대한지리학회지 제49권 제6호 2014(917~934)

GWL을 적용한 공간 헤도닉 모델링

진찬우* · 이건학**

Spatial Hedonic Modeling using Geographically Weighted LASSO Model

Chanwoo Jin* · Gunhak Lee**

요약 : 지리가중회귀 모델(GWR)은 국지적으로 이질적인 부동산 가격을 추정할 수 있는 도구로 폭넓게 활용되어 왔다. 그럼에도 불구하고 GWR은 공간적으로 이질적인 가격결정요인의 선택이나 국지적 추정에서의 관측치 수의 제한 등과 같은 한계를 가지고 있다. 본 연구는 이러한 한계를 극복하기 위한 대안으로 최근 주목받고 있는 지리가중라소 모델(GWL)을 이용하여 국지적으로 다양한 부동산 가격결정요인들을 탐색하고, 부동산 가격 추정에 있어서 GWL 모델의 적용가능성을 살펴보고자 한다. 이를 위해 서울시 아파트 가격을 대상으로 OLS, GWR, GWL의 헤도닉 모델을 구축하였으며, 모델의 설명력, 예측력, 다중공선성 측면에서 이들을 비교·분석하였다. 그 결과, 전역적 모델에 비해 국지적 모델이 전체적인 설명력, 예측력이 우수한 것으로 나타났으며, 특히 국지적 모델 중 GWL 모델은 다중공선성 문제를 자동적으로 해결하면서 공간적으로 이질적인 가격결정요인 집합들을 도출하였고, 다른 모델들에 비해 상당히 높은 설명력과 예측력을 보여주고 있다. 본 연구에서 적용한 GWL 모델은 고차원의 데이터셋에서 유의미한 독립 변수들을 효율적으로 선정하는데 직접적인 도움을 줌으로써 부동산과 같이 대용량의 복잡한 구조를 가진 공간 빅데이터를 위한 유용한 분석 기법으로 활용될 수 있을 것이다.

주요어 : 공간 헤도닉 모델, 공간적 이질성, LASSO, 지리가중회귀 모델, 지리가중라소 모델, 아파트 가격

Abstract : Geographically weighted regression(GWR) model has been widely used to estimate spatially heterogeneous real estate prices. The GWR model, however, has some limitations of the selection of different price determinants over space and the restricted number of observations for local estimation. Alternatively, the geographically weighted LASSO(GWL) model has been recently introduced and received a growing interest. In this paper, we attempt to explore various local price determinants for the real estate by utilizing the GWL and its applicability to forecasting the real estate price. To do this, we developed the three hedonic models of OLS, GWR, and GWL focusing on the sales price of apartments in Seoul and compared those models in terms of model fit, prediction, and multicollinearity. As a result, local models appeared to be better than the global OLS on the whole, and in particular, the GWL appeared to be more explanatory and predictable than other models. Moreover, the GWL enabled to provide spatially different sets of price determinants which no multicollinearity exists. The GWL helps select the significant sets of independent variables from a high dimensional dataset, and hence will be a useful technique for

이 논문은 교육부와 한국연구재단의 BK21플러스 사업(4-Zero지향 국토공간창조 사업단, 서울대학교 지리학과)의 지원을 받아 수행된 연구결과임

* 서울대학교 지리학과 석사과정(MA student, Department of Geography, Seoul National University), cwjin1108@gmail.com

** 서울대학교 지리학과 조교수 및 국토문제연구소 겸무연구원(Assistant Professor, Department of Geography and Institute for Korean Regional Studies, Seoul National University), gunhlee@snu.ac.kr

large and complex spatial big data.

Key Words : spatial hedonic model, spatial heterogeneity, least absolute shrinkage and selection operator (LASSO), geographically weighted regression(GWR), geographically weighted LASSO(GWL), apartment sales price

1. 서론

부동산 가격은 도시의 다양한 사회 경제 활동의 상호작용을 반영하는 지표이자 개인의 토지 거래, 국세 및 지방세의 과세를 위한 기초 자료로 폭넓게 활용되고 있다(이건학 · 김감영, 2013). 이러한 부동산 가격의 정확한 예측과 적절한 정책 수행은 효율적인 도시 및 지역계획에 필수적인 요소라 할 수 있다. 부동산 중에서도 아파트가 주택에서 차지하는 비중이 58%(2010년 기준)에 달하는 우리나라에서는 대부분의 부동산 정책이 아파트 가격 안정화에 초점을 맞추고 있다. 우리나라 부동산은 1997년 외환위기 이후 부동산 가격의 등락 폭이 심해지고, 그 진폭이 지역별로 매우 이질적으로 나타나고 있어 시장예측이 보다 어려워지고 있다(김연미, 2008). 이는 과거에 비해 부동산의 지역적 특성과 맥락이 부동산 가격에 미치는 영향이 강해짐에 따라 국지적 요인들이 보다 중요해졌음을 의미한다.

그 동안 과세와 지역 계획 등의 기초자료로 활용될 부동산 가격을 추정하기 위해 다수의 연구들이 수행되어 왔지만 추정 모델에 사용되는 변수 제한과 부동산 가격의 공간적 맥락에 대한 이해 부족 등으로 인해 추정 결과들이 실제 현실에 적용되는데 제한적이었고, 모델에 의존적인 결과 해석에 치중하였다. 우리나라의 부동산 가격 추정에 관한 주요 연구 동향을 살펴보면, 부동산 가격을 토지 및 주택의 구조적 특성, 입지 요인, 근린 특성 등의 함수적 관계로 정의하는 전통적인 헤도닉 접근(hedonic approach)(Rosen, 1974)에서 공간적 의존성(spatial dependency), 공간적 이질성(spatial heterogeneity)과 같은 외부 공간 효과를 명시적으로 고려하는 공간계량경제학(spatial econometrics) 측면의 연구들이 활발히 이루어졌다. 본 연구에서 주목하는 아파트의 경우, 아파트의 구조적 특성(단지 규모, 평형 등), 지하철과의 접근성, 조망권, 심미적 요인, 아파트 브랜드 가치 등과 같은 아파트 자체의 특성이나 입지적 요인을 전통적인 회귀 모델을 통해 분석한 접근들이 있으며(이번송 등, 2002; 이인화 · 문영기, 2007; 이문숙 등, 2011), 공간적으로 근접한 아파트 가격의 유사한 특성(공간적 의존성)을 명시적으로 고려한 공간한 공간 헤도닉 접근의 연구들이 있다(김성우 · 정건섭, 2010; 김소연 · 김영호, 2013; 이창로 · 박기호, 2013). 공간 헤도닉 접근의 많은 연구들은 여러 실증 사례들을 통해 전통적인 회귀 모델보다 공간 회귀 모델들이 부동산 가격 추정에 있어 보다 정확함을 밝히고 있다. 하지만 이러한 공간 회귀 모델들은 부동산 가격과 설명 변수들의 전역적 상관성에 초점을 맞추고 있어 지역적으로 상이한 부동산의 하부 시장이나 가격결정요인들의 국지적 특성을 반영하는 데에는 한계를 가진다.

부동산 시장의 공간적 이질성을 반영하기 위한 공간 헤도닉 연구에서 자주 사용되는 모델은 지리가중 회귀 모델(GWR)로 각 하위 지역에 따라 상이한 가격 모델을 산출함으로써 국지적으로 이질적인 부동산 가격을 보다 정밀하게 예측할 수 있는 도구로 폭넓게 활용되어 왔다. 국내 아파트 시장의 경우, 강창덕(2010)은 서울시 아파트 실거래가 자료를 GWR을 이용하여 아파트 가격에 대한 설명 변수들의 지역적 차이들을 밝히고, 부동산 감정평가에 있어 공간적 이질성에 대한 고려가 필요함을 주장하고 있다. 또한 김혜영 · 전철민(2012)은 지리적 접근성을 고려한 부동산 가격 추정에서 GWR 모델이 전통적인 최소제곱법(OLS) 기반의 헤도닉 모델에 비해 탁월함을 밝히고 있다.

하지만 GWR 모델 역시 부동산 가격의 지역적 이질성을 효과적으로 다루는 데에는 한계가 있다. 예를 들어, GWR은 부동산 시장의 이질적인 하부시장 구조를 명시적으로 파악하기 어렵다. 왜냐하면 지역별로 상이한 하부시장이 생성되는 원인이 특정 가격형성 요인의 영향력 차이에서 기인할 수도 있지만 가격을 결정하는 요인의 구성 자체가 다를 수도 있기 때문이다(Tu *et al.*, 2007). 또한 일반적으로 국지적 추정에 포함되는 관측치의 수가 적어 추정의 신뢰성 문제를 야기하는 다중공선성(multicollinearity) 문제를 발생시킬 수도 있다(Brunsdon *et al.*, 2012).

이러한 GWR의 한계를 극복할 수 있는 대안으로 최근 지리가중라소 모델(GWL)이 주목받고 있다. GWL은 공간 현상의 공간적 이질성을 고려한 GWR에 다변량 변수 선택에 탁월한 LASSO 알고리즘을 결합한 모델이다. LASSO는 GWR 모델에서 다중공선성 문제가 해결된 최적의 하위 변수 집합을 자동적으로 선택할 수 있도록 하며, 설명 변수에 비해 관측치가 적을 수 있는 한계를 극복할 수 있게 한다. 일반적으로 부동산 가격 추정 모델들은 토지 이용, 주택 구조 변수, 근린 변수, 지역 변수, 하위 시장 구조, 시기적 특성 등 다차원의 수많은 가격결정변수들을 다루기 때문에 변수 간의 다중공선성을 체계적으로 검토하기 어렵고 적절한 변수의 선정이 쉽지 않다. LASSO는 최적의 하위 변수 집합을 선택하는데 도움을 줌으로써 이러한 다차원의 빅데이터를 분석하는데 유용한 것으로 알려져 있다. 따라서 GWR과 LASSO의 특성을 결합한 GWL은 부동산 가격 및 가격결정요인의 공간적 이질성을 명시적으로 반영하면서, 대용량의 다차원 특성을 가지는 부동산 데이터를 보다 과학적이고 체계적으로 분석할 수 있다.

