박만식교수지도

석사학위청구논문

t-분포 하에서의 절단크리깅을 이용한 공간자료의 예측연구

2015

성신여자대학교 대학원

통계학과

최빛나

t-분포 하에서의 절단크리깅을 이용한 공간자료의 예측연구

박만식교수지도

이 논문을 석사학위 논문으로 제출함

2014년 11월

성신여자대학교 대학원

통계학과

최빛나

인 준 서

최빛나의 석사학위 논문으로 인준함.

심사위원 _____인

심사위원 _____인

심사위원 _____인

성신여자대학교 대학원

논 문 개 요

공간통계학이란 미지의 지점의 값을 예측하기위해 관측된 위치자료, 관측자료로 이루어진 공간자료를 분석하는 응용통계학의 한 분야이다. 공간통계학에서 미지의 관측지점의 값을 관측값의 선형결합으로 예측하는 방법을 크리깅이라 한다. 크리깅의 여러 방법들 중 절단 가우시안 크리깅이 있다. 이 방법은 연속형으로 관측되는 공간자료를 범주로 나누어 각 범주에 속할 확률을 표준정규분포의 누적분포함수를 이용하여 예측하는 방법이다.

본 논문에서는 기존의 절단 가우시안 크리깅을 확장하여 각 범주에 속할 확률을 t-분포의 누적분포함수를 이용한 예측방법을 제안하였다. t-분포의 누적분포함수를 적용하기 위하여 공간자료의 공간상관성을 적합한 뒤 자유도를 추정하였다. 모의실험에서 자유도가 각기 다른 t-분포와 표준정규분포를 이용하여 공간자료를 생성하였다. t-분포의 자유도에 따라 본 연구에서 제안한 방법이 기존의 크리깅 방법과 어떤 차이를 보이는지 정분류율, 민감도, 특이도 관점에서 비교하였다. 또한 2012년 연평균 미세먼지(PM_{10})자료를 이용하여 기존의 크리깅과 제안한 크리깅 방법을 각각 적용하였다.

주요어 : t-분포, 공간상관성, 공분산모형, 절단 가우시안 크리깅, 절단 t-분포 크리깅, 일반크리깅, 지시크리깅

목 차

그 림 목 차

표 목 차

제 1 장

서론

공간통계학(spatial statistics)에서 사용되는 지리통계학적 공간자료(geo-statistical data)는 고정된 위치지점과 각각의 위치에서 측정된 관측자료로 이루어져 있다. 대기환경관측망에서 측정된 미세먼지 농도가 한 예이다. 지리통계학적자료를 포함한 공간자료(spatial data)는 관측값들 사이에 상관성을 갖는데 이를 공간상관성(spatial correlation)이라 한다. 공간상관성을 나타내는 대표적인 측도로는 공분산모형(covariance model)과 세미베리오그램(semi-variogram)이 있다. 미지의 지점을 예측할 때, 예측지점 근처의 관측값들의 선형결합으로 예측하는 방법을 크리깅(kriging)이라 한다. 이러한 공간상관성의 모형화를 통하여 미지의 지점에 대한 신뢰성 있는 예측(prediction)이 가능하다.

크리깅을 이용한 선행연구는 다음과 같다. 이상일 (2002)은 지하수 및 토양오염의 추정을 위하여 크리깅 기법을 적용하였다. 조홍래와 정종철 (2006)은 강우자료를 공간보간 기법을 적용하여 보통크리깅(ordinary kriging), 일반크리깅(universal kriging) 예측 결과를 비교하였다. 김선우 등 (2005)은 크리깅 모형과 지리적 가중회귀모형을 일산화탄소자료에 적합하고, 예측오차제곱합(prediction error sum of square)으로 예측성능을 비교하였다. 김동휘 등(2010)은 인천광역시의 송도의 얍밀층 지층분포추정을 위하여 단순크리깅(simple kriging), 보통크리깅, 일반크리깅을 사용하여 얍밀층의 두께를 추정하고, 각 방법의

