3종 주거지역[5]이 서로 중첩되는 지역만을 추출하였다. 즉 토지피복이 시가화 건조지역이면서 동시에 용도지역이 주거지역인 지역을 추출하였다. 그리고 회귀분석 후 잔차에 대한 공분산 구조와 크리깅 분석을 수행하기 위해서는 각 동의 인구를 대변하는 중심점을 추출해야 하는데, 이것 역시 단순히 동의 중심점이 아닌 중첩분석을 통해 도출된 지역이 포함되도록 하여 중심점을 추출하였다(Figure 4)[6].

이렇게 구축된 공간 자료와 센서스 인구 자료를 바탕으로 회귀분석을 수행하였으며, 사용된 회귀모형은 인구 추정 회귀모형 가운데 가장 설명력이 높은 로그변환모형을 사용하였다(Ku, 2008; Kim, 2006; Lo, 1995).

$$\begin{aligned}\log\text{인구}=&\log(0.0536\cdot 1\text{종 주거지역 면적})\\&+\log(0.6432\cdot 2\text{종 주거지역 면적})\\&+\log(0.1003\cdot 3\text{종 주거지역 면적})\end{aligned} \quad (5)$$

식5에 대한 유의성을 나타내는 F값은 440.4 ($p<2.2e-16$)로 인구에 대한 주거지역 설명변수들에 대한 회귀식은 통계적으로 유의하였다. 또한 결정계수 값은 84.79% 였는데, 결정계수는 회귀식이 종속변수인 인구에 대해 설명하는 정도를 나타내므로 식5는 인구에 대해 약 85% 설명하고 있다. 특히 회귀모형에서 상수항의 포함 여부는 설명변수와 종속변수 간의 물리적 관계에 의해 판단할 수 있는데, 이 연구에서는 사용된 독립변수 외의 다른 지역에는 인구가 거주하지 않는다고 가정(Reibel and Agrawal, 2007)한 것이 논리적이라 판단하여, 상수항을 회귀식에서 제외하고 회귀선을 원점에 적합시켰다. 이 경우 R^2가 증가하는 것이 대부분이다(Choi, 2007). 따라서 회귀모형 간의 차이에 대한 비교를 위해서는 R^2보다 잔차제곱합을 비교하는 것이 더 적절하다. 따라서 이 연구에서는 잔차제곱합을 고려한 AIC 통계량을 이용하여 모형간의 비교 · 분석을 수행하였다.

식5와 같이 일반최소자승법을 이용한 회귀모형의 잔차에 대한 공간적 자기상관성을 분석하기 위해 세미베리오그램(semivariogram)을 사용하였으며, 이를 통

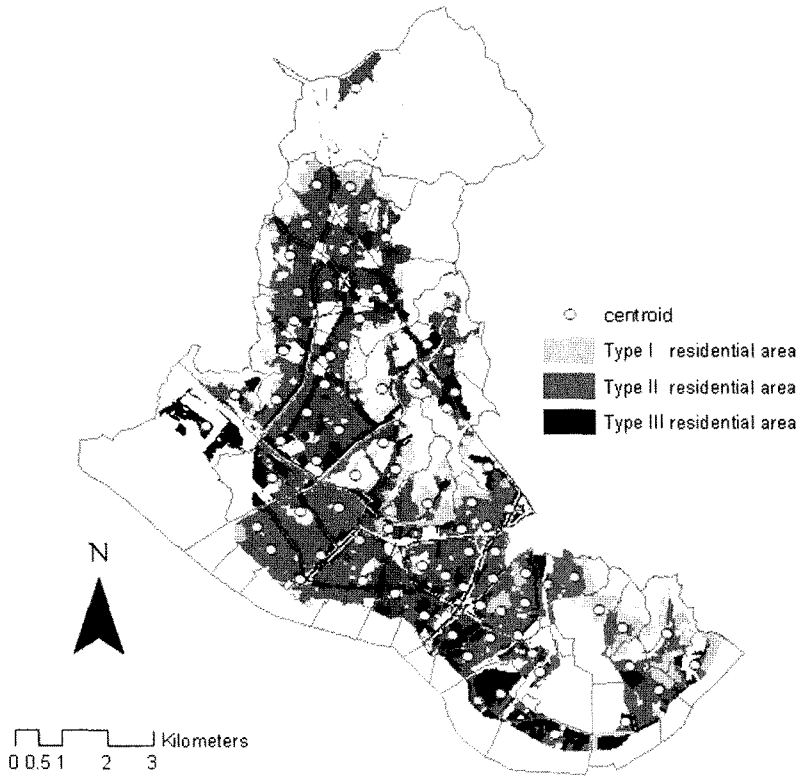

Figure 4. Centroid of residential areas. 주거지역 중심점.