본 연구의 목적은 국지적으로 이질적인 부동산 가격결정요인들을 탐색하기 위해 GWL 모델의 적용가능성을 살펴보고자 한다. 특히, 우리나라 부동산의 바로미터라 할 수 있는 서울시 아파트 시장에 초점을 맞추어 지역적으로 이질적인 아파트 가격결정요인을 도출하고자 한다. 이를 위해 서울시 아파트 단지의 평균 가격을 바탕으로 먼저 전통적인 헤도닉 모델(OLS)을 통해 전역적인 가격 추정 모델의 정확도를 살펴보고, 공간적 이질성을 고려한 국지적 모델들(GWR, GWL)을 비교·분석한다. 본 연구의 주요 분석 모델인 GWL은 대용량의 다차원인 부동산 데이터를 보다 효율적으로 분석할 수 있도록 하여, 기존의 OLS나 GWR 모델에 비해 보다 정확하고 효과적인 부동산 가격 예측 도구로 활용될 수 있을 것으로 기대된다.

2. 방법론 리뷰

앞서 간략히 언급한 것처럼, 지리가중라소 모델은 LASSO와 GWR의 장점을 결합한 형태라 할 수 있다. 따라서 본 장에서는 LASSO와 GWR의 특성을 보다 심층적으로 살펴봄으로써 지리가중라소 모델의 구조와 작동 원리를 고찰하고자 한다.

1) LASSO

LASSO(Least Absolute Shrinkage and Selection Operator)는 축소추정법이라고도 불리는 일종의 벌점화 방법(penalized method)이다. 벌점화 방법은 관측값 n이 변수의 개수 p보다 작을 경우($n<p$) 또는 벡터 간 관계가 스칼라 곱으로 정의되는 경우에 발생하는 특이행렬(singular matrix) 문제를[1] 해결하기 위한 방법으로 모델에 임의의 편의(bias)를 부과함으로써 추정되는 계수값($\hat{\beta}$)을 축소시키는 것이다. 이러한 특이행렬 문제는 상관분석, 차원축소법 등 여러 방법들을 통해 해결할 수 있으나 임의적 기준 설정과 해석상의 난해함 등의 한계를 가지고 있다. 이에 대한 대안으로 OLS의 불편추정(unbiased estimation) 가정을 완화시킨 능형회귀 모델(ridge regression)(Hoerl and Kennard, 1970)이 있다. 능형회귀 모델은 일정 수준의 편의를 부과함으로써 총 오차가 OLS에 비해 작아진다. 즉 편의와 분산 간의 상쇄 관계(trade-off)를 이용하여 추정된 모수가 비록 실제 모수와 차이를 보인다고 할지라도(큰 편의를 갖더라도), 분산의 폭을 줄여(작은 분산을 갖게 하여) 예측력을 높이게 된다. 따라서 능형회귀 모델

은 회귀계수를 추정하는 과정에서 비가역 행렬 X^TX에 일정한 값(편의)을 대각행렬에 더하여 가역행렬을 만든다. 이를 OLS의 회귀계수 추정식에 표현하면 아래 수식 (1)과 같다.

$$\hat{\beta}=(X^TX+\lambda I)^{-1}X^Ty \quad (1)$$

이때 λ는 수식 (3)의 라그랑지안 승수로서 제약조건식을 갖는 수식 (2)의 최적화 모델을 단일 식으로 변형하기 위해 사용된 조정 계수이다. 수식 (2)의 목적함수는 해당식을 최소로 만드는 회귀계수값을 구하는 것으로서 OLS처럼 오차제곱합을 최소화한다. 그러나 0보다 큰 임의의 수 t값이 제약으로 주어지게 될 경우 오차제곱합은 커지게 되며 이에 따라 회귀계수 값이 OLS에 비해 작아지게 된다. 이때의 t값은 교차검정을 통해 선택하게 된다(박창이 등, 2011).

$$\hat{\beta}^{ridge}=\text{argmin}_{\beta}\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_jx_{ij})^2 \quad (2)$$

Subject to

$$\sum_{j=1}^{p}\beta_j^2\le t^2,\ t\ge 0$$

$$\hat{\beta}^{ridge}=\text{argmin}_{\beta}\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_jx_{ij})^2+\lambda\sum_{j=1}^{p}\beta_j^2 \quad (3)$$

그러나 능형회귀 모델은 직교하지 않는 두 벡터(변수)가 존재할 때, 역행렬을 구하는 법을 제시한 것으로 일정한 수준으로 변수들의 회귀 계수를 줄여주지만 정확히 어떤 변수가 선택되어야 모델의 성능이 개선될 수 있는지에 대한 해답을 제시하지 못한다. 이에 대한 대안으로 제시된 것이 LASSO(Tibshirani, 1996)라 할 수 있다. LASSO는 능형회귀 모델과 매우 유사한데, 제약조건에 해당하는 라그랑지안 승수의 항이 절대값으로 정의되는 차이가 있다(수식 4).

$$\hat{\beta}^{lasso}=\text{argmin}_{\beta}\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_jx_{ij})^2+\lambda\sum_{j=1}^{p}|\beta_j| \quad (4)$$

그림 1은 이변량을 가정했을 때 능형회귀 모델과 LASSO가 갖는 제약식의 차이를 기하학적으로 설명한 것으로 두 모델 간의 차이를 직관적으로 이해할 수

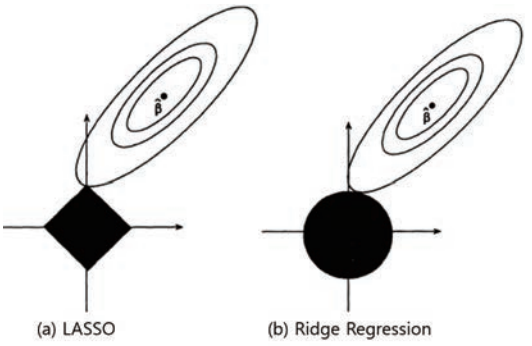


그림 1. LASSO와 능형회귀 모델의 기하학적 비교
출처: Tibshirani(1996)

있다. 음영의 영역은 제약식의 영역을 의미하고, 타원들은 $\hat{\beta}$의 오차제곱합이 같은 값을 연결한 등치선을 의미한다. 능형회귀 모델의 경우(그림 1b), 제약식이 제곱의 형태로 되어 있어 원점을 중심으로 하는 원 모양의 범위를 가지는 반면, LASSO의 경우(그림 1a) 제약식이 절대값으로 되어 정사각형의 범위를 가진다. 이 때, 추정된 $\hat{\beta}$의 오차제곱합은 LASSO의 제약식과는 y절편, 즉 한 변수의 값이 0인 지점에서 형성될 수 있는 반면, 능형회귀 모델의 경우 접점이 절편에서 형성될 수 없다. 이에 따라 LASSO 모델에서는 설명력이 없는 변수를 0으로 추정함으로써 모델에서 자동으로 탈락시킬 수 있게 된다(박창이 등, 2011).

이러한 LASSO 모델은 일반적으로 고차원인 데이터에서 예측력이 다른 방법에 비해 좋은 편으로 알려져 있고, 사례의 개수보다 변수의 개수가 많은 경우에도 활용할 수 있는 장점이 있다(박창이 등, 2011). 따라서 유전 공학에서 소수의 사례인 DNA를 대상으로 이를 구성하는 요인을 추출하는 방법에 유용하게 활용되거나(Usai *et al.*, 2009), 다변량의 빅데이터 분석에 효과적으로 활용될 수 있다(석경하 · 이태우, 2013).

2) 지리가중회귀 모델

GWR(Geographically Weighted Regression)(Brundsdon *et al.*, 1996)은 한 관측치 단위에서 가장 적합한 모델을 추정하는 국지회귀 모델의 일종으로 보

다 높은 예측력을 갖는 것을 목적으로 한다(Fotheringham, *et al.*, 2002). GWR의 일반적인 특징은 전형적인 OLS 방법에 거리 가중치를 부과하여 국지적 공간 자기상관 정도를 암묵적으로 반영하고, 전역적 공간에서의 공간적 이질성을 가정하고 있다는 점이다. 즉, 수식 (5)에서 볼 수 있는 것처럼 위치 좌표 (u_i, v_i)를 갖는 지점 i에서 각기 다른 회귀식과 회귀계수(β_k)가 를 도출함으로써 공간적으로 설명 변수의 영향력이 다를 수 있다는 현실적 요소를 반영한다.

$$y_i=\beta_0(u_i,v_i)+\sum_k \beta_k(u_i,v_i)x_{ik}+\varepsilon_i \quad (5)$$

수식 (5)에서 회귀계수의 추정은 최소오차제곱을 갖도록 한다는 점에서 OLS와 거의 유사하나, 별도로 구해진 공간가중행렬(spatial weight matrix) W를 변수에 곱해준다는 점에서 차이가 있다. 이때 공간가중행렬 W는 각 지점과 다른 지점(데이터)간의 연관성(인접성 또는 거리)에 따라 가중치가 부여된 행렬로, 이를 설명변수 X에 곱해줌으로써 국지적으로 이질적인 회귀계수를 도출할 수 있다(수식 6).