신뢰성을 비교하여 얍밀층 두께를 가장 잘 추정하는 크리깅기법을 선택하는 연구를 하였다. 또한 Elbasiouny *et al*. (2014)은 보통크리깅 방법을 이용하여 토양 표본 탄소(carbon) 와 질소(nitrogen)자료를 이집트의 나일 삼각주의 북쪽(Northern Nile Delta, Egypt) 지역의 연속형 지도로 변환하였고, 이를 통하여 연구지역의 탄소와 질소의 변동성의 차이를 밝혔다. 그리고, 정승환 등 (2010)은 풍속자료에 크리깅 기법을 적용하여 남한지역의 풍속예측지도를 구성하였다. 최지은과 박만식 (2013)은 서로 다른 관측망에서 얻어진 자료를 계층모형을 이용하여 공간분석을 실시하였다. 허태영 등 (2007)은 연평균일교통량 예측을 위하여 공간통계학적 분석을 적용하였다. 이와 같이 공간 분석을 이용한 연구는 Gundogdu *et al*. (2007), Wang *et al*. (2009), Selby *et al*. (2013)이 있다.

기존의 크리깅 기법 외에 새로운 크리깅 방법을 제안한 선행 연구는 다음과 같다. 고혜지(2014)는 비등방성(anistropic) 공간자료를 이용하여 하나 이상의 방향을 고려하는 예측 모형에 관한 연구를 하여 미세먼지 자료를 등방성모형과 하나의 방향을 고려한 비등방성모형, 두 개 방향을 고려한 비등방성 모형을 일반크리깅 기법을 적용하여 비교하였다. 또한 고혜지와 박만식(2014)은 예측성능의 향상을 위하여 예측지점의 분산 추정값이나 순위값의 비교를 통하여 최종적으로 미지의 지점에 대한 예측값을 등방성 모형, 비등방성 모형, 일반크리깅, 지시크리깅(indicator kriging) 중 최적의 방법의 값의 조합으로 새로운 결과값을 사용하는 방법을 제안하였다. Joseph *et al*. (2008)은 모형 내의 평균이 상수인 정규크리깅 대신 미지의 평균 모형을 베이지안 방법을 이용한 수정된 크리깅 방법인 블라인드 크리깅(blind kriging)을 제안하였다. C*á*ceres *et al*. (2010)은 표준정규분포의 누적분포함수를 이용한 공간자료의 예측방법인 절단 가우시안 크리깅(truncated Gaussian

kriging) 방법을 제안하였다.

본 연구에서는 관측자료가 연속형인 경우 특정한 임계값으로 연속형 변수를 범주형 변수로 변환하여 각 범주에 속할 확률을 예측하는 절단 가우시안 크리깅 방법을 확장하여, t-분포의 누적분포함수를 이용하여 공간자료를 예측하는 확장된 방법에 대하여 연구하고자 한다. 이를 위하여 절단 가우시안 크리깅과 t-분포를 이용하여 절단 t-분포 크리깅(truncated t-distribution kriging)을 적용하였고, 모의실험을 통하여 일반크리깅, 지시크리깅 그리고 절단 가우시안 크리깅, 절단 t-분포 크리깅 등 네 가지 크리깅 방법을 적용하고 정분류율, 민감도, 특이도 관점에서 비교하였다. 또한 환경관측공단 (http://www.airkorea.or.kr/)에서 제공하는 2012년 연평균 미세먼지(PM_{10})자료에 적용하여 미지의 6365개의 격자 지점에 예측을 수행하여 환경부에서 정한 연평균 기준치(50$\mu g/m^3$)를 초과하는 지역을 예측하였다.

본 논문의 순서는 다음과 같이 진행된다. 제 2장에서는 본 연구의 기본이 되는 공간자료 및 공간적 성질에 관련된 내용에 대하여 서술하고, 대표적인 크리깅 방법과 절단 가우시안 크리깅에 대하여 설명한다. 또한, 본 연구에서 제안하는 절단 t-분포 크리깅 방법에 대하여 서술한다. 제 3장에서는 모의자료를 생성하여 각 크리깅 방법을 적용한 모의실험에 대하여 설명한다. 제 4장에서는 본 연구에서 제안한 예측방법인 절단 t-분포 크리깅을 실제 관측 자료에 적용시킨 실증 분석에 대하여 서술한다. 마지막으로 제 5장에서는 본 연구의 결론, 그리고 향후 연구 과제에 대하여 서술한다.