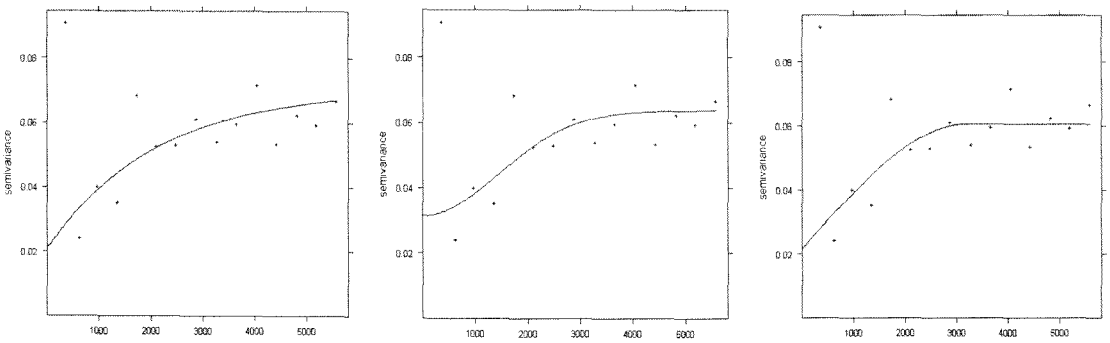

Figure 5. Semi-variograms for regression residuals. 잔차에 대한 이론적 세미베리오그램.

해 회귀모형에 잔차의 공분산 구조를 모델링하였다. 이를 위해 우선 실험적 세미베리오그램을 통해 상관거리(range)를 약 6,000m, 너깃(nugget)은 0.01 그리고 문턱값(sill)을 0.01~0.06으로 설정하였고[7], 이러한 실험적 세미베리오그램의 인자 값을 바탕으로 이론적 세미베리오그램 모델링을 수행했다. 이 연구에서는 지수모형, 가우시안모형, 구형모형 등 세 가지 이론적 세미베리오그램을 사용하였다(Figure 5).

Table 1. Estimated coefficients for each models and AIC. 모형별 추정계수 및 AIC.

Model		Estimated coefficients			AIC
		Type I Residential area	Type II Residential area	Type III Residential area	
OLS		0.0536** (0.023)	0.6432*** (0.000)	0.1003*** (0.000)	14.03
GLS	Exp. Model	0.0445* (0.065)	0.6524*** (0.000)	0.0990*** (0.000)	9.98
	Gaus. Model	0.0457* (0.062)	0.6516*** (0.000)	0.1007*** (0.000)	9.52
	Sph. Model	0.0467** (0.044)	0.6483*** (0.000)	0.1025*** (0.000)	9.19