$$\hat{\beta}^{gwr}=(X^T W_i X)^{-1} X^T W_i Y \quad (6)$$

GWR의 가중행렬은 일정한 폭(bandwidth)을 설정하여 그 이내에 포함된 데이터들과의 관계만 고려하고, 그 이상의 범위에 속하는 데이터들은 회귀계수 추정에 고려하지 않는다. 이러한 근린 폭은 연구자가 임의적 또는 선험적으로 생각하는 고정된 근린의 범위로 설정될 수도 있으며(fixed kernel), 통계적으로 조정된 범위로 설정될 수도 있다(adaptive kernel). 공간적으로 불균등하게 분포하는 데이터들을 효과적으로 포섭하기 위해 일정 수 이상의 사례가 포함되도록 하는 조정 커널(adaptive kernel)의 경우, 모델의 적합도(fit)와 복잡성(parsimony) 간의 상쇄 효과를 고려하여 근린 폭을 설정하는데(수식 7) 추가되는 관측치에 따른 AIC(Akaike Information Criteria)의 비율 변화로 정의된다(Fotheringham *et al.*, 2002).

$$w_i=\frac{\exp(-AIC_i/2)}{\sum_j \exp(-AIC_i/2)} \quad (7)$$

부동산 분야에서 GWR은 부동산 가격의 공간적 이질성을 모델링하는데 활발하게 활용되고 있다. Bitter *et al.*(2007)은 미국 애리조나 투산 지역의 부동산 가격을 GWR 모델로 추정하였고, Löchl and Axhausen (2010)은 스위스 취리히의 주거용 임대료를 추정함에 있어 OLS 모델에 비해 GWR 모델의 결과가 보다 정확함을 밝혔다. 국내에서는 강창덕(2010)이 서울시 아파트 가격에 대한 GWR 분석을 통해 11개의 결정변수가 지역에 따라 다른 영향력을 가짐을 보여주고 있으며, 김혜영 · 전철민(2012)의 경우 도로 네트워크에 기반한 지리적 가중치를 사용하여 그리드 단위의 지가를 추정하였다. 또한 오윤경 등(2014)은 부산 지역 공동주택의 매매가와 전세가 추정에 있어서 GWR 모델을 사용하여 보다 정밀한 추정 결과를 도출하였다. 최근에는 GWR을 이용한 부동산 헤도닉 모델은 단순히 가격을 추정하는데 그치지 않고, 지역별로 상이한 회귀계수를 도출하는 GWR의 특징을 활용하여 부동산 하위시장 구획의 기준을 설정하는데 활용되기도 하였다(Manganelli *et al.*, 2014; 이창로 등, 2014).

그러나 앞서 언급했다시피, GWR은 근린 폭 내에 일정 정도의 사례수가 포함되어야 하는데, 적은 수의 사례로 회귀 계수를 추정하게 될 경우 실제보다 과대 추정되기 때문에 추가적으로 발생되는 사례에 의해 모델이 크게 변하거나, 모델의 예측력이 낮아지기 때문에 모델이 불안정할 수 있다[2]. 또한 변수 간 다중공선성이 클 경우에도 중복 설명으로 인한 과대 추정 문제가 발생할 수 있는데, 이는 변수들의 직관적인 이해를 어렵게 하거나 유의하지 않은 모델을 도출할 수 있기 때문이다.(Bitter *et al.*, 2007). 이러한 문제들에 대해 Holt and Lo(2008)는 앞서 언급한 차원축소법을 활용하여 변수의 개수를 조정하였고, 이창로 등(2014)은 GWR의 결과가 유의하지 않은 지역에 대해 OLS를 함께 사용하는 혼합 GWR 방식을 적용하였다. 하지만 이와 같은 방법들 역시 변수간의 공선성 문제를 사전에 직간접적으로 연구자가 검토해야하기 때문에 매우 고차원의 데이터셋인 경우 여전히 한계

가 존재한다.

3) 지리가중라소 모델

앞서 살펴본 것처럼 GWR과 LASSO는 공간 데이터를 다루는데 한계점을 갖는다. GWR의 경우 자료의 공간적 자기상관을 반영하고 공간적 이질성을 모델링하는데 적합하나, 고차원의 다변량 속성 변수들인 경우 선정한 설명 변수들이 모든 모델에서 유의하지 않을 수 있다는 한계를 지니고 있다(Yu *et al.*, 2007). LASSO의 경우 이러한 문제를 해결해주지만 공간적 효과가 반영되지 않는다는 한계가 존재한다. GWR과 LASSO를 결합한 GWL(Geographically Weighted LASSO)(Wheeler, 2009)은 GWR에서 발생하는 공선성 문제를 벌점화 방식을 통해 직접적으로 해결한 모델이다. 수식 (8)은 GWL에 의한 회귀계수의 추정식으로 공간가중행렬 W가 변수에 곱해짐으로써 공간적 의존성을 반영함과 동시에, 연산상의 문제를 발생시킬 수 있는 다중공선성을 LASSO 모델에서 도출된 조정 계수(λ)를 대각행렬에 더함으로써 해결하였다. 이 모델은 근본적으로 국지적 방법으로 관측치별로 회귀계수가 도출됨에 따라 공간적 이질성을 효과적으로 반영할 수 있다.

$$\hat{\beta}_i^{gwl}=(X^T W_i X+\lambda I)^{-1} X^T W_i Y,\ \lambda \geq 0 \tag{8}$$

GWL을 통해 추정되는 회귀계수는 특정 변수가 해당 지역에서 유의하지 않거나, 다른 변수들로 인해 영향력이 미미한 경우 0으로 수렴하여 제거된다. 즉, 공선성을 띄는 변수가 다른 추가적인 방법이 아닌 모델 자체적으로 제거되는 효과가 있다.

GWL 모델 역시 국지회귀 모델의 일종으로 일정한 범위 설정이 중요한 이슈가 되지만, GWR과 달리 일정한 수준의 사례수를 반드시 확보할 필요는 없다. 다만 모델의 예측력을 높일 수 있는 폭(ϕ)을 설정하는 것이 중요한데, 주로 교차검증을 통한 추정치($\hat{y}$)의 평균예측오차제곱근(root mean square predict error)을 최소화하는 값을 이용하여 수식 (9)를 통해 가중치를 계산한다(Wheeler, 2009). 이를 통해 도출된 거리가중행렬은 각각 종속 변수와 독립 변수에 곱해진 후, LASSO 모델(수식 4)에 대입하여 연산함으로써 회귀계수를 추정할 수 있다.

$$w_i=\sqrt{\exp(-\frac{d_{ij}}{\phi})} \tag{9}$$

한편, GWL 외에도 여러 연구에서 GWR에서 다중공선성 문제를 제기하고 이에 대한 해결방안을 제시하였다. Zhang and Mei(2011)는 GWR의 최소제곱법에 의한 해법의 한계를 극복하기 위해 최소절대값(least absolute deviation)을 이용한 방안을 제안하였으며, Bárcena *et al.*(2014)는 베이지안 방법을 이용하여 다중공선성을 해결하고자 하였다. Brunsdon *et al.*(2012)은 모델에서 공선성이 예측력과 추정 계수의 신뢰도를 약하게 하므로 모델의 해석을 어렵게 한다고 지적하면서, 일정한 임계치를 설정하여 다중공선성을 보이는 지역의 근린 폭을 증가시키거나, 능형회귀 모델을 적용하여 문제를 해결하고자 하였다. 그럼에도 불구하고 GWL은 다중공선성 문제뿐만 아니라 공간적으로 이질적일 수 있는 최적 하위변수 집합을 자동적으로 선택할 수 있다는 점에서 다른 방법들에 비해 부동산 헤도닉 모델에 매우 유용한 것으로 기대할 수 있다. 왜냐하면 부동산 가격의 지역적 차이는 정의된 독립 변수들의 영향력 차이뿐만 아니라, 서로 다른 독립 변수들의 집합에 기인할 수 있기 때문이다. 그러나 GWR의 벌점화 방법에 대한 연구는 아직 미진한 편이며, 대부분 변수들의 다중공선성에만 초점을 맞추고 있어 한계가 있다. 따라서 본 연구에서는 다중공선성 문제뿐 아니라 LASSO의 장점인 최적 변수 집합 선택 능력에 주목하여 고차원의 부동산 데이터 분석에서 GWL의 적용가능성 경험적으로 살펴볼 것이다.

3. 사례 연구

1) 사례 지역 및 데이터

본 연구의 사례 분석에서는 2013년 서울시 아파트 단지의 평균 가격을 일반적 헤도닉 모델과 국지적 공간 헤도닉 모델을 통해 추정하고 이를 비교하였다. 종속 변수로 활용된 아파트 평균 가격은 포털 사이트(네이버, 다음)에서 제공하는 서울시 아파트의 거래 매물별 시세 가격(만원/㎡)을 아파트 단지로 합역하여 평균을 산출한 뒤, 중심점으로 재현하였다.