제 2 장

크리깅의 이론적 배경

2.1 공간통계학의 기본 개념

2.1.1 공간자료의 특성

공간통계학이란 위치정보와 그에 따른 관측 자료로 이루어진 공간자료를 분석하는 응용통계학의 한 분야이다. 위도, 경도로 이루어진 2차원 형태의 공간자료는 다음 식과 같이 표현할 수 있다.

$$\{Z(\mathbf{s}_i); \mathbf{s}_i = (s_{1i}, s_{2i})^T, i = 1, 2, \cdots, n\} \tag{2.1}$$

여기서, $\mathbf{s}_i$는 자료의 위치를 의미하고 $Z(\mathbf{s}_i)$는 위치 $\mathbf{s}_i$에서 얻은 관측값을 의미한다. 임의의 두 관측 지점 $\mathbf{s}_i$와 $\mathbf{s}_j$에 대하여 두 지점의 공간적차이(spatial difference)는 다음과 같이 표현한다.

$$\mathbf{h} = (\mathrm{h}_1, \mathrm{h}_2)^T = (\mathrm{s}_{1i} - \mathrm{s}_{1j}, \mathrm{s}_{2i} - \mathrm{s}_{2j})^T = \mathbf{s}_i - \mathbf{s}_j.$$

본 논문에서는 관측지점의 거리는 $\|\mathbf{h}\| = \sqrt{\mathrm{h}_1^2 + \mathrm{h}_2^2}$으로 유클리드 거리이다.

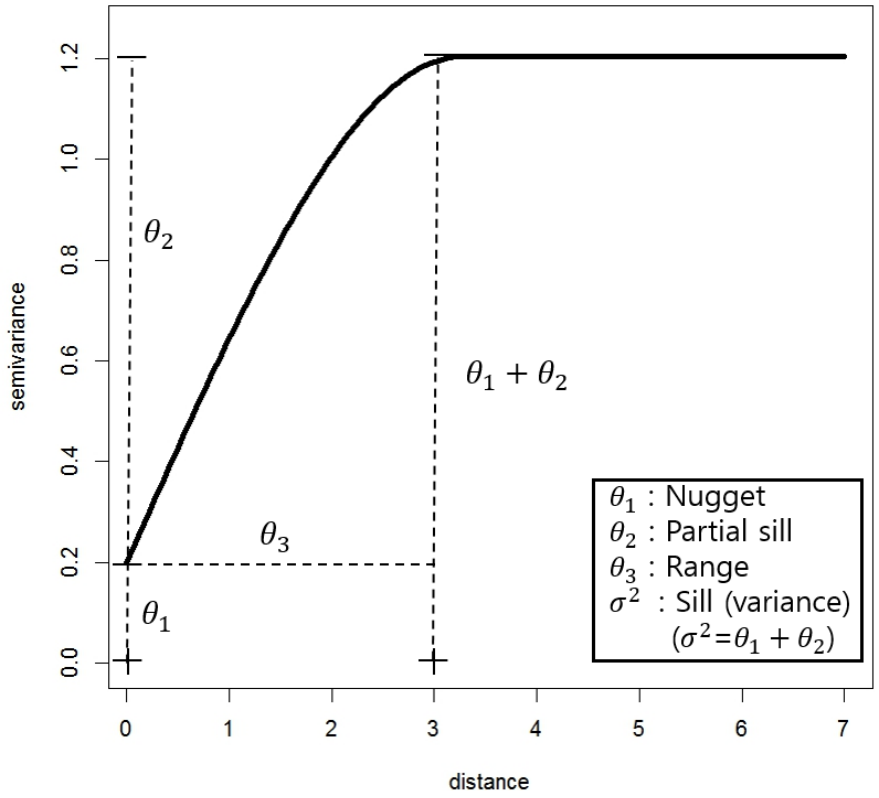


그림 1: 이론적 세미베리오그램(구형모형)

2.1.2 공간 상관성

공간자료는 일반적으로 공간상관성을 가지는데, 이는 공간상에서 얻어진 관측값들 사이의 상관성을 의미하며 관측지점의 거리가 가까울수록 유사성을 갖는 특징을 말한다. 공간자료의 상관정도와 상관을 갖는 패턴을 모형화 한 것이 세미베리오그램이다. 세미베리오그램은 경험적(empirical) 세미베리오그램과 이론적(theoretical) 세미베리오그램이 있다. 그림 1은 이론적 세미베리오그램의 형태를 나타낸 것이다. 그림 1에서 *x*축은 유클리드 거리를 의미하고 *y*축은 세미베리오그

