creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이 학 석 사 학 위 논 문

크리깅을 이용한 서울 4개구 아파트 단지의 매매 시세 예측

Prediction of Price for Apartment Complexes in 4 Districts of Seoul Using Kriging

2017년 2월

서울대학교 대학원

통계학과

김 석 환

크리깅을 이용한 서울 4개구 아파트 단지의 매매 시세 예측

Prediction of Price for Apartment Complexes in 4 Districts of Seoul Using Kriging

지도교수 임 채 영

이 논문을 이학석사 학위논문으로 제출함

2016년 12월

서울대학교 대학원

통계학과

김 석 환

김석환의 이학석사 학위논문을 인준함

2016년 12월

위 원 장	**이 상 열**	**(인)**
부위원장	**임 채 영**	**(인)**
위 원	**임 요 한**	**(인)**

국 문 초 록

크리깅을 이용한 서울 4개구
아파트 단지의 매매 시세 예측
Prediction of Price for Apartment Complexes
in 4 Districts of Seoul Using Kriging

다양한 데이터들이 쏟아져나오는 최근 특히 주목을 받는 데이터가 위치 정보를 담고 있는 공간 데이터이다. 공간 데이터는 공간상의 위치에 따른 의존성이 있기 때문에 그에 맞는 공간통계적 방법론을 적용하여 분석할 필요가 있다. 본 논문에서는 서울 관악구, 금천구, 동작구, 서초구의 4개구에 해당하는 아파트 단지들의 단지별 면적당 매매 시세 자료를 가지고 기본적인 공간통계적 방법론을 적용하여 분석해본다. 이를 이용하여 최종적으로 미지의 장소의 매매 시세를 예측해보는 작업, 즉 공간통계학에서 크리깅(Kriging)이라고 명명하는 작업을 수행한다. 또한 기본적인 방법론에 덧붙여 Zhang and Wang이 방대한 공간자료를 분석할 때 계산적인 측면에서 효율적이라고 제시한 predictive process model의 방법론 또한 이용한다.

주요어 : 공간통계, 크리깅, 배리오그램, 서울 아파트 매매 시세 자료
학 번 : 2015-20290

Contents

1 서론 **1**

2 데이터 설명 **3**

3 공간통계적 방법론 **5**

3.1. 배리오그램의 추정 6

3.1.1. 경험적 배리오그램을 이용한 추정 8

3.1.2. MLE를 이용한 추정 9

3.1.3. predictive process model을 이용한 추정 10

3.2. 크리깅(Kriging) 10

3.2.1. 일반 크리깅(Universal Kriging) 11

3.2.2. 단순 크리깅(Simple Kriging) 11

3.2.3. predictive process model을 이용한 크리깅 11

4 분석 및 결과 **13**

4.1. 단순 크리깅 결과 14

4.2. 일반 크리깅 결과 17

4.3. predictive process model을 이용한 크리깅 결과 22

4.4. 예측 성능 비교 24

5 맺음말 **26**

A R code **30**

List of Tables

4.1 단순 크리깅 - 각 방법별 추정된 계수들 ($\kappa = 1.5$로 고정) 17
4.2 일반 크리깅 - 각 방법별 추정된 계수들 ($\kappa = 1$로 고정) 20
4.3 predictive process model을 이용한 크리깅 - 추정된 계수들 ($\phi = 39, \kappa = 1.5$로 고정) 23
4.4 각 방법별 RMSE 비교 25

List of Figures

2.1 각 아파트 단지의 위치와 단지별 면적당 매매 시세 그림 4

3.1 $\alpha = 1$로 고정되었을 때의 매턴 모형 그래프 7
3.2 $\phi = 1$로 고정되었을 때의 매턴 모형 그래프 7

4.1 훈련자료의 아파트 단지 위치와 단지별 면적당 매매 시세 그림 . 14
4.2 단순 크리깅을 위한 배리오그램의 구름형 그림 15
4.3 단순 크리깅을 위한 경험적 배리오그램 15
4.4 단순 크리깅 - MoM 추정량과 각 방법에 따른 배리오그램 추정 16
4.5 단순 크리깅 - 로버스트 추정량과 각 방법에 따른 배리오그램 추정 16
4.6 단순 크리깅 - 시험자료의 실제 면적당 매매 시세와 예측 결과 . 18
4.7 일반 크리깅을 위한 배리오그램의 구름형 그림 19
4.8 일반 크리깅을 위한 경험적 배리오그램 19
4.9 일반 크리깅 - 로버스트 추정량과 각 방법에 따른 배리오그램 추정 20
4.10 일반 크리깅 - 시험자료의 실제 면적당 매매 시세와 예측 결과 . 21
4.11 훈련자료(파란색 빈 원)와 knot(빨간색 점) 22

4.12 predictive process model을 이용한 크리깅 - 시험자료의 실제 면적당 매매 시세와 예측 결과 24

Chapter 1

서론

다양한 데이터들이 무수히 많이 쏟아져 나오고 있는 최근 그 중에서도 특히 공간상의 위치가 포함된 공간 데이터들이 주목받고 있다. 국내 전역의 위치에 따른 강수량 데이터, 각 지역에 따른 특정 정당의 지지율 데이터 등과 같이 이차원상의 좌표에 의한 의존성이 있는 공간 데이터들은 기존의 통계학이 접근하는 방식과는 다른 방식으로 접근해야 올바른 분석이 가능하며 이에 따라 공간 자료를 다루는 공간통계학이 발전해 왔다.

국내에서도 지역과 위치에 따라 영향을 받는 지표들은 많이 있으나 그 중에서도 특히 집값은 지역에 따른 영향을 많이 받는 것으로 알려져 있다. 수도권과 지방은 물론 수도권 안에서도 서울과 비서울, 또 서울 안에서도 강남과 비강남에 따른 집값은 천차만별이다. 따라서 집값과 같은 데이터는 집이 소재하고 있는 위치에 따른 의존성이 존재하며 공간통계적 방법론을 적용해야 더 의미 있는 분석이 가능하다.

본 논문에서는 서울 안에서도 관악구, 금천구, 동작구, 서초구 4개구의 아파트 단지의 단지별 면적당 매매 시세 자료를 분석해본다. 기본적인 공간통계적 방법론을 적용하여 자료를 분석한 후 최종적으로 미지의 장소의 매매 시세를

예측해보는 과정을 수행해보고자 한다. 공간통계학에서는 이러한 예측을 크리깅(Kriging)이라고 명명한다. 또한 기본적인 공간통계적 방법론에 덧붙여 Zhang and Wang (2010)이 방대한 공간자료를 분석할 때 계산적인 측면에서 효율적이라고 제시한 predictive process model의 방법론 또한 이용하여 크리깅을 수행해보고자 한다.

본 논문의 2장에서는 분석에 사용한 아파트 매매 시세 데이터를 설명한다. 위치에 따른 매매 시세의 정도를 살펴보고 공간통계적 방법론을 적용해야할 필요성에 대하여 언급할 것이다. 3장에서는 공간통계적인 방법론에 대하여 소개한다. 공간통계 분석에서 핵심적인 역할을 수행하는 배리오그램에 대하여 소개하고 배리오그램을 추정할 수 있는 방법들을 소개한다. 또한 이를 이용하여 최종적으로 수행하고자 하는 예측과정, 즉 크리깅에 대하여서도 이야기한다. 4장에서는 이를 이용한 분석 결과에 관하여 논하고 그 의미를 고찰한다.

Chapter 2

데이터 설명

본 논문에서 분석한 데이터는 KB 국민은행이 운영하는 KB 부동산 사이트 (http://nland.kbstar.com/)에서 취득한 데이터로서 2016년 12월 2일 기준 서울특별시 관악구, 금천구, 동작구, 서초구 4개 구의 총 498개 아파트 단지의 단지별 면적당($1m^2$) 매매 시세를 담고 있다. 데이터는 아파트 단지의 공간상 좌표를 나타내는 X, Y와 단지별 면적당 매매 시세를 나타내는 info1, 단지별 총 세대수를 뜻하는 info2, 그리고 단지가 속한 행정동을 의미하는 info3와 단지가 속한 구를 의미하는 district로 구성되어 있다. 이 때 공간상 좌표를 나타내는 X, Y 변수는 구글 geocoding API를 이용하여 주소를 경도-위도 좌표(WGS-84 좌표계 혹은 EPSG 4326)로 변환한 후 이를 다시 중부원점 좌표계(EPSG 5181)로 변환하는 과정을 거쳐서 얻은 값들이다.