그림 2는 2013년 서울시 아파트 가격의 연속적인 공간 분포를 살펴보기 위해 개별 아파트 단지 가격을 역거리가중(inverse distance weighting) 인터폴레이션을 통해 연속면으로 변환한 지도이다. 일반적으로 알려진 바와 같이 강남 3구로 불리는 강남·서초·송파구에서 전체적으로 높은 값을 보이고 있다. 특히 강남구 개포동의 경우 ㎡당 평균가격이 1,500만원(평당 약 5,000만원)을 상회하는 값을 가지며, 압구정동 일대의 한강 조망권역에서 역시 ㎡당 1,200에서 1,500만원 사이에서 아파트가 거래되었음을 알 수 있다. 강남구의 평균 단가는 약 736만원으로, 전체 평균인 435만원에 비해 약 1.7배 가량 높은 값을 보인다. 용산구와 서초구가 그 뒤(각각 666만원/㎡, 644만원/㎡)를 잇고 있으며, 송파구 역시 높은 값(507만원/㎡)을 보인다. 반면 서울 외곽지역이라고 인식되는 금천구와 도봉구는 낮은 거래가(각각 284만원/㎡, 302만원/㎡)를 형성하고 있음을 시각적으로도 확인할 수 있다.

한편, 아파트 가격의 영향을 줄 수 있는 설명 변수들은 크게 세 개의 범주로 수집하였다. 우선 구조 변수(structure variable)로 아파트의 건축적 특징들을 근거로 산출하였으며, 아파트 건물의 평균 층수, 가구

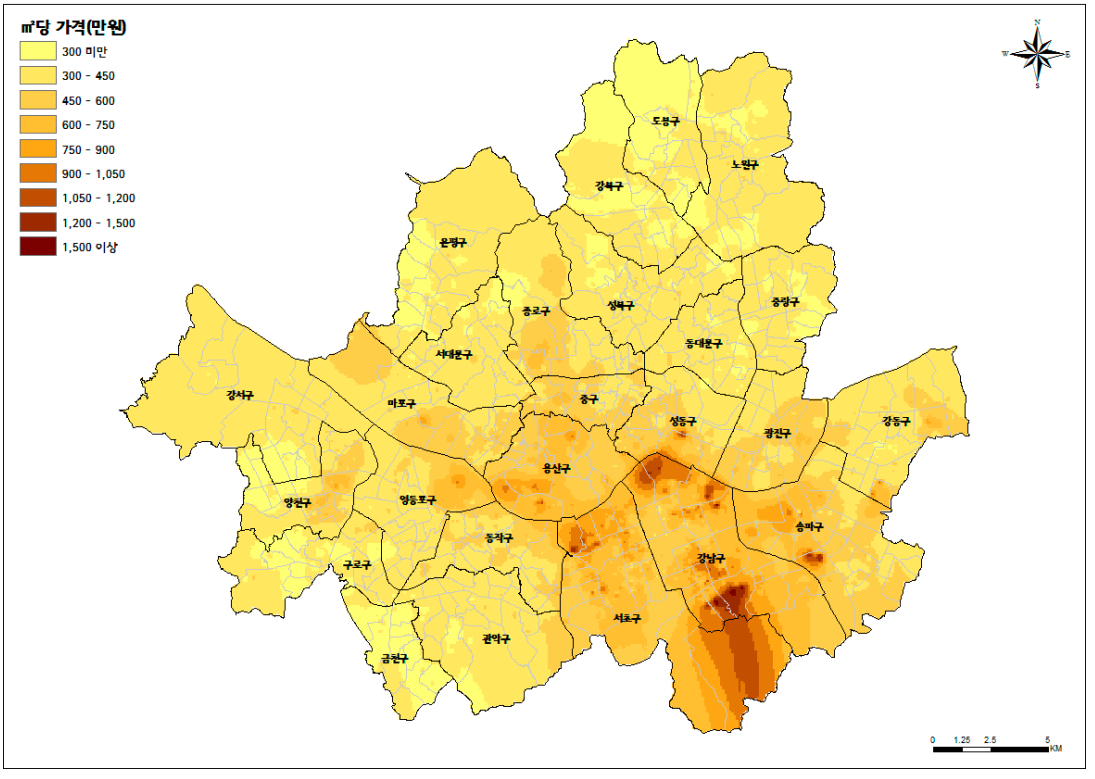


그림 2. 2013년 아파트 단지 평균 가격 분포

당 평균면적, 용적률, 건폐율, 가구당 주차대수, 난방방식, 건설업체의 인지도의 총 8개의 변수를 사용하였다. 두 번째 범주인 입지 변수(locational variable)는 아파트의 입지적 특징을 반영하는 것으로 주변의 각종 서비스 시설물과의 거리, 서비스 특성 등을 근거로 교육, 교통, 상업, 환경, 보건 부문에서 총 19개의 변수를 사용하였다(진영남 · 손재영, 2005; 정수연, 2006; 오홍운 · 김태호, 2009; 김경민 등, 2010; 박나예 · 이상경, 2013; 이진순 등, 2013). 마지막 근린 변수(neighborhood variable)는 해당 아파트가 속한 지역의 속성 또는 하부 주택 시장의 특성을 의미하는 것으로, 통계청의 2010년 인구총조사 자료와 2012년 사업체 통계를 바탕으로 동 단위로 합역하여 기본 인구학적 특성(인구밀도, 노령화인구비율), 소득수준(학력, 주거 소유 유형, 아파트가 차지하는 비율), 지역의 발전 정도(80년대 단독주택 비율, 도로 포장면적) 등을 활용하였다. 이상 39개 독립 변수에 대한 세부 설명은 아래 표 1과 같다.

표 1. 서울시 아파트 단지 가격 모델 변수

범주	변수명	변수 설명	단위
종속 변수	가격	㎡당 아파트 거래가격	만원/㎡
구조 변수	건축연한	건축 연차(2014년 기준)	년
	평균층수	평균층 수	층
	평균면적	가구 당 평균 면적	㎡
	용적률	용적률	%
	건폐율	건폐율	%
	주차	가구 당 주차대수	대
	난방방식	난방방식	1=개별난방 / 0=나머지
	브랜드	건설업체	1=10대 건설사 / 0=나머지
입지 변수	초교거리	최근린 초교와의 거리	m
	초교종류	최근린 초교의 종류	1=사립/0=공립
	중교거리	최근린 중교와의 거리	m
	중교진학	최근린 중교의 특목고 진학률	%
	고교거리	최근린 고교와의 거리	m
	고교종류	최근린 고교의 종류	1=특목고 또는 자사고 / 0=일반고
	고교학군	고교 학군	1=8학군 / 0=나머지
	지하철	최근린 지하철역과의 거리	m
	버스정류장	도보5분 거리(350m) 버스정류장 개수	개
	쇼핑	최근린 백화점(대형마트)과의 거리	m
	녹지	최근린 녹지공간과의 거리	m
	도심	최근린 고용중심지와의 거리	m
	혐오시설	최근린 혐오시설과의 거리	m
	간선도로	최근린 간선도로와의 거리	m
	도로인접	도로 인접 여부	1=인접 / 0=아님
	어린이집	도보5분 거리(350m) 어린이집의 수	개
	학원	도보5분 거리(350m) 학원의 수	개
	병의원	도보5분 거리(350m) 병 · 의원의 수	개
	종합병원	최근린 종합병원과의 거리	m

근린 변수 (동 기준)	인구밀도	인구 밀도	인/㎡
	노인비	노령화인구 비율	%
	고졸비	고졸자 비율	%
	자가비	자가 소유 비율	%
	아파트비	아파트 비율	%
	단독주택비	80년대에 건설된 단독주택 비율	%
	도로면적비	도로면적 비율	%
	도로연장비	도로연장 비율	%
	복지시설	인구 만 명당 의료복지 시설 수	‰
	문화시설	인구 만 명당 문화시설 시설 수	‰
	의료복지종사자	의료복지산업 종사자 비율	%
	문화종사자	문화산업 종사자 비율	%

주: 혐오시설(변전소, 철도기지, 대규모공단, 하수처리장)

2) 분석 방법

앞서 언급한 종속 변수와 독립 변수를 토대로 본 연구에서는 세 가지 유형(OLS, GWR, GWL)의 서울시 아파트 가격의 헤도닉 모델을 구축하였다. 국지적 모델인 GWR과 GWL의 근린 폭은 두 모델 모두 교차검정 오차(Cross Validate Error)를 최소화하는 지수함수 커널에 의한 범위로 설정하여 추정하였다. 이는 특정한 지역에 공간적 의존성이 일정한 거리에 의해 감소하는 것을 합리적으로 가정할 수 있고, 두 모델 근린 폭을 일치시킴으로써 결과 비교 시 근린 설정에 의한 오차를 막을 수 있기 때문이다. 서울시 아파트 가격 모델의 통계적 유의성과 정확성을 평가하기 위해 일반적으로 널리 사용되는 수정 결정계수(adjusted R^2), 평균절대오차(Mean Absolute Error: MAE), 평균제곱근오차(Root Mean Square Error: RMSE)를 사용하였다. MAE는 i지점에서 추정치와 관측치 차이의 절대값을 산출하여 그 평균을 구하는 비교적 단순한 평가 지표로 비교 대상의 사례수가 동일하여 관측치 개수에 의한 영향이 없을 경우 단순하고 직관적인 해석을 제공하여 유용하게 사용될 수 있다(Hydman and Koehler, 2006). 한편, RMSE는 i 지점에서 추정치와 관측치 차이를 제곱한 평균의 제곱근 값으로서 모델을 예측력을 평가함에 있어 데이터의 스케일에 무관하게 사용될 수 있다는 점에서 많이 활용되는 방법이다(Wheeler, 2009). MAE의 경우 오차의 기준으로 중앙값을 사용하기 때문에 오차의 평균값을 포함하는 RMSE에 비해 극값에 따른 영향을 덜 받는 경향이 있다.