램값을 의미한다. 그림 1에서 거리가 멀어질수록 세미베리오값은 점점 증가하여 더 이상 공간상관성이 존재하지 않게 되는데, 이 때의 y값을 문턱(sill, σ^2)이라 하며 이는 공간자료의 분산을 의미한다. 문턱은 너겟(nugget, θ_1)과 부분문턱(partial sill, θ_2)으로 이루어져있다. 너겟은 자료의 측정오차를 의미하며, 부분문턱은 공간상관성이 없어지는 지점에서의 측정오차를 제외한 분산을 의미한다. 범위(range, θ_3)는 자료들이 공간상관성을 갖는 최대한의 거리를 의미한다. 이와 반대로, 공분산모형은 공간자료의 거리가 가까울수록 상관성이 높으므로, 거리가 멀어질수록 공분산 모형은 값이 감소하다가 일정한 거리가 되면 공간상관성이 없어져 거의 0에 가까운 값을 갖는다. 따라서, 공분산모형은 그림 1를 뒤집어놓은 형태가 될 것이다. 대표적인 이론적 세미베리오그램 모형에는 지수모형, 가우시안모형, 구형모형, 마틴모형 등이 있다. 이에 관한 식은 다음과 같다. 여기서, $\gamma(\mathbf{h})$은 세미베리오그램이다.

(1) **지수모형(exponential model)**

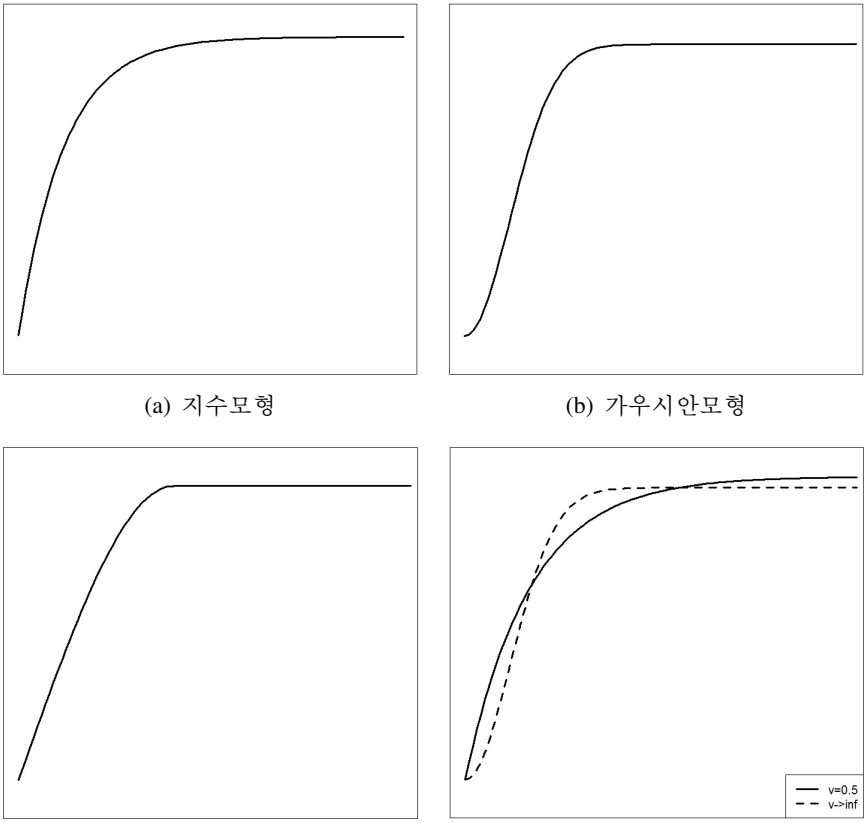
$$\gamma(\mathbf{h}) = \theta_1^2 + \theta_2^2 \left[1 - exp\left(-\frac{\|\mathbf{h}\|}{\theta_3} \right) \right], \quad \|\mathbf{h}\| > 0.$$