각 아파트 단지의 위치와 그에 따른 단지별 면적당 매매 시세를 원의 크기를 통하여 나타낸 그림을 다음 Fig 2.1에서 확인할 수 있다. 매매 시세의 평균은 591만원/m^2이며 최솟값은 187만원/m^2 (금천구 시흥동 소재), 최댓값은 2219만원/m^2 (서초구 반포동 소재)이었다. 전반적으로 서초구에 1000만원/m^2이 넘는 고가의 아파트 단지들이 분포해 있으며 동작구를 거쳐 관악구, 금천구

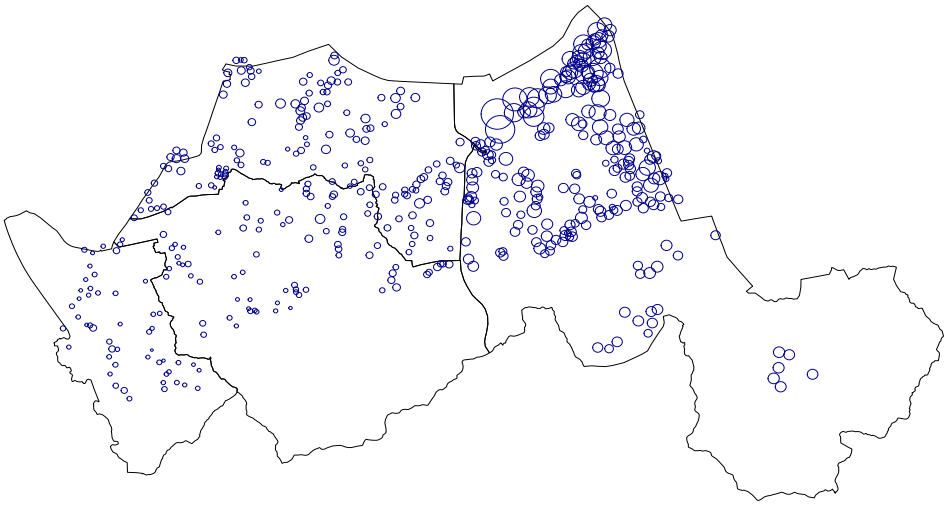

Figure 2.1: 각 아파트 단지의 위치와 단지별 면적당 매매 시세 그림

쪽으로 갈수록 매매 시세가 낮아지는 경향이 있음을 확인할 수 있다. 이러한 점을 보았을 때 매매 시세를 설명하는 데 있어서 아파트 단지의 공간상의 위치와 이에 따른 의존성을 고려한 공간통계 모형을 적합하는 것이 필요함을 알 수 있다.

Chapter 3

공간통계적 방법론

공간 자료들은 각 데이터마다 위치에 따른 의존성이 있기 때문에 이에 적절한 모형을 세워서 분석해야 한다. 주어진 데이터의 공간(위치 정보)이 연속적인 연속형 데이터에 대하여 아래와 같이 공간 확률 과정을 나타낼 수 있다.

$$Z(s),\ s \in D \subseteq \mathbb{R}^d,\ d \geq 1$$

단, 이 때 D는 관심의 대상이 되는 연속 공간이다. 확률 과정 $\{Z(s),\ s \in D\}$에 대하여 평균 함수 $\mu(s) = E[Z(s)]$ 라고 할 때, 두 지점 $s_1, s_2 \in D$에 대한 공분산 함수 $C(s_1, s_2)$는

$$C(s_1, s_2) = E[Z(s_1) - \mu(s_1)]E[Z(s_2) - \mu(s_2)]$$

로 표현할 수 있다. 일반적으로 확률 과정 $\{Z(s),\ s \in D\}$는 (이차) 정상성 조건을 만족한다고 가정한다. 즉, 임의의 $h \in D \subseteq \mathbb{R}^d$에 대하여

$$E[Z(s)] = E[Z(s+h)] = \mu$$

$$Cov(Z(s), Z(s+h)) = Cov(Z(0), Z(h)) = C(h)$$

와 같은 조건을 만족한다고 가정한다. 이는 두 지점의 데이터의 공분산은 두 지점의 절대적인 위치가 아닌 두 지점 간의 거리에 의하여 결정된다는 것을 의미한다.

정상성 조건에는 다음과 같은 고유 정상성(Intrinsic Stationary) 조건도 있다. 확률 과정 $\{Z(s),\ s \in D\}$이 임의의 $h \in D \subseteq \mathbb{R}^d$에 대하여

$$E[Z(s+h) - Z(s)] = 0$$

$$Var(Z(s+h) - Z(s)) = 2\gamma(h)$$

를 만족하면 고유 정상성 조건을 만족한다고 한다. 이 때 $\gamma(h)$는 세미배리오그램(semi-variogram)이라고 하고 $2\gamma(h)$는 배리오그램(variogram)이라고 한다. 이차 정상성 조건을 만족하는 확률 과정은 고유 정상성 조건도 만족하며 아래와 같은 식이 성립한다.

$$\gamma(h) = C(0) - C(h)$$

또한 공분산함수 $C(h)$가 아래와 같이 벡터 h의 유클리드 거리에 의해서만 영향을 받는 경우 $C(h)$를 등방성(isotropy)이라고 한다.

$$C(h) = C_0(\|h\|) \text{ or } \gamma(h) = \gamma_0(\|h\|)$$

3.1. 배리오그램의 추정

공간통계에서 주로 쓰이는 모수 등방성 배리오그램 모형 (parametric isotropy variogram model)은 여러가지가 있으나 여기에서는 그 중에서도 제일 널리 쓰이는 매턴 모형(Matern model)을 소개한다. 매턴 모형의 식은 아래와 같다.

$$C(h) = \sigma^2 \frac{2^{1-\alpha}}{\Gamma(\alpha)} \left(\frac{\|h\|}{\phi}\right)^{\alpha} K_{\alpha}\left(\frac{\|h\|}{\phi}\right), \quad \phi, \alpha > 0$$

위의 식에서 ϕ는 척도모수(scale parameter), α는 평활모수(smoothness pa-

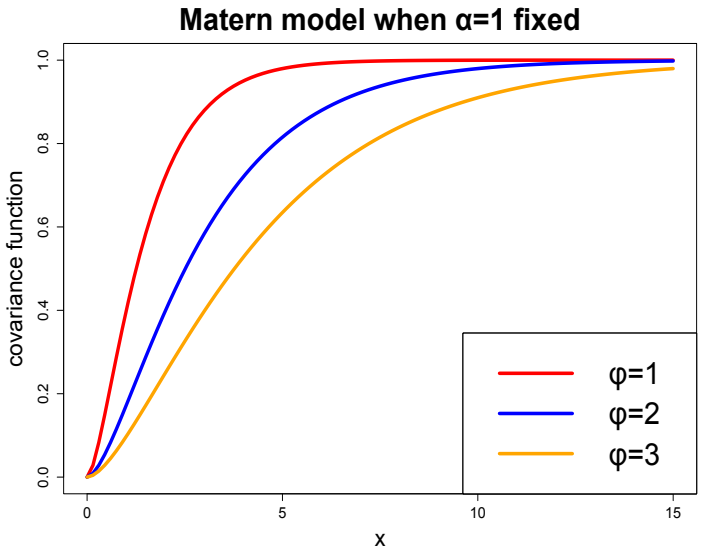

Figure 3.1: $\alpha = 1$로 고정되었을 때의 매턴 모형 그래프

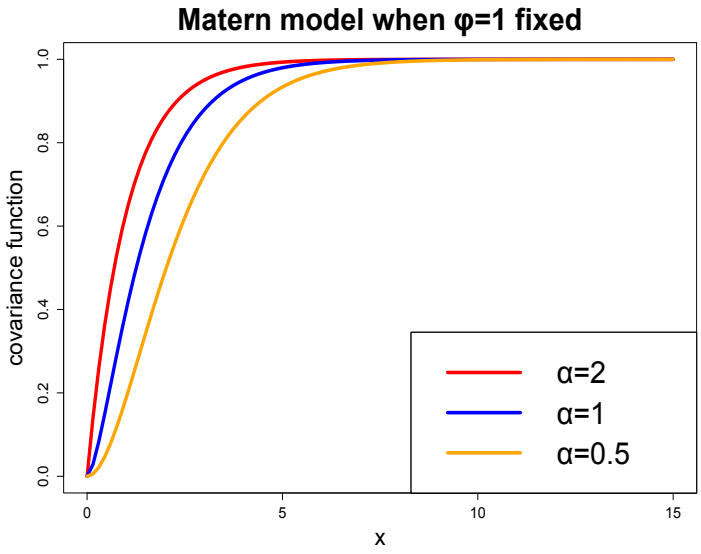

Figure 3.2: $\phi = 1$로 고정되었을 때의 매턴 모형 그래프

rameter) 이고 K_α는 차수가 α인 제2종 변형 Bessel 함수이다. Fig 3.1과 3.2에서는 ϕ와 α에 따른 다양한 매턴 모형의 그래프를 확인할 수 있다. 매턴 모형은 $\alpha = 0.5$인 경우 지수 모형이 되며, $\alpha \to \infty$인 경우 가우시안 모형이 된다.

공간 데이터를 모델링하는데 있어서 또 하나 특징적인 점으로 nugget 효과를 들 수 있다. 소규모 변동성(micro-scale variability)과 측정 오류 등의 이유로 인하여 모델로 설명되지 않는 부분을 nugget 효과라고 하는데, 공분산함수 $C(h)$가 nugget 효과 τ^2를 지니는 경우 다음과 같이 나타낼 수 있다.

$$C(h) = \sigma^2 \rho(h, \phi) + \tau^2 I_{\{h=0\}}$$

여기서 $I_{\{h=0\}}$은 $h = 0$일 때 1이고 그 외의 경우에는 0의 값을 나타내는 지시함수이다.

일반적으로 배리오그램은 비모수적 추정량인 경험적 배리오그램(emperical variogram)을 이용한 방법과 MLE를 이용한 방법으로 추정할 수 있다. 본 논문에서는 이 두 가지 방법에 덧붙여 Zhang and Wang (2010)의 논문에서 제시하고 있는 predictive process model을 이용한 방법도 소개한다.

3.1.1. 경험적 배리오그램을 이용한 추정

경험적 배리오그램은 MoM 추정량(Matheron, 1963)과 로버스트 추정량(Cressie and Hawkins, 1980)의 두 가지 경우가 많이 쓰인다.

MoM 추정량 $2\hat{\gamma}(h)$의 경우 좌표 $s_1, \cdots, s_n$ 상의 데이터에 대하여

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(s_i, s_j) \in N(h)} (Z(s_i) - Z(s_j))^2$$

와 같이 구할 수 있다. 단, $N(h) = \{(s_i, s_j) | s_i - s_j = h\}$ 이다.