한편, 국지적 모델에서 설명력은 각 사례별로 도출되므로, 3,210개의 수정 결정계수 값이 도출될 수 있다. 이러한 국지적 결정계수 값은 전역적 모델과의 직관적인 비교를 어렵게 하여 OLS에 비해 국지적 모델이 개선되었는지 여부를 판단하는데 어려움을 가져올 수 있다. 따라서 본 연구에서는 GWR과 GWL과 같은 국지적 모델의 설명력은 모든 국지회귀식의 평균값으로 대체하였다. 또한 예측력의 지표인 MAE와 RMSE의 경우 절대적인 기준이 존재하지 않아, 상대적인 평가를 실시하였다. 모델 간 MAE와 RMSE를 비교함으로써 상대적으로 예측력이 우수한 모델을 선정하였고, 이들 간의 차이가 유의한 수준인지 판단하기 위한 통계 분석으로 분산 분석(ANOVA)과 3개의 모델 쌍(OLS-GWR, OLS-GWL, GWR-GWL)에 대한 T-검정을 실시하였다.

모델의 다중공선성 평가는 각 변수별 분산팽창인자(Variance Inflation Factor: VIF)를 이용하였다. 일반적으로 10 이상의 값을 가질 때, 다중공선성이 문제가 될 수 있으므로 본 연구에서 VIF의 임계치는 10으로 설정하였다. 국지적 모델에서 공선성 검증은 국지적 VIF(Wheeler, 2007), 국지적 계수 상관, 전역

적 계수 상관(Wheeler and Tiefelsdorf, 2005) 등을 통해 가능하지만 본 연구에서는 전역적 계수 상관만을 기준으로 국지적 모델의 공선성을 검증하였다. 왜냐하면 데이터의 크기가 커서 국지적 계수를 도출할 경우 계수의 수가 매우 많아지기 때문이다. 총 741개의 변수 쌍에 대한 전체 계수 상관성을 계산하여 GWR과 GWL의 공선성을 검증하였으며, 상관계수의 임계치, ±0.7을 기준으로 살펴보았다. 이러한 분석을 위한 데이터의 가공 및 편집은 ArcGIS 10.1을 활용하였으며, 통계프로그램 SPSS 21과 R 3.1.1을 이용하여 모델을 구축하였다. 특히, GWR과 GWL 모델은 R 패키지 중 *gwrr*(Wheeler, 2013)을 이용하여 구축하였다.

4. 연구 결과

본 장에서는 OLS, GWR, GWL 모델간의 성능을 설명력, 다중공선성, 예측력의 측면에서 비교 평가하고자 한다. 우선 설명력을 살펴보면, OLS의 경우 95% 신뢰수준에서 유의한 27개의 변수로 약 67%의 설명력을 도출하였다. 채택된 설명 변수는 구조 변수 중 아파트 단지의 평균 층수와 가구당 면적, 용적률, 난방방식, 건설사의 인지도 등이 가장 유의한 수준에서 선택되었고, 입지 변수 중에서는 인접 고등학교의 학군, 지하철 · 쇼핑시설 · 녹지공간과의 거리, 도로인접성 여부, 근린 학원 · 병의원의 수가 가장 유의한 것으로 나타났다. 또한 근린 변수들은 노령화인구, 대졸자 비율, 자가소유 비율, 80년대에 건축된 단독주택 비율, 만 명당 문화시설 수, 의료복지산업 종사자 비율이 가장 유의한 변수로 도출되었다(표 2).

반면 국지적 모델인 GWR과 GWL은 3,210개의 모델이 각기 다른 설명력을 도출함에 따라 평균적인 설명력을 중심으로 살펴보고자 한다. 먼저 GWR은 모든 아파트 단지에 대해 39개 변수가 모두 포함되었으나 각기 다른 회귀계수를 갖는 모델로 약 92%의 매우 높은 설명력을 보여주고 있다. GWL의 경우 각 지역별로 다른 변수를 포함하는 모델로 약 91%의 현상을 설명하였다. GWL 모델의 설명 변수는 평균적으로 19개가 포함되지만 변수별로 모델에 포함되는 개수가 다르며(표 3) 변수의 종류 역시 상이하게 나타난다. 예를 들어, 매우 작은 수준의 추정오차값을 갖는 강서구 내발산동의 C아파트(−0.016)의 경우 20개의 변수를 설명 변수로 채택하였지만, 동대문구 장안평동의 J아파트는 11개의 변수만 선택하였다. 또한 노원구 공릉동의 D아파트(−1.116)와 동작구 신대방동의 S아파트(−1.163)의 경우 똑같이 변수 5개를 설명 변수 집합으로 선택하였지만, 전자의 경우 평균층수 · 가구당 면적 · 고용중심지와의 거리 · 노령인구비율 · 대졸자비율을 채택한 반면 후자의 경우 가구당 면적 · 노령인구비율 · 대졸자비율 · 자가소유비율 · 80년대 단독주택 비율을 설명 변수 집합으로 선정하였다. 이렇게 지역별로 다른 변수를 채택함에 따라 특정 변수가 설명하는 지역이 공간적으로 상이하게 나타난다. 그림 3은 변수별로 설명되는 지역이 다르게 표시되는 것을 보여주는 지도로 주황색 부분이 해당 변수에 의한 설명이 가능한 지역임을 의미한다. 가장 많은 모델에 포함된 변수인 고졸비(그림 3a)는 서울 전역에서 설명 변수로 작용하지만, 특히 서초구와 강남구 일대(검은색 지역)에서 유의하지 않은 것을 확인할 수 있다. 대신 서초구와 강남구의 경우 평균층수(그림 3b), 평균면적(그림 3c), 브랜드(그림 3d)와 같은 구조변수가 고졸비, 단독주택비(그림 3e)와 같은 근린변수에 비해 유의하게 나타나며 입지변수(간선도로, 그림 3f)는 혼합적인 양상으로 나타난다.

다음으로 다중공선성을 살펴보면 OLS의 경우 VIF 검정을 통해 분석한 결과 선택된 변수들의 VIF값이 임계치로 설정한 10미만으로 나타나 공선성이 없는 것으로 판단된다(표 2). 또한 GWL 모델 역시 다중공선성이 발생하지 않은 것으로 볼 수 있다. 이는 GWL 모델의 회귀계수 간 상관계수의 상위 5개 쌍을 포함하는 행렬인 표 4에서 임계치인 절대값 0.7을 넘는 값이 발견되지 않음을 통해 확인할 수 있다(최고값: −0.698). 반면 GWR 모델의 다중공선성 검정 결과 3개의 변수쌍이 임계치를 넘어 앞서 가장 높은 설명력을 보였던 GWR의 결과를 완전히 신뢰하기 어렵다는 것을 보여주고 있다(표 5).

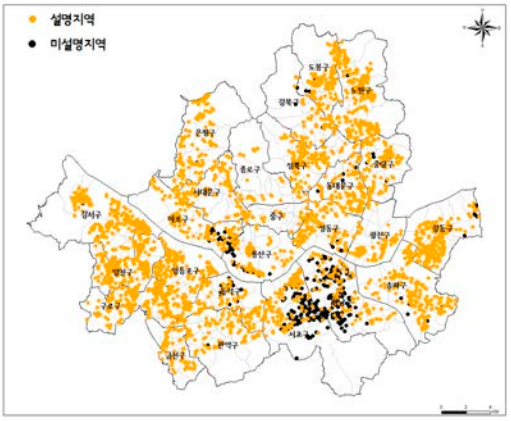
표 2. 서울시 아파트 가격 모델 결과(OLS)

변수명	Estimate	Std. Error	t value	P value	VIF
(상수항)	368.700	34.500	10.687	0.000 ***	
건축연한	−0.671	0.289	−2.321	0.020 *	1.478
평균층수	4.783	0.495	9.655	0.000 ***	1.968
평균면적	0.524	0.072	7.278	0.000 ***	1.407
용적률	−0.148	0.013	−11.392	0.000 ***	1.842
난방방식	−30.040	5.472	−5.490	0.000 ***	1.421
브랜드	37.430	4.657	8.039	0.000 ***	1.147
중교진학	0.634	0.206	3.075	0.002 **	1.253
고교종류	−11.310	5.508	−2.053	0.040 *	1.106
고교학군	178.600	8.518	20.962	0.000 ***	2.585
지하철	−0.029	0.006	−4.962	0.000 ***	1.366
버스정류장	−0.173	0.074	−2.347	0.019 *	1.301
쇼핑	−0.013	0.003	−4.287	0.000 ***	1.455
녹지	−0.026	0.003	−7.742	0.000 ***	1.108
도심	0.002	0.000	2.250	0.025 *	1.413
혐오시설	−0.004	0.002	−2.140	0.032 *	1.798
도로인접	−18.440	5.068	−3.638	0.000 ***	1.276
어린이집	−0.920	0.385	−2.392	0.017 *	1.239
학원	0.467	0.135	3.456	0.000 ***	1.517
병의원	−0.740	0.223	−3.313	0.000 ***	1.114
종합병원	0.007	0.002	3.216	0.001 ***	1.406
인구밀도	−468.800	231.200	−2.028	0.043 *	1.468
노인비	−74.540	20.390	−3.656	0.000 ***	1.585
고졸비	508.300	31.410	16.182	0.000 ***	3.169
자가비	−242.800	21.340	−11.378	0.000 ***	1.462
단독주택비	226.500	19.730	11.482	0.000 ***	1.492
문화시설수	4959.000	1111.000	4.465	0.000 ***	2.499
의료복지종사자	−447.300	74.320	−6.019	0.000 ***	2.587