(2) **가우시안모형(Gaussian model)**

$$\gamma(\mathbf{h}) = \theta_1^2 + \theta_2^2 \left[1 - exp\left(-\frac{\|\mathbf{h}\|^2}{\theta_3^2} \right) \right], \quad \|\mathbf{h}\| > 0.$$

(3) **구형모형(spherical model)**

$$\gamma(\mathbf{h}) = \begin{cases} \theta_1^2 + \theta_2^2 \left[\dfrac{3\|\mathbf{h}\|}{2\theta_3} - \dfrac{\|\mathbf{h}\|^3}{2\theta_3^3} \right], & \|\mathbf{h}\| \le \theta_3 \\ \theta_1^2 + \theta_2^2, & \|\mathbf{h}\| > \theta_3. \end{cases}$$



(a) 지수모형 (b) 가우시안모형

(c) 구형모형 (d) 마턴모형

그림 2: 이론적 세미베리오그램

(3) **마턴모형(Matérn model)**

$$\gamma(\mathbf{h}) = \theta_1^2 + \theta_2^2 \left[1 - \frac{(2\sqrt{\nu}\|\mathbf{h}\|)^\nu}{2^{\nu-1}\rho^\nu \Gamma(\nu)} K_\nu \left(\frac{2\sqrt{\nu}\|\mathbf{h}\|}{\rho}. \right) \right]$$

여기서, Γ는 감마함수, K_ν는 ν의 수정된 베셀함수(bessel function)이다. ν가 0.5이면 지수모형의 세미베리오그램과 동일하고, ν이 무한한 값을 가지면 가우시안모형과 동일하다.

경험적 세미베리오그램 식은 다음과 같다.

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2.$$

여기서, $N(\mathbf{h})$는 유클리드 거리 $\|\mathbf{h}\|$를 갖는 자료 전체 쌍의 집합이다. $|N(\mathbf{h})|$ 은 $N(\mathbf{h})$에 속하는 쌍의 갯수를 의미한다. 이러한 세미베리오그램은 거리가 가까울수록 작은 값을 가지고, 거리가 멀어질수록 큰 값을 가진다. 또한 일정거리 이상이 되면 세미베리오 값은 거의 증가하지 않게 되는데 이는 자료간의 공간상관성이 존재하지 않는 것을 의미한다. 공분산모형은 이와 반대로 거리가 가까울수록 상관성이 높아 두 지점의 공분산은 큰 값을 갖게 되고, 거리가 멀어질수록 공간상관성이 사라져서 거의 0에 가까운 값을 가지게 된다. 세미베리오그램과 공분산모형의 관계는 다음과 같이 표현 할 수 있다.

$$\gamma(\mathbf{h}) = \sigma^2 - C(\mathbf{h}).$$

여기서, $C(\mathbf{h})$은 공분산 모형을 의미한다.

2.2 크리깅

공간통계학에서 공간자료를 분석하여 미지의 지점에 대한 예측값을 도출하는 것을 크리깅이라 한다. 이를 위하여 크리깅은 예측 지점 주위의 관측값의 선형결합을 이용한다. 크리깅을 적용하기 위하여 공간적 상관관계를 나타내는 세미베리오그램이나 공분산모형이 사용되며 이에 대한 정의가 필요하다. 본 절에서는 크리깅의 기본개념과 크리깅의 종류에 대해서 다룬다. 공간자료는 식 (2.1)과 같이 나타낼 수 있고, 이는 $\mathbf{z} = [Z(\mathbf{s}_1), Z(\mathbf{s}_2), ..., Z(\mathbf{s}_n)]^T$와 같이 표현할 수 있다. 이러한 공간자료는 다음과 같은 선형모형으로 나타낼 수 있다.