로버스트 추정량 $2\bar{\gamma}(h)$의 경우는 편향 수정 계수(biased correction term)를 포함하고 있으며 식은 아래와 같다.

$$2\bar{\gamma}(h) = \frac{1}{0.457 + \frac{0.494}{|N(h)|}} \left\{ \frac{1}{|N(h)|} \sum_{(s_i,s_j)\in N(h)} |Z(s_i) - Z(s_j)|^{\frac{1}{2}} \right\}^4$$

위와 같은 방식으로 경험적 배리오그램을 구한 후 이를 모수적 모형 $\gamma(h;\theta)$에 적합하여 배리오그램을 추정할 수 있다. 여러가지 방법 중에서도 주로 쓰이는 것은 AWLS 방법으로도 불리는 Approximated Weighted Least Square 방법 (Cressie, 1985) 이다. MoM 추정량 $\hat{\gamma}(h)$를 예시로 생각할 때 모수의 추정량 $\hat{\theta}$을 구하는 AWLS 방법은 아래와 같다. 고정된 m에 대하여 $\hat{\gamma}(h_1), \cdots, \hat{\gamma}(h_m)$을 안다면

$$\hat{\theta} = argmin \sum_j \frac{|N(h_j)|}{\gamma^2(h_j;\theta)} \{\hat{\gamma}(h_j) - \gamma(h_j;\theta)\}^2$$

와 같이 $\hat{\theta}$을 구할 수 있다.

3.1.2. MLE를 이용한 추정

다음과 같은 공간 확률 과정을 생각하자.

$$Z(s) = x^T(s)\beta + \epsilon(s), \quad \epsilon(s) \sim GP(0, C(\cdot,\theta))$$

그러면 $s_1, \cdots, s_n$에 대응되는 $Z(s_1), \cdots, Z(s_n)$에 대하여 $\mathbf{Z} = (Z(s_1), \cdots, Z(s_n))^T$라고 하면

$$\mathbf{Z} = X^T\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 V(\boldsymbol{\rho}))$$

와 같이 나타낼 수 있다. 따라서 로그 가능도 함수는

$$\log l(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}; \mathbf{Z}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2 V(\boldsymbol{\rho})| - \frac{1}{2\sigma^2}(\mathbf{Z} - X\boldsymbol{\beta})^T V^{-1}(\boldsymbol{\rho})(\mathbf{Z} - X\boldsymbol{\beta})$$

가 되며 위의 로그 가능도 함수를 최대로 만드는 $(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}, \hat{\boldsymbol{\rho}})$이 공분산함수의 MLE 추정량이 된다.

3.1.3. predictive process model을 이용한 추정

Zhang and Wang (2010)의 논문에서는 low rank model의 일종인 predictive process model을 이용, drop-one 예측을 바탕으로 구한 몇 가지 예측 지표들을 최소로 만드는 방식으로 공분산함수의 모수를 추정하는 알고리즘을 소개하고 있다. $W(s)$를 평균이 0이고 공분산함수가 $\sigma^2\rho(h,\phi)$인 정상 확률 과정이라고 하자. 단 ϕ는 벡터일수도 있다. 이제 Benerjee *et al.*(2008)에서 언급한 knot이라는 이름의 m개의 위치 $\mathbf{u}_1, ..., \mathbf{u}_m$을 안다고 하면 predictive process model은 아래와 같이 정의된다.

$$Z(s) = \mu + \mathbf{k}(s)^T R^{-1}\mathbf{W} + \varepsilon(s)$$

단 $\mathbf{W} = (W(\mathbf{u}_1), ..., W(\mathbf{u}_m))^T$이고, R은 (i,j)번째 원소가 $\rho(\mathbf{u}_i - \mathbf{u}_j, \phi)$인 $\mathbf{W}$의 상관행렬이며, $\mathbf{k}(s) = Corr(\mathbf{W}, W(s))$는 j번째 원소가 $\rho(s - \mathbf{u}_j, \phi)$인 행렬이다. 이제 아래와 같은 3단계를 거쳐서 공분산함수의 모수를 추정할 수 있다.

1. 최적선형불편추정량(BLUE)으로 μ를 추정
2. MSE(Mean Square Error)를 최소화함으로써 $w = \tau^2/\sigma^2$와 ϕ를 추정
3. 로그 지표(logarithmic score)를 최소화함으로써 τ^2를 추정

3.2. 크리깅(Kriging)

배리오그램을 추정한 후에는 주어진 지점들의 데이터를 이용하여 새로운 지점에 대한 예측(prediction)을 수행할 수 있다. 공간통계학에서는 이러한 예측 과정을 크리깅(Kriging)이라고 명명한다.

3.2.1. 일반 크리깅(Universal Kriging)

다음과 같은 공간 확률 과정을 생각하자.

$$Z(s) = x^T(s)\boldsymbol{\beta} + \epsilon(s)$$

그러면 $\mathbf{Z} = (Z(s_1), \cdots, Z(s_n))^T$라는 데이터를 알고 있을 때 새로운 지점 s_0에 대하여

$$Z_0 \equiv Z(s_0) = x^T(s_0)\boldsymbol{\beta} + \epsilon(s_0)$$

를 예측하는 것이 목표가 된다. 단, $x^T(s_0)\boldsymbol{\beta} \equiv x_0^T$이고 $\epsilon(s_0) \equiv \epsilon_0$이라고 하자.

이제 $\Sigma \equiv Cov(\mathbf{Z})$, $\boldsymbol{\tau} \equiv Cov(\mathbf{Z}, Z_0)$, $\sigma_0^2 \equiv Var(Z_0)$ 와 같이 나타내자. $\hat{\boldsymbol{\beta}}_{GLS} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{Z}$ 이고 따라서 라그랑즈 승수법을 통하여 다음과 같이 $\hat{Z}_0$를 구할 수 있다.

$$\hat{Z}_0 = x_0^T\hat{\boldsymbol{\beta}}_{GLS} + \boldsymbol{\tau}^T\Sigma^{-1}(\mathbf{Z} - X^T\hat{\boldsymbol{\beta}}_{GLS})$$

이와 같은 과정을 일반 크리깅(universal Kriging)이라고 한다.

3.2.2. 단순 크리깅(Simple Kriging)

단순 크리깅(simple Kriging)은 일반 크리깅의 모형에서 $x^T(s)\boldsymbol{\beta}$ 가 μ로 바뀐 경우로서 아래와 같이 $\hat{Z}_0$를 구할 수 있다.

$$\hat{Z}_0 = \hat{\mu} + \boldsymbol{\tau}^T\Sigma^{-1}(\mathbf{Z} - \hat{\mu}\mathbf{1})$$

단, $\hat{\mu} = \bar{\mathbf{Z}}$ (표본평균)이며 $\mathbf{1}$은 성분이 모두 1인 열벡터이다.

3.2.3. predictive process model을 이용한 크리깅

3.1.3.에서 제시한 predictive process model을 Zhang and Wang (2010)의 논문에 나와 있는 low rank model의 크리깅 예측량(Kriging predictor)

을 구하는 방법에 적용하면 크리깅을 수행할 수 있다. 주어진 데이터로 $\mathbf{Z} = (Z(s_1), \cdots, Z(s_n))^T$를 알고 있으며 $n \times m$ 행렬 $K = (\mathbf{k}(s_1), ..., \mathbf{k}(s_n))^T$ 라고 나타낼 때 아래와 같이 $\hat{Z}_0$ 을 구할 수 있다.

$$\begin{aligned} \hat{Z}_0 &= \hat{\mu} + \mathbf{k}(s_0)^T R^{-1} \hat{\mathbf{W}} \\ &= \hat{\mu} + \mathbf{k}(s_0)^T (\hat{w} R + K^T K)^{-1} K^T \mathbf{Z} \end{aligned}$$

Chapter 4

분석 및 결과

2장에서 소개한 총 498개 아파트 단지의 단지별 면적당($1m^2$) 매매 시세 데이터를 사용, 매매 시세 데이터가 정상성 조건을 만족한다고 가정하고 공간통계 모형을 적합한 후 이를 이용하여 미지의 장소의 면적당 매매 시세를 예측하는 크리깅을 수행하였다. 이를 위하여 전체 데이터를 400개의 훈련자료(training data)와 98개의 시험자료(test data)로 구분하였다. 훈련자료 400개만을 이용하여 배리오그램을 추정, 모형을 적합한 후에 98개 시험자료의 위치를 대상으로 몇 가지 크리깅을 실시하고 예측값과 실제 매매 시세를 이용, RMSE(root mean square error)를 구해 각 방법 별로 예측 성능을 비교해보았다. 훈련자료의 400개 아파트 단지의 위치와 그에 따른 단지별 면적당 매매 시세를 원의 크기를 통하여 나타낸 그림을 Fig 4.1에서 확인할 수 있다.

총 세 가지 방법으로 크리깅을 실시하였다. 첫 번째는 단순 크리깅, 두 번째는 공변량으로 단지별 세대수와 단지가 속한 구를 포함시킨 일반 크리깅, 세 번째는 predictive process model을 이용한 크리깅이다. 크리깅을 실시하기 전에 계산 상의 편의를 위하여 공간 정보가 입력되어 있는 변수 X, Y를 100분의 1로 스케일링하여 분석에 사용하였다.