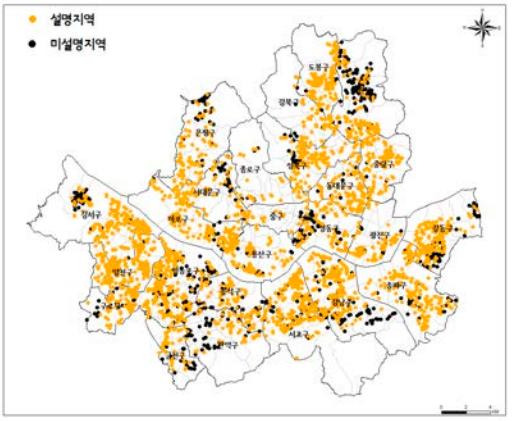
***, **, *는 각각 신뢰수준 0.001, 0.01, 0.05를 의미함

마지막으로 평균절대오차(MAE)와 평균제곱근오차(RMSE)로 평가한 예측력의 경우 평가 지표에 따라 다른 결과가 도출되었다. GWR은 RMSE가 50.459로 가장 낮은 값을 보인 반면, GWL은 MAE를 기준으로 약 25,000원 정도의 오차가 발생하여 가장 높은 예측력을 보였다. 이러한 예측력의 불일치를 통계적으로 검정한 결과 RMSE의 GWR과 GWL간의 차이가 유의하지 않은 것으로 나타났다(표 6). 즉 통계적으로 유의한 평균절대오차를 근거로 예측력을 비교했을 때, GWL이 가장 우수한 모델이라고 할 수 있다.

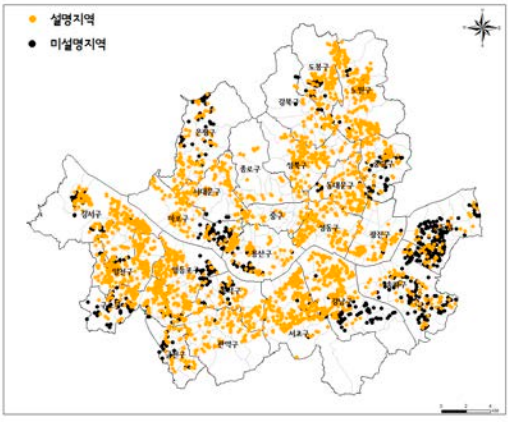
한편, 예측력의 공간적 분포는 그림 4와 5를 통해 확인할 수 있다. 그림 4는 각 모델을 이용하여 아파트 가격을 추정한 결과로, 모델별 지역적 패턴의 차이가 두드러지게 나타나고 있다. 우선 그림 4a는 OLS의 결과로 흔히 말하는 강남 3구(서초~강남~송파)가 상대적으로 높은 값을 나타내며 서울시 외곽(도봉구, 금천구)에 낮은 값을 형성하고 있다는 매우 전역적인 경향만 포착할 수 있다. 반면 GWR(그림 4b)은



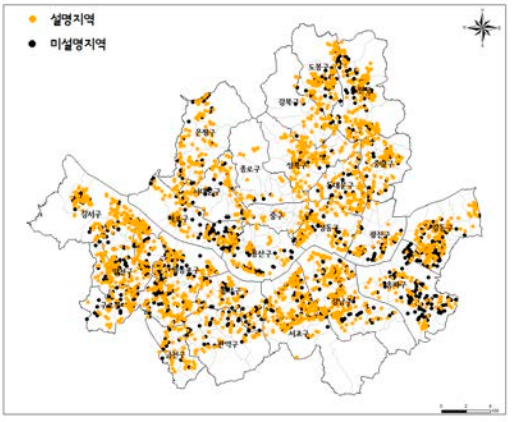
(a) 고졸비



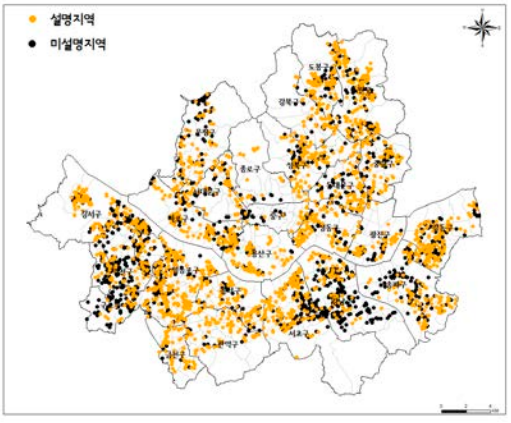
(b) 평균층수



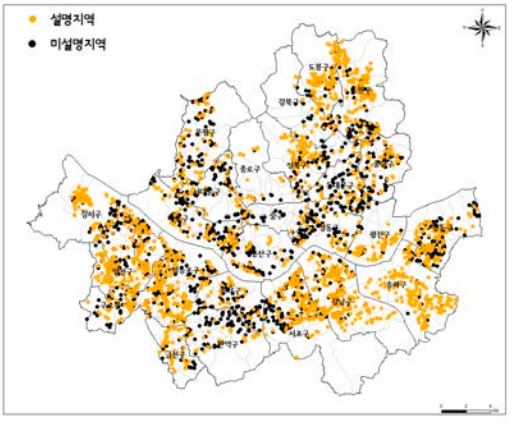
(c) 평균면적



(d) 브랜드



(e) 단독주택비



(f) 간선도로

그림 3. GWL 모델에 포함되는 상위 6개 변수의 공간적 분포

표 3. 독립 변수별 GWL 모델 개수와 포함 순위

순위	변수명	모델 수	순위	변수명	모델 수	순위	변수명	모델 수
1	고졸비	2,857	14	버스정류장	1,612	27	문화시설수	1,413
2	평균층수	2,541	15	복지시설수	1,604	28	난방방식	1,379
3	평균면적	2,530	16	초교거리	1,581	29	초교종류	1,361
4	브랜드	2,159	17	병의원	1,570	30	문화종사자	1,344
5	단독주택비	1,891	18	종합병원	1,565	31	고교거리	1,282
6	간선도로	1,880	19	고교종류	1,560	32	인구밀도	1,280
7	건축연한	1,815	20	혐오시설	1,559	33	건폐율	1,250
8	중교진학	1,708	21	의료복지종사자	1,539	34	쇼핑	1,246
9	주차	1,698	22	고용중심지	1,538	35	지하철	1,234
10	고교학군	1,686	23	중교거리	1,486	36	학원	1,229
11	도로인접	1,680	24	녹지	1,474	37	도로연장비	1,132
12	노인비	1,640	25	어린이집	1,458	38	도로면적비	1,112
13	아파트비	1,630	26	용적률	1,434	39	자가비	1,014

표 4. 회귀계수 간 상관계수 상위 5개 행렬(GWL)

	평균층수	용적률	초교종류	지하철	녹지	도심	도로인접	인구밀도	복지시설
평균층수									
용적률	−.698**								
초교종류	.212**	−.221**							
지하철	−.216**	.304**	.154**						
녹지	−.437**	.528**	.248**	.173**					
도심	.138**	.082**	.138**	.197**	−.067**				
도로인접	−.031	.160**	.187**	.508**	.042*	.275**			
인구밀도	.107**	0.006	.141**	.407**	−.089**	.580**	.399**		
복지시설	−.104**	.286**	−.542**	.065**	.287**	.033	.097**	.108**	

***, **, *는 각각 신뢰수준 0.001, 0.01, 0.05를 의미함

표 5. 회귀계수 간 상관계수 상위 5개 행렬(GWR)

	건축연한	평균층수	용적률	도심	간선도로	도로인접	인구밀도	자가비	아파트비
건축연한									
평균층수	.636**								
용적률	−.476**	−.713**							
도심	−.219**	.340**	−.219**						
간선도로	.180**	.413**	−.206**	.392**					
도로인접	.180**	.319**	−.152**	.318**	.735**				
인구밀도	.283**	.355**	−.205**	.703**	.518**	.462**			
자가비	−.243**	−.190**	.321**	.360**	−.160**	−.019	.224**		
아파트비	.063**	.362**	−.452**	.011	.308**	.063**	.155**	−617**	

***, **, *는 각각 신뢰수준 0.001, 0.01, 0.05를 의미함

표 6. 모델 간 통계적 유의성 검정 결과

지표	방법	비교 모델	F(또는 T) 값	P value
MAE	ANOVA	OLS-GWR-GWL	611.038	0.00000
	T-Test	OLS-GWR	25.703	0.00000
		OLS-GWL	28.74	0.00000
		GWR-GWL	6.007	0.00000
RMSE	ANOVA	OLS-GWR-GWL	70.1137	0.00000
	T-Test	OLS-GWR	9.036	0.00000
		OLS-GWL	8.649	0.00000
		GWR-GWL	-0.57	0.56000

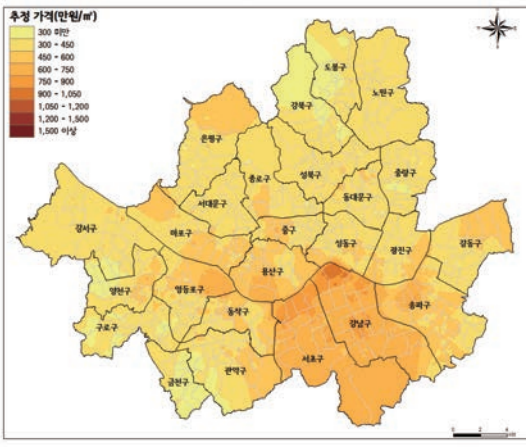
OLS 결과에 비해 지역적 편차가 두드러지며, 보다 세부적인 지역적 차이를 보여주고 있다. 아울러 OLS 모델에서 확인할 수 없었던 지역들(마포구 서교동, 영등포구 여의동, 용산구 이촌2동, 송파구 가락1동 등)이 높은 아파트 가격을 형성하고 있음을 확인할 수 있으며, 실제 아파트 가격 분포(그림 2)에서 확인할 수 있는 지역간 대소관계를 바르게 파악할 수 있다. GWL(그림 4c) 역시 실제 분포와 유사한 패턴을 포착하지만 극단적으로 높은 값들에 대한 추정은 GWR에 비해 두드러지지 않다. 다만 중위값(380.77만원/㎡)에 가까운 지역에 대한 추정은 앞서 살펴본 두 모델에 비해 정밀하게 이뤄졌음을 확인할 수 있다.