Signif. codes: ***: 0, **: 0.05, *: 0.1

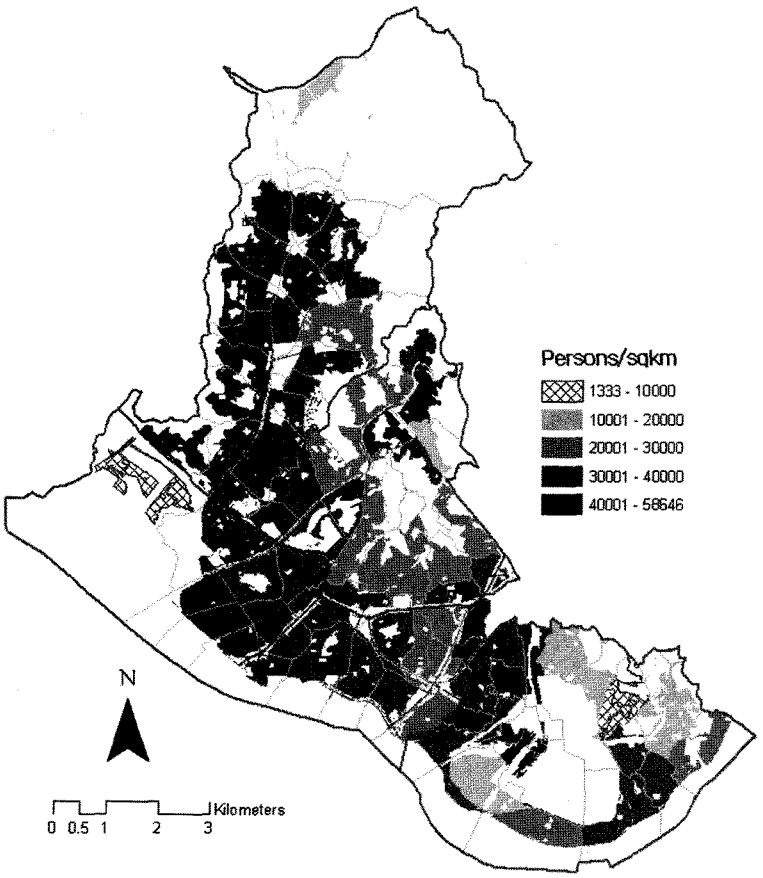

Figure 6. Population distribution for residential area using RK model. RK모형을 이용한 주거지역 인구분포도.

이러한 3가지 모형에서 나온 잔차의 공분산 구조를 가중치로 하여 일반 최소자승법을 수행하였으며, 각각의 모형 가운데 최적의 모형을 선택하기 위해 모형의 설명변수의 개수와 잔차제곱합을 고려한 Akaike 정보기준(Akaike Information Crietrion, 이하 AIC)[8]값을 사용하였다. 각각의 모형에 대한 분석 결과를 요약하면 Table 1과 같다.

지수 모형과 가우시안 모형의 경우 1종 주거지역 변수가 0.1 수준에서 유의적으로 나타났으며, 구형 모형과 최소자승법 모형은 0.05 수준에서 유의적인 것으로 분석됐다. 반면에 AIC 값을 비교했을 때, 최소자승법 모형이 가장 낮게 나왔으며 구형 모형이 가장 높게 나왔다. 즉 구형 모형을 이용한 일반 최소자승법이 네 가지 모형 가운데 상대적으로 가장 우수한 예측 결과를 나타내는 모형이라 할 수 있겠다.

마지막으로 구형 모형 기반의 일반 최소자승법을 이용한 추정값의 잔차값에 대하여 단순 크리깅을 수행한다. 이를 위해 잔차에 대한 세미베리오그램을 모델링하였다. 이를 위해 실험적 세미베리오그램을 수행하였으며, 그 형태가 Figure 5와 매우 유사했다. 이것은 최소자승법에서 추출된 잔차의 공분산 구조는 일반 최소자승법에서 나온 잔차의 공분산 구조와 거의 유사한 형태를 보인다는 다른 RK모형 연구에서 나온 결과와 동일했다(Hengl *et al.*, 2007). 따라서 세미베리오그램 인자값은 최소자승법과 동일하게 설정하였으며, 구형 모형을 이용하여 이론적 세미베리오그램을 모델링한 후 이를 바탕으로 단순 크리깅을 수행했다. 단순 크리깅을 수행한 결과 인구수가 음수로 나온 경우 모두 영(零)으로 변환하였으며, 이렇게 변환된 최종 크리깅 결과를 일반 최소자승법에서 추정한 결과와 합산하여 주거지역 인구와 인구분포를 추정하였다. RK모형을 이용한 최종 인구분포 추정도는 Figure 6과 같다.

2) 정확도 평가

이 연구는 기존의 회귀모형을 이용한 인구추정 모형의 정확도를 향상시키는데 목적이 있다. 이러한 관점에서 정확도 평가는 RK모형과 회귀모형(식5) 두 가지 모형에 대하여 인구추정 결과를 바탕으로 정확도를 비교하였다. 이를 위해 주거지역 단위로 추정한 인구를 센서스 단위로 집계한 후 그 결과를 센서스 자료와 비교하였으며, 비교 방법으로는 인구 간의 관계를 나타내는 상관관계 계수(ρ) 그리고 일반적으로 많이 사용하고 있는 평균제곱근오차(RMSE: root mean squared error)와 평균절대오차(MAE: mean absolute error) 통계량을 사용했다(식6).