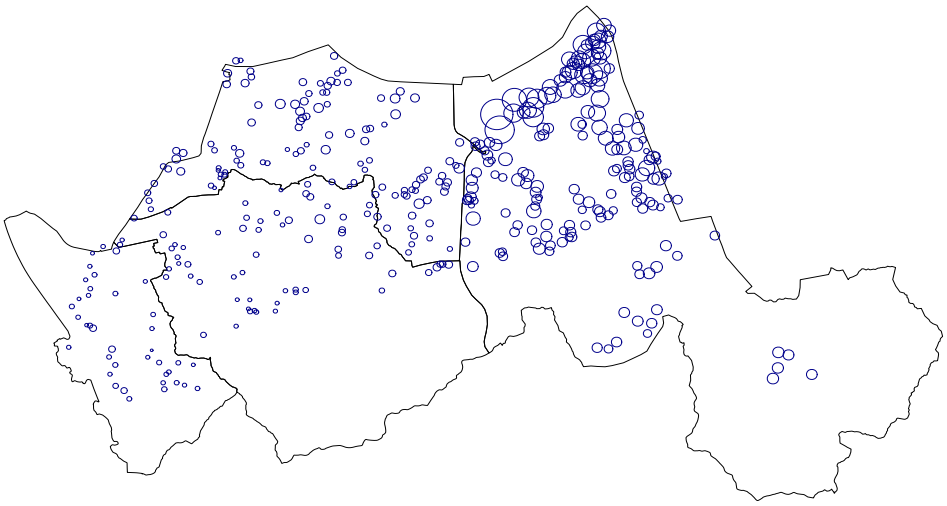

Figure 4.1: 훈련자료의 아파트 단지 위치와 단지별 면적당 매매 시세 그림

4.1. 단순 크리깅 결과

훈련자료를 대상으로 배리오그램의 경향을 살펴보기 위한 구름형 그림을 그려보면 Fig 4.2와 같다. 그림을 보면 거리가 12km 이상인 데이터 쌍의 수가 적은 것을 알 수 있다. 따라서 최대거리를 12km로 제한하여 MoM 추정량과 로버스트 추정량의 그림을 그려보면 Fig 4.3과 같다. 두 경우 모두 전체적인 경향 자체를 보았을 때 매턴 모형에 적합하는 것이 적절하다 생각되어 각 추정량에 $\kappa = 1/\alpha = 1.5$로 고정하고 AWLS 방법을 적용, 매턴 모형을 적합하여 각각 배리오그램을 추정함과 동시에 MLE를 이용하여 배리오그램을 추정하였다. 그 결과는 Fig 4.4, Fig 4.5 및 Table 4.1에 제시되어 있다.

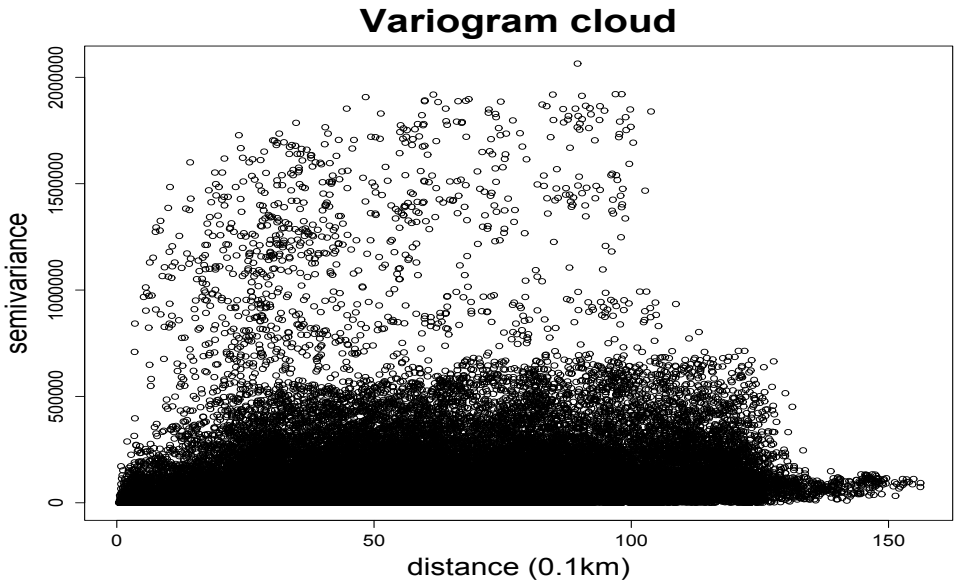

Figure 4.2: 단순 크리깅을 위한 배리오그램의 구름형 그림

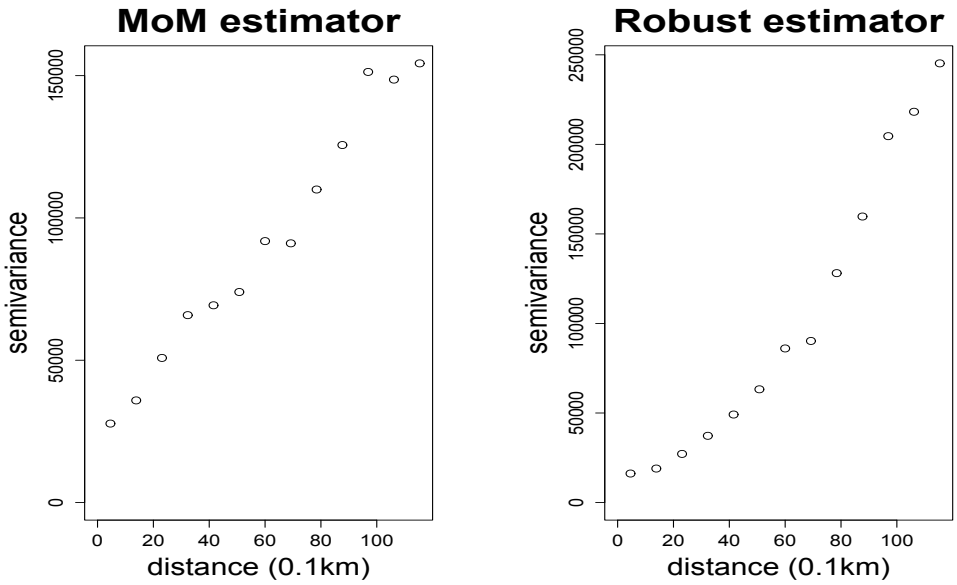

Figure 4.3: 단순 크리깅을 위한 경험적 배리오그램

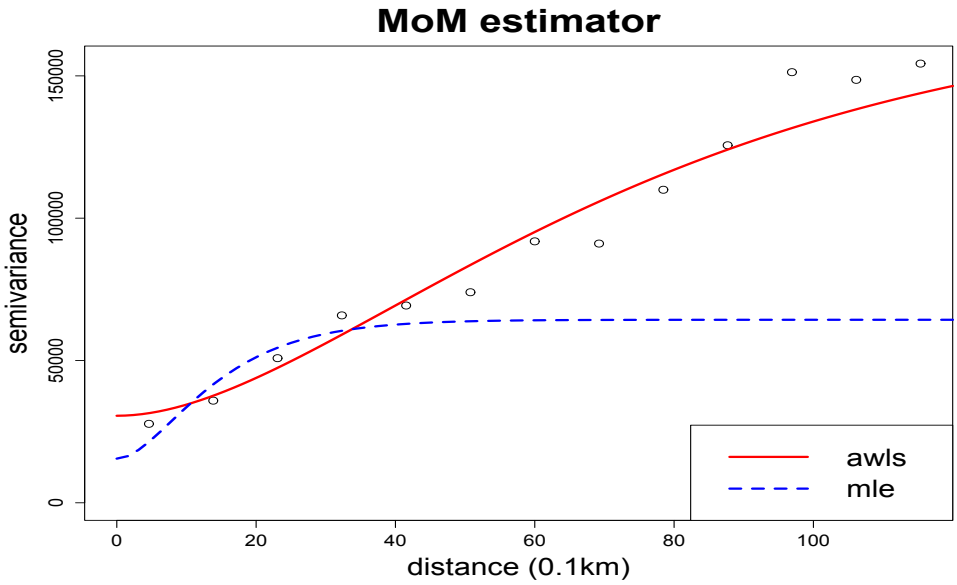

Figure 4.4: 단순 크리깅 - MoM 추정량과 각 방법에 따른 배리오그램 추정

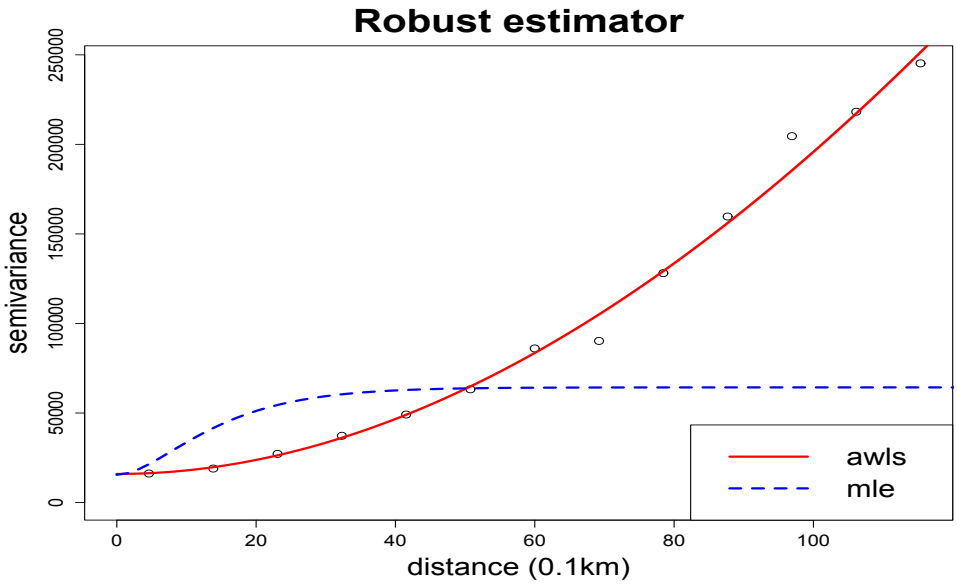

Figure 4.5: 단순 크리깅 - 로버스트 추정량과 각 방법에 따른 배리오그램 추정

Table 4.1: 단순 크리깅 - 각 방법별 추정된 계수들 ($\kappa = 1.5$로 고정)

	τ^2	σ^2	ϕ
MoM	30576.04	143439.33	39.38
Robust	15959.08	14999428.10	611.69
MLE	15514.64	48786.99	7.73

추정된 배리오그램을 토대로 하여 98개의 시험자료를 대상으로 크리깅을 실시하였다. 크리깅 결과는 다음 페이지의 Fig 4.6에서 확인할 수 있다. 파란색은 시험자료 98개 아파트 단지의 실제 면적당 매매 시세를 나타내며 빨간색은 MoM 방법, 초록색은 Robust 방법, 주황색은 MLE 방법을 이용하여 추정한 면적당 매매 시세를 나타낸다. 모든 경우에 대하여 대체적으로 추정이 실제 매매가와 비슷하게 이루어지나 서초구의 경우 실제 매매가가 비싼 단지를 다소 싼 값으로 추정하거나 반대로 실제 매매가가 높지 않은 단지를 실제 매매가보다 더 비싸게 추정하는 경우가 자주 보이는 것으로 나타난다.