한편 그림 5의 지도는 통계적으로 유의한 차이를 보인 절대오차의 공간적 분포를 나타낸 지도로써, 직관적인 이해가 쉬운 수준(0~10, 10~50, 50~100, 100~150, 150~최대값)으로 급간을 나누고, 빨간색이 짙어질수록 큰 오차값을 나타내도록 지도화하였다. 이를 살펴보면 OLS(그림 5a)는 전체적으로 낮은 예측력을 보이는 반면, GWL(그림 5c)이 가장 높은의 예측력을 보여주고 있다. 특히 OLS에서 한강 조망권(영등포구 여의동~용산구 이촌동~성동구 성수1가동~송파구 잠실3동)을 따라 발생한 큰 오차가 GWL에서 상당히 완화되었음을 시각적으로 확인할 수 있다. 그러나 최고오차 지역은 오차가 상당히 개선된 GWR과 GWL 모델에서 모두 높고 발생 지점이 동일하다는 결과는 이 지역의 아파트 가격이 본 연구에서 설정한 변수에 의해 설명되지 않는 것을 의미하며, 이는 해당 지역에 대해 보다 심층적인 조사 또는 후속 연구가 필요함을 시사한다.

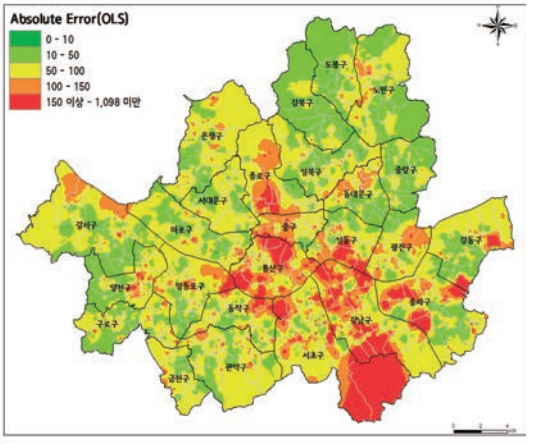
이상의 결과를 요약하면 표 7과 같다. OLS는 서울시 아파트 가격 형성 요인에 대해 다중공선성이 없고 해석이 용이한 단순한 모델을 제시하지만, 설명력이나 예측력 측면에서 그 효용성을 장담할 수 없다. 반면 국지적 모델들인 GWR과 GWL은 전역적 모델에 비해 설명력이나 예측력에서 우월하지만, GWR의 경우 다중공선성으로 인한 모델 해석이 어렵다는 한계를 가진다. 대신에 GWL은 국지적으로 설명 변수들의 영향력뿐만 아니라 설명 변수의 집합을 다르게 함으로써 GWR에 비해 보다 단순한 모델을 만들면서도 예측력의 유의한 개선을 이루어, 가장 유용한 방법으로 고려될 수 있다.

5. 결론

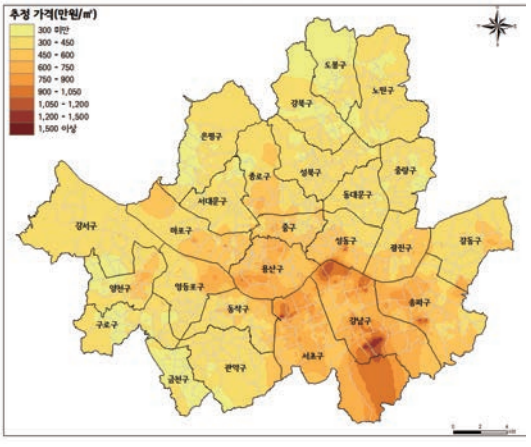
토지 및 주택 가격의 추정에 있어 많은 연구들은 다양한 가격결정요인과 하부 시장의 특성뿐 아니라 공간적 의존성 및 이질성과 같은 공간 효과들을 적절히 반영한 공간 헤도닉 접근법들이 전통적인 헤도닉 모델에 비해 보다 정확한 가격 추정에 이용될 수 있음을 보여주고 있다. 이러한 공간 효과 중 본 연구는 부동산 시장의 공간적 이질성에 주목하여 부동산 가격 추정에서 새로운 국지적 공간회귀 모델의 적용가능성



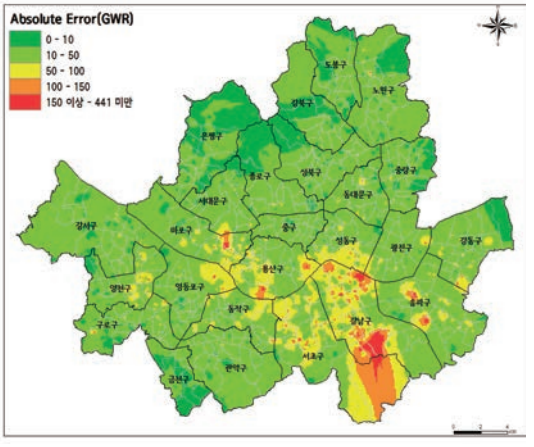
(a) OLS



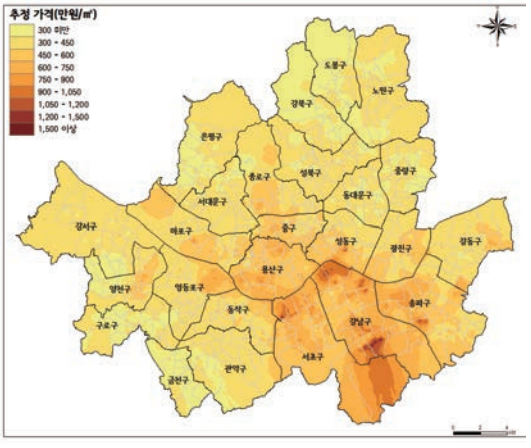
(a) OLS



(b) GWR

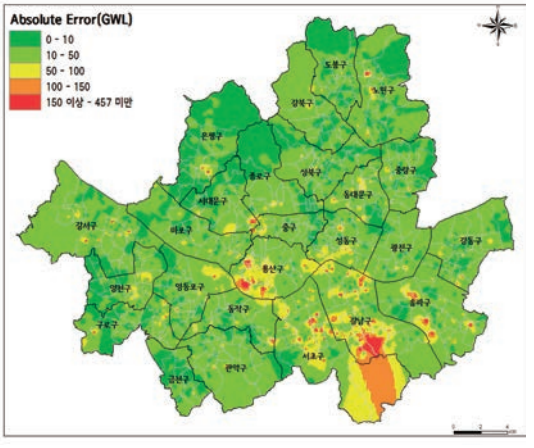


(b) GWR



(c) GWL

그림 4. 각 모델별 추정결과의 공간적 분포



(c) GWL

그림 5. 각 모델별 절대오차의 공간적 분포

표 7. 모델별 결과 비교

	변수의 개수	수정 결정계수	MAE	RMSE	다중공선성
OLS	27	0.6646	71,146	105,211	없음
GWR	39	0.921(평균)	31,729	50,459	있음
GWL	19(평균)	0.913(평균)	25,271	52,982	없음

을 살펴보았다. 구체적으로 헤도닉 모델링에서도 자주 사용되는 공간 모델인 GWR과 벌점화 방법의 일종인 LASSO를 결합한 GWL 모델을 아파트 가격 추정에 적용하였다. GWL은 부동산 가격의 공간적 이질성을 명시적으로 고려하는 한편, 고차원의 속성 데이터에서 지역적으로 다른 가격결정요인들의 최적화된 조합을 체계적으로 선정함으로써 다양한 변수를 포함하는 부동산 데이터를 보다 효율적으로 분석할 수 있는 가능성을 보여주고 있다.

GWL의 효용성을 평가하기 위해 2013년도 서울시 아파트 가격을 바탕으로 모델 변수 개수, 설명력과 예측력, 다중공선성 측면에서 GWR과 GWL 모델을 비교·분석하였으며, 아울러 전통적인 OLS 기반의 전역적 헤도닉 모델의 적합도도 함께 비교하였다. OLS 모델은 27개의 변수로 약 67%의 현상을 설명하지만, 비교적 큰 예측 오차가 나타났다. 다중공선성 문제를 탐색한 결과, 유의한 수준에서 다중공선성이 없음이 확인되었다. OLS를 통해 추정한 서울시 아파트 가격은 실제 분포와 전반적인 패턴은 유사하였지만, 추정된 가격의 범위와 지역적 범위 측면에서 큰 오차가 발생하였다. 게다가 서울시 전체적으로 아파트 가격의 편차가 작게 나타나 지역적 이질성이 충분히 반영되지 않고 있다.