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{P}_i-P_i)^2}\ ,$$

$$MAE=\sqrt{\frac{1}{P}\sum_{i=1}^{N}|P_i-\hat{P}_i|} \tag{6}$$

이와 함께 모형의 예측 효율성을 평가하는 방법으로 식7과 같은 G통계량을 사용했다(Eldeiry and Garcia, 2009). G값은 모형의 예측 결과가 샘플의 평균값을 사용했을 경우와 비교해서 어느 정도 효율성을 가지고 있는가를 평가하는 통계량이다. 만약 G값이 1과 같으면 모형의 예측 결과가 매우 정확하다는 것을 의미하고, G값이 양수이면 평균을 사용한 것 보다 더 신뢰성이 높고 반면에 음수이면 신뢰성이 오히려 샘플의 평균을 사용했을 때 보다 더 낮다는 것을 의미한다. 그리고 G값이 0일 경우 샘플의 평균을 사용한 것과 동일하다는 것을 의미한다.

$$G=\left[1-\left\{\sum_{i=1}^{N}(\hat{P}_i-P_i)^2/\sum_{i=1}^{n}(\hat{P}_i-\bar{P})^2\right\}\right] \tag{7}$$

이러한 네 가지 통계량을 사용하여 RK모형과 회귀모형의 정확도를 평가한 결과를 요약하면 Table 2와 같다. 평균제곱근오차와 평균절대오차 값을 보면 전체적으로 회귀모형에 비해서 RK모형을 이용하여 추정한 결과의 정확도가 비교적 높게 나왔다. 그리고 센서스 자료와의 상관관계를 비교했을 때 RK모형은 0.84로 매우 높은 상관관계가 나왔으며 회귀모형과 비교 했을 때 약 0.2 정도 높게 나왔다. 또한 모형의 효율성을 평가하는 G값의 경우 회귀모형과 RK모형 모두 양의 값을 가지고 있었으나, RK모형이 회귀모형에 비해 1에 더 가깝게 나온 것을 확인할 수 있었다.

마지막으로 연구지역별로 추정 오차의 크기를 살펴

Table 2. Accuracy assessment of RK model and regression model estimations.
RK모형과 회귀모형에 대한 정확도 평가.

Model	Accuracy assessment methods			
	ρ	RMSE	MAE	G
RK Model	0.838	4,272.80	18.9%	0.436
Regression Model	0.623	6,488.65	30.0%	0.144

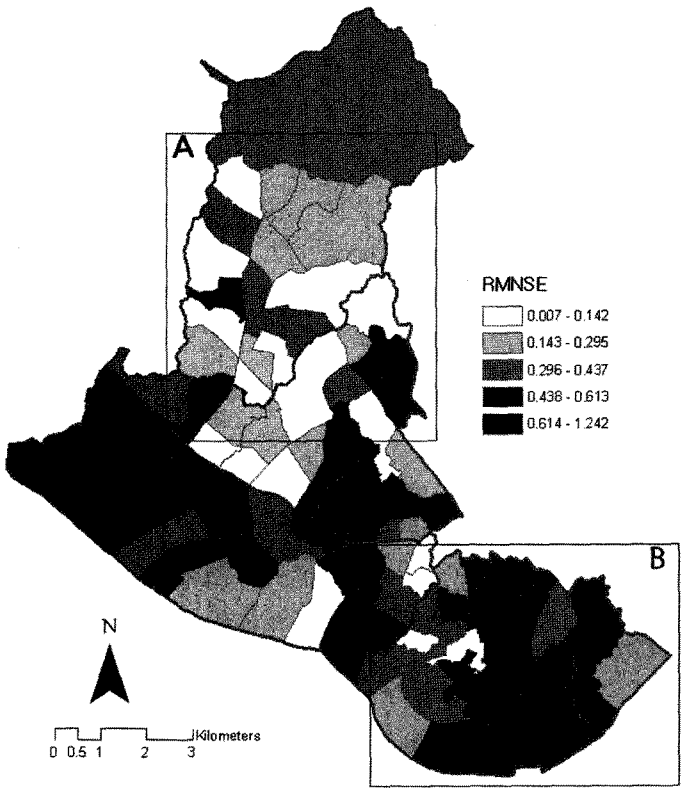

Figure 7. High-density area(A) and low-density area(B). 연구지역의 고밀도지역(A)과 저밀도지역(B).