4.2. 일반 크리깅 결과

훈련자료를 대상으로 단지별 세대수와 단지가 속한 구를 공변량으로 포함하여 배리오그램의 경향을 살펴보기 위한 구름형 그림을 그려보면 Fig 4.7과 같다. 단순 크리깅에서와 마찬가지로 거리가 12km 이상인 데이터 쌍의 수가 상대적으로 많지 않은 것을 알 수 있다. 따라서 최대 거리를 12km로 제약하고 단지별 세대수와 단지가 속한 구를 공변량으로 포함하여 MoM 추정량과 로버스트 추정량 그림을 그려보면 Fig 4.8과 같다. 구의 경우 4가지 범주로 이루어진 범주형 자료이므로 5가지의 가변수로서 모형에 포함되었다. 두 경우

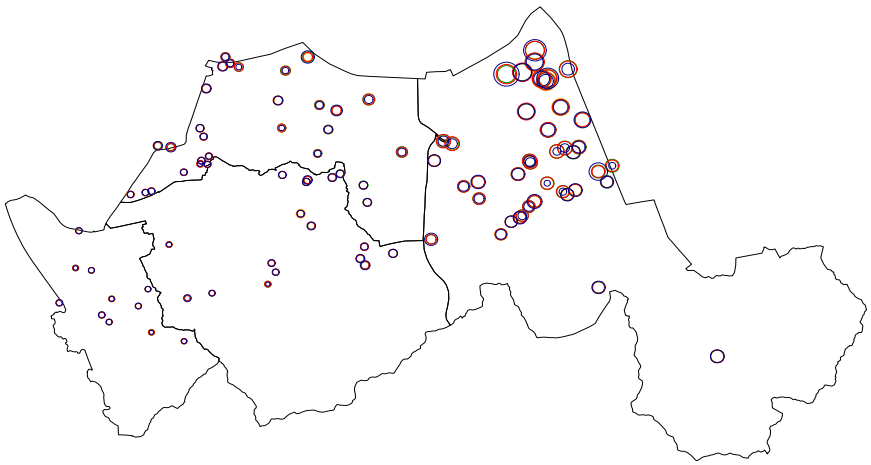

Figure 4.6: 단순 크리깅 - 시험자료의 실제 면적당 매매 시세와 예측 결과

모두 nugget 효과가 분명하게 있는 것으로 드러나는데, MoM 추정량의 경우 거리가 늘어남에 따라 값이 감소하는 특이한 경향을 보이고 있다. 이에 비해 로버스트 추정량은 감소 경향이 없는 것으로 보여 로버스트 추정량을 대상으로 $\kappa = 1/\alpha = 1$로 고정하고 ALWS 방법을 적용, 매턴 모형을 적합하여 배리오그램을 추정함과 동시에 MLE를 이용한 추정 역시 진행하였다. 그 결과는 Fig 4.9 및 Table 4.2 에 제시되어 있다.

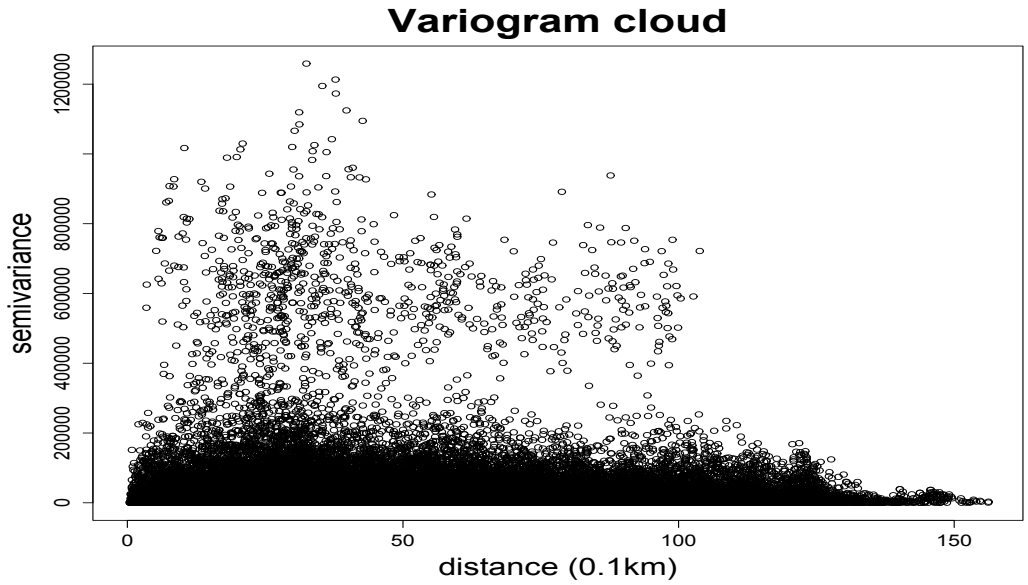

Figure 4.7: 일반 크리깅을 위한 배리오그램의 구름형 그림

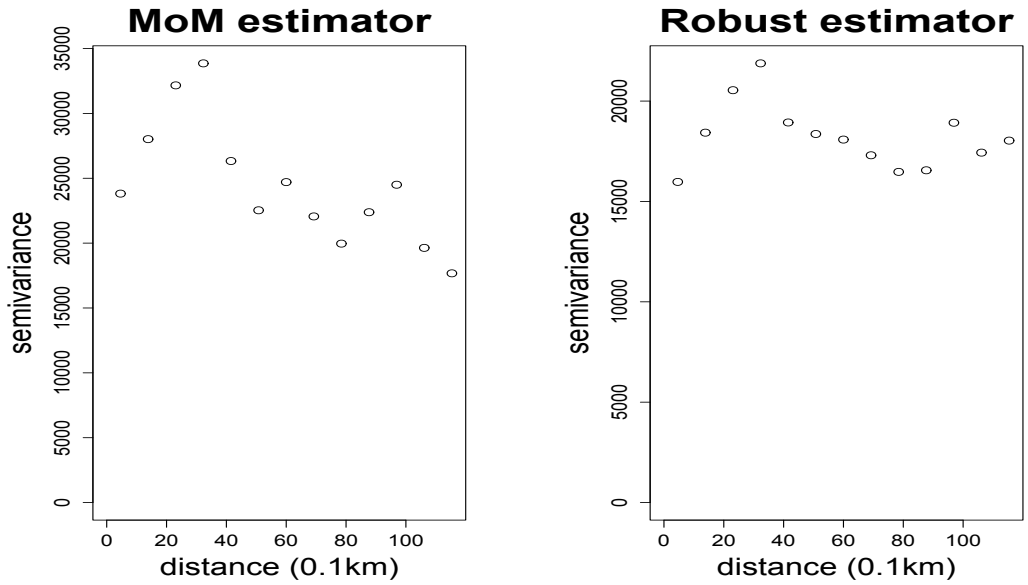

Figure 4.8: 일반 크리깅을 위한 경험적 배리오그램

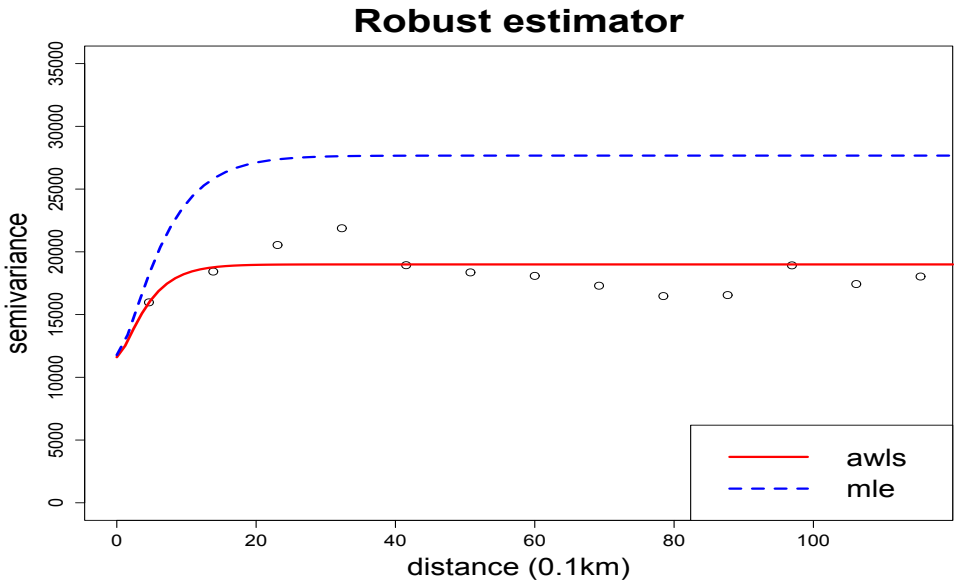

Figure 4.9: 일반 크리깅 - 로버스트 추정량과 각 방법에 따른 배리오그램 추정

Table 4.2: 일반 크리깅 - 각 방법별 추정된 계수들 ($\kappa = 1$로 고정)

	τ^2	σ^2	ϕ
Robust	11601.34	7394.28	3.05
MLE	11773.20	15894.96	4.55

추정된 배리오그램을 토대로 98개의 시험자료에 대한 크리깅을 실시하였고 Fig 4.10에서 그 결과를 확인할 수 있다. 파란색은 시험자료 98개의 실제 면적당 매매 시세를 나타내며 Robust 방법을 통한 예측 시세는 초록색, MLE 방법을 통한 예측 시세는 주황색으로 각각 나타냈다. 단순 크리깅에서와 마찬가지로 대체적으로 실제 매매가와 비슷하게 추정이 이루어지는 경향이 있는 것으로 보이나, 단순 크리깅에서도 살펴볼 수 있었듯이 주로 서초구의 아파트

단지를 중심으로 간간히 실제 매매가가 비싼 단지를 다소 싼 값으로 추정하거나 반대로 실제 매매가가 낮은 단지를 실제보다 더 비싸게 추정하는 경우가 발생하는 것을 알 수 있다.