반면, 국지적 모델인 GWR과 GWL의 경우에는 전역적 모델에 비해 높은 설명력과 예측 결과가 도출되었다. GWR의 경우 OLS에 비해 높은 설명력과 예측력을 보였으나, 전역적 계수 상관 분석을 통해 검증한 다중공선성에서 문제가 발견되어 모델의 결과를 충분히 신뢰하기 어려웠다. 반면 GWL의 경우 다중공선성 문제가 해결된 상태로 높은 설명력을 띠면서도 평균적으로 19개 변수로 구성되어 다른 두 모델에 비해 간명하면서 유용한 모델이 도출되었다. 이러한 국지적 모델의 결과를 공간적으로 살펴보면, 강남구와 용산구의 평균 가격을 비교적 정확하게 추정하였으며, 국지적으로 나타나는 극값들을 OLS 모델에 비해 정밀하게 추정하였다. 한편 GWR과 GWL의 결과를 비교해보면, GWL의 결과는 중위값을 중심으로 높은 예측력을 보이는 반면 극값에 대한 예측은 GWR이 보다 높은 것으로 나타났다. 그러나 전반적인 패턴을 살펴보면 GWL 모델에서 보다 현실에 부합되는 추정이 이루어졌음을 알 수 있다.

그러나 본 연구는 예측력 분석에서의 한계가 있다. 예측 평가를 위한 데이터셋을 별도로 구축하지 않고, 모델을 구축하는데 활용한 전체 데이터셋을 예측 평가에 재사용했다는 점에서 예측력이 과대평가되었을 수 있다. 그러나 본 연구에서는 절대적인 수준에서의 예측력을 다루고 있는 것이 아니라, 세 모델간의 비교를 위해 상대적인 차이를 논의하고 있기 때문에 크게 무리가 없으리라 판단된다. 다만, MAE와 RMSE의 대소 관계가 일치하지 않는다는 측면에서 예측력의 일관성 부분에 대한 문제가 제기될 수 있으며, 이는 추후에 보완된 데이터와 예측력 평가 방법과 관련한 추가적인 연구를 통해 보완될 필요가 있을 것이다. 그럼에도 불구하고 본 연구에서 제안하는 GWL은 고차원의 데이터셋에서 유의미한 독립 변수들을 효율적으로 선정하는데 직접적인 도움을 줌으로써 대용량의 복잡한 데이터 구조를 가지고 있는 부동산 데이터를 분석하기 위한 유용한 기법으로 활용될 수 있을 것으로 기대된다.

주

1) 변수로 구성된 행렬 X가 특이행렬일 경우, X^TX의 역행렬을 구할 수 없기 때문에 OLS에서 회귀계수를 추정하는 방정식의 해가 유일하지 않게 된다.
2) 조정 커널의 경우 역시 일정한 사례 수를 모델에 포함시키기 위해 밴드 폭을 확대하는데, 그 과정에서 근린 지역의 설정이 지나치게 넓어질 수 있다는 한계가 있다.

참고문헌

강창덕, 2010, "GWR 접근법을 활용한 부동산 감정평가 모델 연구," 부동산연구, 20(2), 107-132.

김경민 · 이의준 · 박대권, 2010, "초중고등학교 수요가 서울시 구별 아파트 가격에 미치는 영향," 국토연구, 65, 99-113.

김성우 · 정건섭, 2010, "부산 아파트 실거래가를 이용한 전통적 헤도닉모델과 공간계량모델간의 적합도에 관한 비교 연구," 부동산학연구, 16(3), 41-55.

김소연 · 김영호, 2013, "주거지 인문환경의 공간 속성을 고려한 주택 가격 결정 모델," 한국지도학회지, 13(3), 41-56.

김연미, 2008, "서울시 아파트가격 불균등에 관한 공간통계적 분석," 지리학논총, 52, 17-47.

김혜영 · 전철민, 2012, "공간구문론 및 지리적 가중회귀기법을 이용한 지가분석," 한국지리정보학회지, 15(2), 35-45.

박나예 · 이상경, 2013, "지역 및 근린생활환경이 주상복합아파트 가격에 미치는 영향 연구," 부동산연구, 23(2), 153-170.

박창이 · 김용대 · 김진석 · 송종우 · 최호식, 2011, R을 이용한 데이터마이닝, 교우사, 서울.

석경하 · 이태우, 2013, "데일리 렌즈 데이터를 사용한 데이터마이닝 기법 비교," 한국데이터정보과학학회지, 24(6), 1341-1348.

오윤경 · 강정규 · 김종민, 2014, "지리가중회귀모델을 이용한 주택가격 결정요인의 지역별 특성에 관한 연구," 세무회계연구, 40, 1-17.

오흥운 · 김태호, 2009, "고속도로 인터체인지 이격거리와 주변 아파트 가격의 관계 연구," 대한교통학회지, 27(6), 89-96.

이건학 · 김감영, 2013, "개별공시지가와 주택실거래가의 공간적 불일치에 관한 연구," 대한지리학회지, 48(6), 879-896.

이문숙 · 허종호 · 박승배, 2011, "아파트 브랜드가 아파트 가격에 미치는 영향에 관한 연구," 상품학연구, 29(1), 139-149.

이번송 · 정의철 · 김용현, 2002, "아파트 단지특성이 아파트 가격에 미치는 영향 분석," 국제경제연구, 8(2), 21-45.

이인화 · 문영기, 2007, "공동주택단지의 심미적 디자인요인이 아파트가격형성에 미치는 영향," 주택연구, 15(3), 169-194.

이진순 · 김종훈 · 손양훈, 2013, "환경특성이 아파트 가격에 미치는 영향에 관한 연구," 부동산연구, 23(3), 99-121.

이창로 · 박기호, 2013, "지가형성요인의 다수준 종단 분석," 대한지리학회지, 48(2), 272-287.

이창로 · 엄영섭 · 박기호, 2014, "부동산 하부시장 구획," 대한지리학회지, 49(3), 405-422.

정수연, 2006, "교육요인이 서울아파트가격에 미치는 영향에 관한 연구," 국토계획, 41(2), 153-166.

진영남 · 손재영, 2005, "교육환경이 주택가격에 미치는 효과에 관한 실증분석," 주택연구, 13(2), 125-148.

Bárcena, M. J., Menéndez, P., Palacios, M. B., and Tusell, F., 2014, Alleviating the effect of collinearity in geographically weighted regression, *Journal of Geographical Systems*, 16(4), 441-466.

Bitter C., Mulligan G. F., and Dall'erba S., 2007, Incorporating spatial variation in housing attribute prices, a comparison of geographically weighted regression and the spatial expantion method, *Journal of Geographical System*, 9(1), 7-27.

Brunsdon, C., Charlton, M., and Harris, P., 2012, Living with collinearity in local regression models, *In Proceedings of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.*

Fotheringham S., Brunsdon C., and Charlton M., 2002, *Geographically weighted regression*, Wiley, Chichester.

Hoerl, A. E., and Kennard, R. W., 1970, Ridge regression,

Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Holt, B. J., and Lo, C. P., 2008, The geography of mortality in the Atlanta metropolitan area, *Computer, Environment and Urban Systems*, 32(2), 149-164.

Hyndman, R. J., and Koehler, A. B., 2006, Another look at measures of forecast accuracy, *International journal of forecasting*, 22(4), 679-688.

Löchl, M., and Axhausen, K. W., 2010, Modelling hedonic residential rents for land use and transport simulation while considering spatial effects, *Journal of Transport and Land Use*, 3(2), 39-63.

Manganelli, B., Pontrandolfi, P., Azzato, A., and Murgante, B., 2014, Using geographically weighted regression for housing market segmentation, *International Journal of Business Intelligence and Data Mining*, 9(2), 161-177.

Rosen, S., 1974, Hedonic prices and implicit markets: Product differentiation in pure competition, *Journal of Political Economy*, 82(1), 34-55.

Tibshirani, R., 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1), 267-288.

Tu, Y., Sun, H., and Yu, S. M., 2007, Spatial autocorrelations and urban housing market segmentation. *The Journal of Real Estate Finance and Economics*, 34(3), 385-406.

Usai, M. G., Goddard, M. E., and Hayes, B. J.,2009,LASSO with cross-validation for genomic selection, *Genetics Research*, 91(6), 427-436.

Wheeler, D., and Tiefelsdorf, M., 2005, Multicollinearity and correlation among local regression coefficients in geographically weighted regression, *Journal of Geographical Systems*, 7(2), 161-187.

Wheeler, D. C., 2007, Diagnostic tools and a remedial method for collinearity in geographically weighted regression, *Environment and Planning A,* 39(10), 2464-2481.

Wheeler, D. C., 2009, Simultaneous coefficient penalization and model selection in geographically weighted regression, the geographically weighted lasso. *Environment and planning A*, 41(3), 722-742.

Wheeler, D., 2013, Package gwrr, geographically weighted regression with penalties and diagnostic tools.

Yu, D., Wei, Y. D., and Wu, C., 2007, Modeling spatial dimensions of housing prices in Milwaukee, WI, *Environment and Planning B*, 34(6), 1085-1102.

Zhang, H., and Mei, C., 2011, Local least absolute deviation estimation of spatially varying coefficient models, *International Journal of Geographical Information Science*, 25(9), 1467-1489.

교신: 이건학, 151-742, 서울특별시 관악구 관악로 1, 서울대학교 사회과학대학 지리학과(이메일: gunhlee@snu.ac.kr, 전화: 02-880-4019, 팩스: 02-876-9498)

Correspondence: Gunhak Lee, Department of Geography, College of Social Sciences, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 151-742 Korea (e-mail: gunhlee@snu.ac.kr, phone: +82-2-880-4019, fax: +82-2-876-9498)

최초투고일 2014. 11. 25
수정일 2014. 12. 19
최종접수일 2014. 12. 24