보기 위해 평균제곱근오차를 표준화한 표준화된 평균제곱근 오차값(RMNSE: root mean normalized squared error)을 이용하여 연구지역을 도식화하였으며, 그 결과는 Figure 7과 같다.

Figure 7에서 상자A는 은평구와 서대문구 일대의 고밀도 지역을 나타내고 상자B는 저밀도 지역인 용산구 지역이다. 고밀도 지역과 저밀도 지역을 시각적으로 비교했을 때, 동의 면적이 크고 밀도가 낮은 저밀도 지역에 오차값이 높게 나왔다. 특히 이 지역은 연구지역 가장자리에 위치하거나 또는 다른 지역에 비해 접하는 면이 적어서 주변에 표본의 수가 상대적으로 적은 지역들이다. 이러한 현상은 RK모형 외에도 다른 보간기법에서도 빈번하게 발생하는 현상으로 RK모형 역시 이와 같은 일반적인 보간기법의 문제점을 내포하고 있는 것으로 판단된다. 따라서 인구추정 모형에 있어서 이러한 문제점을 해결하기 위해서는 RK모형을 비롯한

다양한 면 보간기법(areal interpolation)들 간의 장단점을 면밀히 비교·분석하여 보간기법에서 발생하는 이상의 문제점을 보완할 수 있는 새로운 형태의 보간기법을 개발하는 연구가 앞으로 진행되어야 할 것이다.

5. 결론

인구자료는 다양한 분야에서 빈번하게 사용하는 자료이지만 센서스 단위로 집계하여 제공되는 인구자료를 적절한 가공 과정이 없이 사용할 경우, 분석 시 다양한 문제점이 발생하고 있다. 이를 위해 토지이용도나 위성영상 등을 이용하여 회귀모형을 통해 토지이용단위로 인구를 추정하고 있으나, 회귀모형만을 이용할 경우 자료의 공간적자기상관의 문제, 잔차의 추정 문제 등으로 인해 높은 예측 정확도를 기대하기 어렵다. 이러한 맥락에서 본 연구에서는 기존의 회귀모형에서 공간적자기상관을 고려할 수 있도록 하고 추출된 잔차를 단순 크리깅을 통해 보간하는 RK모형을 제안하였다. 이러한 RK모형의 활용성을 검증하기 위해 서울에 인접한 네 곳의 자치구인 은평구, 서대문구, 마포구, 용산구를 대상으로 실증 분석을 수행하였다. 그리고 RK모형과 기존의 회귀모형의 예측 정확도를 비교하기 위해 정확도 평가를 수행하였다. 정확도 평가 결과 평균제곱근오차와 평균절대오차 값이 회귀모형에 비추어 RK모형이 높게 나왔으며, 모형의 효율성을 나타내는 G값 역시 RK모형이 더 적합하게 나온 것을 확인할 수 있었다. 즉 도시지역의 인구추정에 있어서 모형의 예측 정확도 그리고 모형의 효율성 측면에서 RK모형이 회귀모형에 비해 우수하게 나왔다.

RK모형이 기존의 회귀모형에 비해 예측 정확도를 향상시킨다는 점에서 그 활용성이 높을 것으로 판단되나, RK모형이 회귀모형에 비해 분석과정이 복잡하다는 단점이 있다. 즉 회귀분석과 크리깅 분석 과정에서 개별적으로 베리오그램 모델링을 수행해야 하고, 그 결과를 바탕으로 회귀분석과 크리깅 분석을 해야 하는 복잡한 과정을 거쳐야 한다. 이러한 계산 과정의 복잡성은 RK모형을 비롯한 대부분의 하이브리드 모형이 가지고 있는 공통적인 문제점이라 할 수 있겠다. 또한 RK모형은 아직까지 지리통계나 GIS 응용프로그램에서 완벽하게 지원하지 않는다는 점도 RK모형의 활용성에 있어서의 문제점이라 할 수 있겠다. 실제로 RK모형을 비롯한 하이브리드 형태의 모형이 단일 모형에 비해 예측 정확도가 높다는 것은 이미 다수의 연구에서 검증 되었으나, 그 구조가 복잡하고 지원하는 응용프로그램이 많지 않기 때문에 다양한 연구에서 활발하게 활용되고 있지 않는 것이 현실이다.