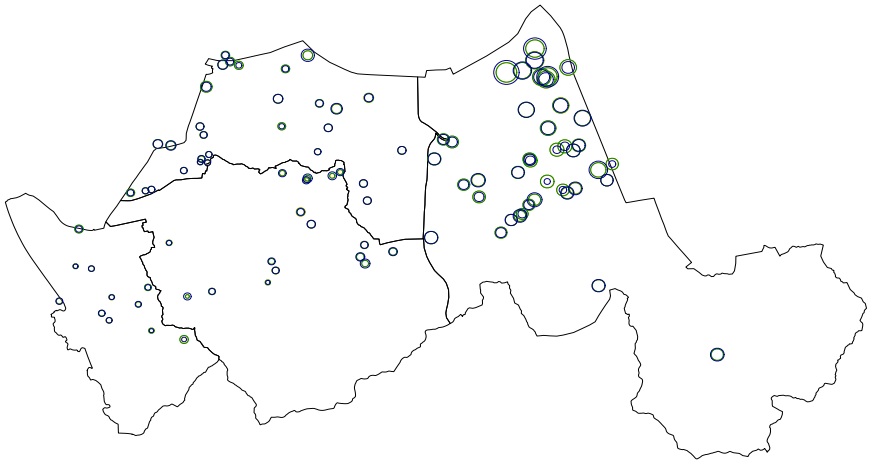

Figure 4.10: 일반 크리깅 - 시험자료의 실제 면적당 매매 시세와 예측 결과

4.3. predictive process model을 이용한 크리깅 결과

predictive process model을 이용하여 배리오그램을 추정하기 위하여 우선 51개의 knot을 설정하였다. Zhang and Wang (2010)의 논문에서의 추천에 따라 훈련자료의 위치 사이사이에 일정한 간격으로 knot을 지정하였다. Fig 4.11에서 훈련자료의 공간 위치와 knot의 배치를 확인할 수 있다. 훈련자료는 파란색 빈 원으로 표시되어 있고 51개의 knot은 빨간색 점으로 표시되어 있다.

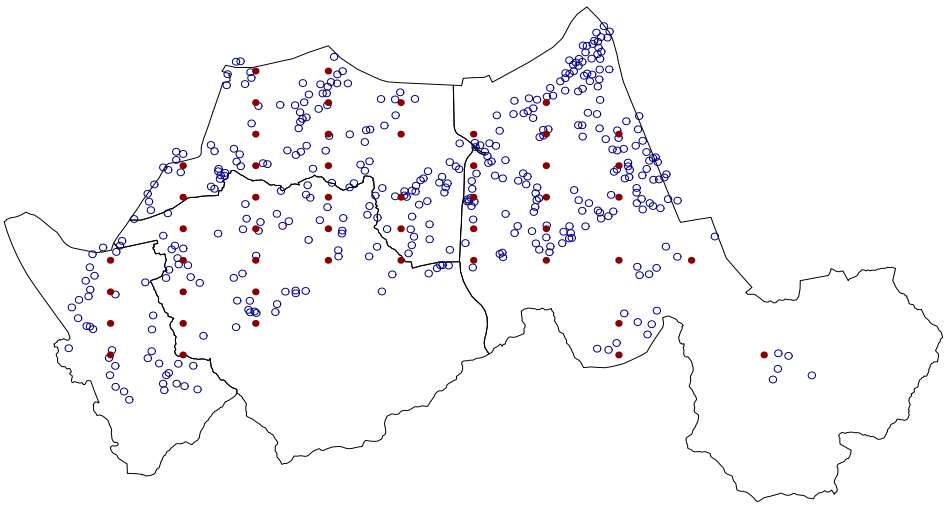

Figure 4.11: 훈련자료(파란색 빈 원)와 knot(빨간색 점)

단순 크리깅에서 MoM 추정량을 이용하여 매턴 모형에 배리오그램을 적합한 것을 토대로 $\phi = 39, \kappa = 1.5$ 이라는 값을 적합에 사용하는 것이 타당하다고 판단하였다. 따라서 $\phi = 39, \kappa = 1.5$로 고정하고 predictive process model의 다른 모수들을 추정한 결과는 Table 4.3과 같다.

Table 4.3: predictive process model을 이용한 크리깅 - 추정된 계수들 ($\phi = 39, \kappa = 1.5$로 고정)

τ^2	σ^2	μ
33763.61	230261.70	332.79

이를 토대로 98개의 시험자료에 대한 크리깅을 실시하였고 Fig 4.12에서 그 결과를 확인할 수 있다. 파란색은 시험자료 98개의 실제 면적당 매매 시세를 나타내며 빨간색은 predictive process model을 이용한 크리깅으로 예측한 시세를 나타낸다. predictive process model을 이용한 크리깅은 단순 크리깅 및 일반 크리깅의 경우와는 다르게 전반적으로 모든 지역의 시세를 실제 매매 시세보다 더 높게 예측하는 경향이 있음을 확인할 수 있다.

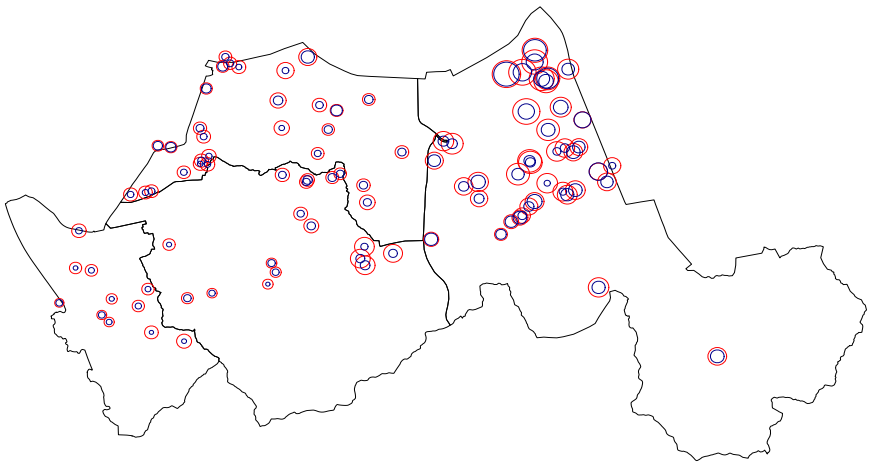

Figure 4.12: predictive process model을 이용한 크리깅 - 시험자료의 실제 면적당 매매 시세와 예측 결과

4.4. 예측 성능 비교

위에서 수행한 세 가지 크리깅의 예측 성능을 비교해보았다. 예측 성능을 비교하는 지표로는 아래에 소개되어 있는 RMSE(root mean square error)를 사용하였다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Z}(s_i) - Z(s_i)\right)^2}$$

비교 결과는 다음 Table 4.4에서 확인할 수 있다.

Table 4.4: 각 방법별 RMSE 비교

단순 MoM	단순 로버스트	단순 MLE
128.04	136.43	120.90
일반 로버스트	일반 MLE	PPM
121.69	118.84	402.13

전반적으로 단지별 세대수와 단지가 속한 구를 공변량으로 넣은 일반 크리깅의 경우가 RMSE가 작은 편으로 나타났다. 경험적 배리오그램을 사용하여 배리오그램을 추정한 경우와 MLE를 이용하여 추정한 경우 둘 다 각각 단순 크리깅에 비하여 공변량을 포함하여 예측을 한 일반 크리깅이 더 예측 능력이 높았다고 할 수 있다. 이는 지어진지 얼마 되지 않아 매매 시세가 높은 아파트 단지의 경우 대규모로 지어져 세대수가 많은 경향이 있다는 점과 더불어, 비슷한 지역에 아파트 단지가 위치하더라도 구의 경계를 기점으로 서초구는 매매 시세가 높고 동작구나 관악구는 매매 시세가 낮은 것과 같이 속해있는 구에 따른 아파트 단지의 시세 차이가 분명히 존재하기 때문인 것으로 파악된다.

반면 predictive process model을 이용한 크리깅의 경우에는 이 방법이 전반적으로 매매 시세를 실제보다 높게 예측했던 경향에서도 짐작할 수 있듯이 다른 두 가지 크리깅에 비하여 RMSE가 꽤 높게 나타나 예측 능력이 많이 떨어지는 것으로 드러났다. Zhang and Wang (2010)이 제시하였듯이 이 방법은 방대한 공간자료를 대상으로 크리깅을 수행할 때 계산상의 효율성을 높이는 것을 목적으로 고안되었다. 따라서 다른 방법들에 비하여 이 방법의 예측 능력이 떨어지는 것은 자료의 수가 그다지 많지 않은 본 케이스에 predictive process model을 사용하는 것이 다소 부적절하기 때문이라고 생각할 수 있다.

Chapter 5

맺음말

본 논문에서는 서울 4개구 아파트 단지의 면적별 매매 시세 데이터를 이용하여 공간통계적 방법론을 적용, 모형을 적합하고 크리깅을 하였다. 본 논문에서 분석에 사용한 데이터는 아파트 단지의 매매 시세로써 이는 아파트 단지가 소재하고 있는 위치에 따른 의존성이 있는 데이터이다. 따라서 일반적인 통계적 방법론이 아닌 공간통계적 방법론을 이용하여 모형을 세우고 예측을 할 필요가 있었다. 498개의 데이터를 400개의 훈련자료와 98개의 시험자료로 나누어 단순 크리깅, 세대수와 단지가 속한 구를 공변량으로 모형에 포함한 일반 크리깅, predictive process model을 이용한 크리깅을 시도한 결과 일반 크리깅의 예측 성능이 가장 우수했다. 따라서 본 논문에서는 세대수와 단지가 속한 구를 공변량으로 모형에 포함한 일반 크리깅을 이용하여 아파트 단지의 면적별 매매 시세를 예측하는 것이 좋다고 제안하는 바이다.

Bibliography

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **70(4)**, 825–848.

Bivand, R.S., Pebesma, E., Gomez-Rubio, V. (2013). *Applied Spatial Data Analysis with R.* Springer.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology*, **17**, 563–586.

Cressie, N., Hawkins, D.M. (1980). Robust estimation of the variogram: I *Mathematical Geology*, **12(2)**, 115-125.

Cressie, N., Johannesson, G. (2008). Fixed rank kriging for very large data sets. *Journal of the Royal Statistical Society*, **B 70**, 209-226.

Gelfand, A.E., Diggle, P.J., Fuentes, M., Guttorp, P. (2010). *Handbook of Spatial Statistics.* CRC Press.

Heo, S., Lee, S. (2008). Land price differentials between Gangnam and Gangbuk areas in Seoul. *Korea Real Estate Review*, **18(2)**, 125-144.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, **58**, 1246–1266.

Zhang, H., Wang, Y. (2010). Kriging and cross-validation for massive spatial data. *Environmetrics*, **21**, 209-304.

Abstract

Suk Hwan Kim
The Department of Statistics
The Graduate School
Seoul National University

With the advent of various and enormous data, there exists a great interest in spatial data which contains the information of locational coordination. Since spatial data have dependency on the spatial location where the data are observed, the data need to be analyzed using appropriate spatial statistical method. This thesis analyze the price data for apartment complexes in 4 districts in Seoul – Gwanak-gu, Geumcheon-gu, Dongjak-gu, and Seocho-gu - with elementary spatial statistical methods, which leads to the prediction of price for apartment complexes of unknown location. This process is called Kriging in spatial statistics. In addition to basic spatial statistical analysis, the data will be analyzed using the predictive process model proposed by Zhang and Wang, which can be applied to the massive data for efficient computation.

Keywords : *Spatial statistics, Kriging, Variogram, Spatial data.*
Student Number : 2015-20290