그러나 최근에 들어서는 R통계 패키지를 비롯한 다양한 응용프로그램들이 웹 커뮤니티를 통한 오픈소스를 지향하는 추세이므로, 이러한 플랫폼에 RK모형과 같은 다수의 하이브리드모형을 지리통계 학자들과 전문 프로그래머들이 개발하고 보급한다면 다양한 분야에서 활발하게 활용될 수 있을 것으로 판단된다.

주

1) MAUP은 공간 단위가 수정 가능하다는 의미로 보통 분석 단위 혹은 분석 스케일의 선택이 분석 결과에 매우 유의적인 영향을 미친다는 것을 의미한다(Openshaw, 1984).

2) 하이브리드 보간법(hybrid interpolation)은 지능형 보간법처럼 보조자료를 사용함과 동시에 서로 다른 보간 기법을 하나로 융합한 형태의 보간법을 말한다.

3) RK모형은 1987년 Ahmed and de Marsily(1987)가 처음으로 이러한 개념을 연구에 도입한 후 Odeh *et al.*(1994)가 Regression-Kriging이라는 이름으로 모형에 명칭을 부여했다.

4) R은 다양한 통계 기법과 수치해석을 지원하는 공개소프트웨어 통계 분석 도구로써, 각 분야의 통계 전문가들이 개발한 프로그램을 라이브러리 형태로 무료로 다운받아 사용할 수 있다는 장점이 있다. 특히 최근에는 지리학에서 많이 사용하는 공간통계와 관련된 다수의 라이브러리를 제공하고 있으며, 기존의 g-stat도 하나의 라이브러리로 제공하고 있다.

5) 일반주거지역은 주거 형태와 밀도에 따라 1~3종으로 구분하는데, 1종일반주거지역은 저층 중심의 저밀도 주거지역이며, 2종일반주거지역은 중층중심의 중밀도 주거지역, 3종일반주거지역은 종/고층 중심의 고밀도 주거지역을 의미한다.

6) 인구분포와 같이 모든 지역에 균등하게 분포하지 않고 특정 지역에 밀집되어 나타나는 자료를 면 보간하기 위해서는 단순히 주거지역이 차지하는 면적을 이용하여 인구를 추정하는 면적 가중방법을 사용하기 보다는 폴리곤의 중심점을 이용하여 점 간의 거리를 고려한 연속적인 인구 밀도면을 생성하는 면 보간 방법(density surface creation)을 사용해야 한다.

7) 실험적 세미베리오그램 분석 과정에서 방향에 따른 비등방성(anisotrophy)을 45도 단위로 측정하였으나, 각 방향에 대하여 세미베리오그램에 큰 변화는 보이지 않았다. 따라서 등방성(isotrophy)이란 가정 하에 실험적 세미베리오그램을 모델링하였다.

8) Akaike 정보기준은 1973년 Akaike가 제안한 지수로 모형의 설명력과 크기를 동시에 고려하여 모형 간의 우수성을 판단하는 모형 선택 기준으로 오래전부터 다양한 분야에서 사용되어 왔다. AIC 지수 값이 작을수록 적은 수의 독립변수로도 양호한 예측결과를 산출하는 간명하고 좋은 모형을 의미하는데, AIC 지수에 구체적인 형태는 다음과 같다.

$AIC=\log(\frac{RSS}{n})+\frac{2}{n}k$, RSS: 잔차 제곱의 합, n: 관측치의 수, k: 설명변수의 개수

참고문헌

Ahmed, S. and de Marsily, G., 1987, Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity, *Water Resources Research*, 23(9), 1717-1737.

Bishop, T. F. A. and McBratney, A. B., 2001, A comparison of prediction methods for the creation of field-extent soil property maps, *Geoderma*, 103, 151-162.

Burrough, P. and McDonnell, R., 1998, *Principles of Geographical Information Systems*, Oxford University Press, Oxford.

Chen, K., 2002, An approach to linking remotely sensed data and areal census data, *International Journal of Remote Sensing*, 23, 37-48.

Choi, J. G., 2007, *Geostatistics*, Sigmapress, Seoul (최종근, 2007, 지구통계학, 시그마프레스).