Appendix A

R code

```
##################### 자료 최종 정리 ##########################
final <- read.csv("final.csv")[,-1]

##################### 지도 위에 그림 그리기 ###################
list.files()

library(maptools)
library(PBSmapping)

Shape <- importShapefile("seoul",readDBF=TRUE)
sp.Shape<-readShapeSpatial("seoul")
plot(sp.Shape)
regions<-slot(sp.Shape,"polygons")
regions<-lapply(regions,function(x){SpatialPolygons(list(x))})

plotPolys(Shape)
```

```
finalevent <- as.EventData(final[,1:3], projection = NA)
map <- rbind(regions[[5]], regions[[8]],
regions[[6]], regions[[4]])
plot(map)
addPoints(finalevent,col="blue4",cex=final[,4]/500,pch=1)
plot(map)
addPoints(finalevent,col="red4",cex=final[,5]/500,pch=1)

##################### X, Y 축 scaling 하기 ####################
final2 <- final
final2[,2] <- final[,2]/100
final2[,3] <- final[,3]/100



##################### train set, test set 나누기 ###############
set.seed(34)
train <- sample(1:498, 400)
test <- setdiff(1:498, train)
trainset <- as.geodata(final2[train,-1], covar.col=c(4,6))
testset <- as.geodata(final2[test,-1], covar.col=c(4,6))


finalevent1 <- as.EventData(final[train,1:3], projection = NA)
finalevent2 <- as.EventData(final[test,1:3], projection = NA)


plot(map)
addPoints(finalevent1,col="blue4",cex=final[train,4]/500,pch=1)
```

```
plot(map)
addPoints(finalevent2,col="blue4",cex=final[test,4]/500,pch=1)

#### ANALYSIS 1 : Simple Kriging with no covariate #############
## variogram 기본 적합 ##
anal1.v1 =variog(trainset, option='cloud')
# MoM
anal1.v2 =variog(trainset,  max.dist=120,estimator.type='classical')
# Robust
anal1.v3 =variog(trainset,  max.dist=120,estimator.type='modulus')

par(mfrow=c(1,1), cex=1, cex.lab=1.5, cex.main=2, mai=c(1,1.2,0.7,0.7))
plot(anal1.v1, main='Variogram cloud')

par(mfrow=c(1,2), cex=1, cex.lab=1.5, cex.main=2, mai=c(1,1.2,0.7,0.7))
plot(anal1.v2, main='MoM estimator', cex=1.3)
plot(anal1.v3, main='Robust estimator', cex=1.3)


## ALWS, MLE 적합 ##
dev.off()

anal1.alws=variofit(anal1.v2, ini=c(120000, 20),
cov.model="matern", fix.kappa=T, kap=1.5,
nug=10000, weights='cressie') # ALWS (MoM)
anal1.alws2=variofit(anal1.v3, ini=c(120000, 20),
cov.model="matern", fix.kappa=T, kap=1.5,
nug=10000, weights='cressie') # ALWS (Robust)
anal1.mle=likfit(trainset, ini=c(120000, 20),
cov.model='matern', fix.kappa=T, kap=1.5,
```

```
fix.nugget=FALSE, nug=10000) # MLE



par(mfrow=c(1,1), cex=1, cex.lab=1.5,
cex.main=2, mai=c(1,1.2,0.7,0.7))
plot(anal1.v2, main="MoM estimator", col='black', lwd=1, cex=1.3)
lines(anal1.alws, lty=1, lwd=3, col="red")
lines(anal1.mle, lty=2, col='blue', lwd=3)
legend('bottomright', c('awls', 'mle'),
col=c('red', 'blue'), lwd=c(3,3),lty=c(1,2), cex=1.5)

par(mfrow=c(1,1), cex=1, cex.lab=1.5,
cex.main=2, mai=c(1,1.2,0.7,0.7))
plot(anal1.v3, main="Robust estimator", col='black', lwd=1, cex=1.3)
lines(anal1.alws2, lty=1, lwd=3, col="red")
lines(anal1.mle, lty=2, col='blue', lwd=3)
legend('bottomright', c('awls', 'mle'),
col=c('red', 'blue'), lwd=c(3,3),lty=c(1,2), cex=1.5)



## Simple Kriging ##
KC.anal1.alws=krige.control(type='SK', obj.mod=anal1.alws)
sk.anal1.alws=krige.conv(trainset,
krige=KC.anal1.alws, loc=final2[test, 2:3])

KC.anal1.alws2=krige.control(type='SK', obj.mod=anal1.alws2)
sk.anal1.alws2=krige.conv(trainset,
krige=KC.anal1.alws2, loc=final2[test, 2:3])
```

```
KC.anal1.mle=krige.control(type='SK', obj.mod=anal1.mle)
sk.anal1.mle=krige.conv(trainset,
krige=KC.anal1.mle, loc=final2[test, 2:3])

plot(map)
addPoints(finalevent2,col="orange",cex=sk.anal1.mle$pred/400,pch=1)
addPoints(finalevent2,col="green4",cex=sk.anal1.alws2$pred/400,pch=1)
addPoints(finalevent2,col="red",cex=sk.anal1.alws$pred/400,pch=1)
addPoints(finalevent2,col="blue4",cex=final[test,4]/400,pch=1)


## RMSE 측정을 위한 예측값 정리 ##
prediction <- data.frame(final[test, 4],
sk.anal1.alws$pred,
sk.anal1.alws2$pred,
sk.anal1.mle$pred)


########## ANALYSIS 2 : Univesal Kriging with covariates ########
## variogram 기본 적합 ##
anal2.v1 =variog(trainset, option='cloud',
trend=~trainset$coords[,1]+trainset$coords[,2]+
trainset$covariate[,1]+trainset$covariate[,2])
# MoM
anal2.v2 =variog(trainset, max.dist=120,estimator.type='classical',
trend=~trainset$coords[,1]+trainset$coords[,2]+
trainset$covariate[,1]+trainset$covariate[,2])
# Robust
anal2.v3 =variog(trainset, max.dist=120,estimator.type='modulus',
trend=~trainset$coords[,1]+trainset$coords[,2]+
```

```
trainset$covariate[,1]+trainset$covariate[,2])

par(mfrow=c(1,1), cex=1, cex.lab=1.5, cex.main=2, mai=c(1,1.2,0.7,0.7))
plot(anal2.v1, main='Variogram cloud')

par(mfrow=c(1,2), cex=1, cex.lab=1.5, cex.main=2, mai=c(1,1.2,0.7,0.7))
plot(anal2.v2, main='MoM estimator', cex=1.3)
plot(anal2.v3, main='Robust estimator', cex=1.3)




## ALWS, MLE 적합 ##
dev.off()

anal2.alws=variofit(anal2.v3, ini=c(5000, 10),
cov.model="matern", fix.kappa=T, kap=1,
fix.nugget=F,  nug=15000, weights='cressie') # ALWS
anal2.mle=likfit(trainset,trend=~trainset$coords[,1]+
trainset$coords[,2]+trainset$covariate[,1]+
trainset$covariate[,2], ini=c(500, 1),
cov.model='matern', fix.kappa=T, kap=1,
fix.nugget=FALSE, nug=15000) # MLE


par(mfrow=c(1,1), cex=1, cex.lab=1.5,
cex.main=2, mai=c(1,1.2,0.7,0.7))
plot(anal2.v3, main="Robust estimator",
col='black', lwd=1,  cex=1.3, ylim=c(0,35000))
```

```
lines(anal2.alws, lty=1, lwd=3, col="red")
lines(anal2.mle, lty=2, col='blue', lwd=3)
legend('bottomright', c('awls', 'mle'),
col=c('red', 'blue'),lwd=c(3,3), lty=c(1,2), cex=1.5)


## Universal Kriging ##
KC.anal2.alws=krige.control(type='OK', obj.mod=anal2.alws,
trend.d=~trainset$coords[,1]+trainset$coords[,2]+
trainset$covariate[,1]+trainset$covariate[,2],
trend.l=~testset$coords[,1]+testset$coords[,2]+
testset$covariate[,1]+testset$covariate[,2])
sk.anal2.alws=krige.conv(trainset,
krige=KC.anal2.alws, loc=final2[test, 2:3])


KC.anal2.mle=krige.control(type='OK', obj.mod=anal2.mle,
trend.d=~trainset$coords[,1]+trainset$coords[,2]+
trainset$covariate[,1]+trainset$covariate[,2],
trend.l=~testset$coords[,1]+testset$coords[,2]+
testset$covariate[,1]+testset$covariate[,2])
sk.anal2.mle=krige.conv(trainset,
krige=KC.anal2.mle, loc=final2[test, 2:3])


plot(map)
addPoints(finalevent2,col="orange",cex=sk.anal2.mle$pred/400,pch=1)
addPoints(finalevent2,col="green4",cex=sk.anal2.alws$pred/400,pch=1)
addPoints(finalevent2,col="blue4",cex=final[test,4]/400,pch=1)
```

```
## RMSE 측정을 위한 예측값 정리 ##
prediction <- data.frame(prediction,sk.anal2.alws$pred, sk.anal2.mle$pred)


############ ANALYSIS 3 : Predictive Process Model #######
# 기본 설정 #
data<-final2
a<-seq(min(final2[,2])+10,max(final2[,2])-10,length=10)
b<-seq(min(final2[,3])+10,max(final2[,3])-10,length=10)

ax <- rep(a, 10)
by <- rep(b, each=10)
knot <- data.frame(ax, by)
names(knot) <- c("Easting", "Northing")

index <- c(1,2,8,10, 11, 12, 13, 18, 21, 22, 23,
31, 32, 33, 34, 35, 36, 37, 38, 39,
42, 43, 44, 45,46, 47, 52, 53, 54,
55, 56, 57, 58, 62, 63, 64, 66, 67, 68,
73, 74, 75, 76, 77, 78, 83, 84, 85, 87, 93, 94)

knot <- knot[index,]


knotevent <- data.frame(1:51, knot[,1:2]*100)
colnames(knotevent) <- c("EID", "X", "Y")
knotevent <- as.EventData(knotevent, projection = NA)

plot(map)
```

```
addPoints(finalevent1,col="blue4",cex=1,pch=1)
addPoints(knotevent,col="red4",cex=1,pch=16)

plot(trainset$coords[,1],trainset$coords[,2],
pch='x',xlab="", ylab="",  axes=F)
points(knot$Easting,knot$Northing,pch='v',  col="red")


spatialdata<-trainset
targetdata<-testset

c1=c(1:length(spatialdata$data))
c2=c(1:dim(knot)[1])



Y=spatialdata$data


CVcheck=function(phi){

one=as.matrix(rep(1, length=length(Y)))
In=diag(rep(1, length=length(Y)))


R=matrix(0, length(c2), length(c2))
for (i in 1:length(c2)) {
for (j in 1:length(c2)) {
h=sqrt((knot[c2[i],1]-knot[c2[j],1])^2 +
(knot[c2[i],2]-knot[c2[j],2])^2)
```

```
R[i,j]=1-matern(h,phi, kappa=1.5)
}
}
Rinv=solve(R)

Kt=matrix(0, length(c2), length(Y))
for (j in 1:length(Y)) {
for(i in 1:length(c2)){
h=sqrt((knot[c2[i],1]-spatialdata$coords[c1[j],1])^2
+ (knot[c2[i],2]-spatialdata$coords[c1[j],2])^2)
Kt[i,j]=1-matern(h,phi, kappa=1.5)
}
}
K=t(Kt)

# mse
mse = function(w) {
G = K%*%solve(w*R+ Kt%*%K)%*%Kt
muhat= as.vector((t(one)%*%(In-G)%*%Y)/(t(one)%*%(In-G)%*%one))
f=sum((solve(In-diag(G))%*%(In-G)%*%(Y-muhat*one))^2)
return (f)
}


# estimation of w
what= optimize(f = mse, lower=0, upper=0.2,
maximum=FALSE, tol=0.00000000001)
#what=nlm(f=mse, p=0.1)

# estimation of mu
```

```
G = K%*%solve(what$minimum*R+ Kt%*%K)%*%Kt
#optimize -> what$minumum
bunza <- (t(one)%*%(In-G)%*%Y)
bunmo <- (t(one)%*%(In-G)%*%one)
muhat= as.vector((t(one)%*%(In-G)%*%Y)/(t(one)%*%(In-G)%*%one))

# estimation of tau square
#tausqhat=sum((sqrt(solve(In-diag(G)))%*%(In-G)%*%Y)^2)/length(Y)
alphasq=1/(1-diag(G))-(1-rowSums(G))^2/((1-diag(G))^2
*(length(Y)-sum(G)))
errorsq=(solve(In-diag(G))%*%(In-G)%*%(Y-muhat*one))^2
tausqhat=mean(errorsq*(1/alphasq))

sigmasq=tausqhat*alphasq
zsq=errorsq*(1/sigmasq)
#return(alphasq)
#return(tausqhat)
#return(c(sigmasq=sigmasq, zsq=zsq))

# prediction score
RMSE=sqrt(sum(( (solve(In-diag(G))) %*% (In-G)
%*% (Y-muhat*one))^2)/length(Y))
LogS=sum(log(2*pi*sigmasq)/2+zsq/2)/length(Y)
CRPS=mean(sqrt(sigmasq)*(sqrt(zsq)*(2*pnorm(sqrt(zsq))-1)
+2*pnorm(sqrt(zsq))-(1/pi)^(1/2)) )
return(c(phi=phi, what=what$minimum, tausqhat=tausqhat,
RMSE=RMSE, LogS=LogS, CRPS=CRPS, muhat=muhat))



} # optimize -> what$minimum
```

```
estimates <- CVcheck(39)
round(estimates,2)

# 최종 적합 #
phi=39
muhat=estimates[7]
tausq=estimates[3]
what=estimates[2]



R=matrix(0, length(c2), length(c2))
for (i in 1:length(c2)) {
for (j in 1:length(c2)) {
h=sqrt((knot[c2[i],1]-knot[c2[j],1])^2 +
(knot[c2[i],2]-knot[c2[j],2])^2)
R[i,j]=1-matern(h,phi, kappa=1.5)
}
}
Rinv=solve(R)

Kt=matrix(0, length(c2), length(Y))
for (j in 1:length(Y)) {
for(i in 1:length(c2)){
h=sqrt((knot[c2[i],1]-spatialdata$coords[c1[j],1])^2 +
(knot[c2[i],2]-spatialdata$coords[c1[j],2])^2)
Kt[i,j]=1-matern(h,phi, kappa=1.5)
}
}
K=t(Kt)
```

```
A=K%*%Rinv



ks=matrix(0, ncol=98, nrow=length(c2))
for (j in 1:98)
{
for (i in 1:length(c2))
{
h=sqrt((knot[c2[i],1]-targetdata$coords[c1[j],1])^2 +
(knot[c2[i],2]-targetdata$coords[c1[j],2])^2)
ks[i,j]=1-matern(h,phi, kappa=1.5)
}
}

muhat+t(ks)%*%Rinv%*%solve(what*Rinv + t(A)%*%A)%*%t(A)%*%Y
muhat+t(ks)%*%solve(what*R + Kt%*%K)%*%Kt%*%Y

# plot
plot(map)
addPoints(finalevent2,col="red",cex=(muhat+t(ks)
%*%Rinv%*%
solve(what*Rinv + t(A)%*%A)
%*%t(A)%*%Y)/400,pch=1)
addPoints(finalevent2,col="blue4",cex=final[test,4]/400,pch=1)


## RMSE 측정을 위한 예측값 정리 ##
prediction <- data.frame(prediction,muhat+t(ks)
```

```
%*%Rinv
%*%solve(what*Rinv + t(A)%*%A)%*%t(A)%*%Y)
colnames(prediction) <- c("TRUE", "SK_ALWS1",
"SK_ALWS2", "SK_MLE",
"UK_ALWS", "UK_MLE", "PPM")


# RMSE 구해보기 #
RMSE <- data.frame(matrix(rep(0, 6), nrow=1))
colnames(RMSE) <- c("SK_ALWS1", "SK_ALWS2",
"SK_MLE", "UK_ALWS", "UK_MLE", "PPM")
for (i in 2:7)
{
RMSE[1,i-1] <- sqrt(sum((prediction[,i]-prediction[,1])^2)/98)
}

RMSE
```