Donnay, J. P. and Unwin, D., 2001, *Modelling Geographical Distributions in Urban Areas, Remote Sensing and Urban Analysis*, Taylor and Francis, New York, 205-224.

Eldeiry, A. and Garcia, L. A., 2009, Comparison of regression kriging and cokriging techniques to estimate soil salinity using landsat images, *Hydrology Days*, 27-38.

Fisher, P. F. and Langford, M., 1995, Modeling the errors in areal interpolation between zonal systems by Monte Carlo simulation, *Environment and Planning A*, 27, 211-224.

Goodchild, M., Anselin, L., and Deichmann, U., 1993, A framework for the areal interpolation of socioeconomic data, *Environment and Planning A*, 25, 383-397.

Griffith, D. A. and Can, A., 1996, Spatial statistical/econometric version of simple urban population density models, in Arlinghaus, S. L. and Griffith, D. A. (eds.), *Practical Handbook of Spatial Statistics*, CRC Press.

Harvey, J. T., 2002, Estimating census district populations from satellite imagery: Some approaches and limitations, *International Journal of Remote Sensing*, 23, 2071-2095.

Hengl, T. 2009, *A Practical Guide to Geostatistical Mapping*, Lulu Enterprises, Inc.

Hengl, T., Heuvelink, G. B. M., and Rossiter D. G., 2007, About regression-kriging: From equations to case studies, *Computer & Geosciences*, 33, 1301-1315.

Hengl, T., Heuvelink, G. B. M., and Stein, A., 2003, *Comparison of Kriging with External Drift and Regression-kriging*, Technical note, ITC.

Holt, J. B., Lo, C. P., and Hodler, T. W., 2004, Dasymetric estimation of population density and areal interpolation of census data, *Cartography and Geographic Information Science*, 31(2), 103-121.

Jensen, J. R., 1983, Estimating census district populations from satellite imagery: Some approaches and densities, *Transactions in GIS*, 4(3), 217-234.

Kim, H., 2006, Population estimation using land use and land cover data from Landsat TM images, *The*

Geographical Journal of Korea, 40(4), 489-496.

Ku, C. Y., 2008, A Study on estimating the population in urban area with high resolution satellite image, *The Geographical Journal of Korea*, 42(1), 137-148.

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*, 7(2), 53-67.

Liu, X. H., Kyriakidis, P. C., and Goodchild, M. F., 2008, Population-density estimation using regression and area-to-point residual kriging, *International Journal of Geographical Information Science*, 22(4), 431-447.

Liu, X. H., Clark, K., and Herold, M., 2006, Population density and image texture: A comparison study, *Photogrammetric Engineering & Remote Sensing*, 72(2), 187-196.

Lo, C. P., 1995, Automated population and dwelling unit estimation from high-resolution satellite images: A GIS approach, *International Journal of Remote Sensing*, 16(1), 17-34.

Odeh, I., McBratney, A., and Chittleborough, D., 1994, Spatial prediction of soil properties from landform attributes derived from a digital elevation model, *Geoderma*, 63(3-4), 197-214.

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(1), 93-106.

Openshaw, S., 1984, The modifiable areal unit problem, *Concepts and Techniques in Modern Geography*, 39(Norwich, UK: Geobooks).

Reibel, M. and Bufalino, M. E., 2005, Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems, *Environment and Planning A*, 37(1), 127-139.

Triantafilis, J., Odeh, I. O. A., and McBrantney, A. B., 2000, Five geostatistical models to predict soil salinity from electromagnetic induction data across irrigated cotton, *Soil Science Society of America Journal*, 65(3), 869-878.

Wu, C. and Murray, A. T., 2005, A cokriging method for estimating population density in urban areas, *Computer, Environment and Urban Systems*, 29, 558-579.

교신: 최진무, 110-743, 서울시 종로구 홍지동 7 상명대학교 인문사회과학대학 지리학과(이메일: jmchoi@smu.ac.kr, 전화: 02-2287-5328)

Correspondence: Jinmu Choi, Department of Geography, College of Humanities & Social Science, Sangmyung University, 7, Hongji-dong, Jongno-gu, Seoul, 110-743, Korea (e-mail: jmchoi@smu.ac.kr, phone: +82-2-2287-5328)

최초투고일 2010. 11. 9
수정일 2010. 12. 14
최종접수일 2010. 12. 15