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.2}$$

여기서, $\mathbf{X}$는 크기 $(n \times p)$인 예측인자벡터, $\boldsymbol{\beta} = [\beta_1, ..., \beta_p]^T$는 회귀계수벡터이고, $\mathbf{X}\boldsymbol{\beta}$는 평균함수(mean function)로 방향(direciton) 혹은 추세(trend)를 의미한다. 이 때, $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma})$는 오차항을 의미한다. $\boldsymbol{\Sigma}$은 분산-공분산행렬(variance-covariance matrix)이다. 이를 이용하여 미지의 지점 $\mathbf{s}_0$의 예측값을 구하기 위한 식은 다음과 같다.

$$Z(\mathbf{s}_0) = \sum_{i=1}^{n} a_i Z(\mathbf{s}_i) = \mathbf{a}^T \mathbf{z}. \tag{2.3}$$

여기서, $\mathbf{s}_0$은 예측하고자 하는 미지의 지점, $Z(\mathbf{s}_i)$는 이미 관측된 지점 $\mathbf{s}_i$의 관측값이다. $\mathbf{a} = [a_1, a_2, ..., a_n]^T$는 가중치벡터이고, 이때, 가중치벡터는 $\mathbf{a}^T\mathbf{1} = 1$을 만족한다. 이러한 가중치벡터를 구하기 위한 크리깅의 종류에는 단순크리깅, 보통크리깅, 일반크리깅, 지시크리깅 등이 있으며, 다음 절에서 이에 대하여 다룬다.

2.2.1 일반크리깅

일반크리깅은 공간통계학에서 가장 보편적으로 사용되는 크리깅 방법으로, 단순크리깅과 보통크리깅은 일반크리깅의 특수한 경우이다. 일반크리깅을 적용하기 위하여 공간자료 $\mathbf{z} = [Z(\mathbf{s}_1), Z(\mathbf{s}_2), ..., Z(\mathbf{s}_n)]^T$는 식 (2.3)과 같이 선형결합 형태로 표현 가능해야한다. 이 때, $Var(\mathbf{z}) = \mathbf{\Sigma}$, $Var[Z(\mathbf{s}_0)] = \sigma^2$라 가정한다. 또한 관측값과 예측값이 공간적 상관을 가지면, $Cov[\mathbf{z}, Z(\mathbf{s}_0)] = \boldsymbol{\sigma}$으로 표현할 수 있고, $\boldsymbol{\sigma}$는 $(n \times 1)$의 벡터이다. 일반크리깅은 불편성(unbiasedness)과 최소분산(minimum variance)을 가지는 최적의 선형예측인자를 찾는데 목적이 있다. 그러므로, 식 (2.3)의 형식의 예측인자를 고려한다. 여기서 평균제곱오차 $E[(\mathbf{a}^T\mathbf{z} - Z(\mathbf{s}_0))^2]$를 최소화하는 가중치벡터 $\mathbf{a}$를 찾음으로써 $\mathbf{a}^T\mathbf{z}$는 최량선형불편예측량(best linear unbiased predictor, BLUP)이 된다. 이를 구하기 위하여 필요한 평균제곱오차는 식 (2.4)와 같다.

$$\begin{aligned} E[(\mathbf{a}^T\mathbf{z} - Z(\mathbf{s}_0))^2] &= Var[\mathbf{a}^T\mathbf{z}] + Var[Z(\mathbf{s}_0)] - 2Cov[\mathbf{a}^T\mathbf{z}, Z(\mathbf{s}_0)] \\ &= \mathbf{a}^T\mathbf{\Sigma}\mathbf{a} + \sigma^2 - 2\mathbf{a}^T\boldsymbol{\sigma}. \end{aligned} \tag{2.4}$$

식(2.4)는 $E[\mathbf{a}^T\mathbf{z}] = E[Z(\mathbf{s}_0)]$라는 조건 하에서 성립한다.

$$\mathbf{a} = \mathbf{\Sigma}^{-1}[\boldsymbol{\sigma} - \mathbf{X}(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}\mathbf{\Sigma}^{-1}\boldsymbol{\sigma} - x(\mathbf{s}_0))]. \tag{2.5}$$

여기서, $x(\mathbf{s}_0)$는 지점 $\mathbf{s}_0$에서 관측되는 $(p \times 1)$ 크기의 예측인자벡터이다. $\boldsymbol{\beta}$의 일반화최소제곱추정량(generalized least-squares estimator), $\hat{\boldsymbol{\beta}}_G = (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{z}$이다. 따라서, $Z(\mathbf{s}_0)$의 최량선형불편예측량은 다

음과 같이 표현 할 수 있다.

$$\mathbf{a}^T\mathbf{z} = \hat{Z}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)^T\hat{\boldsymbol{\beta}}_G + \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}_G).$$

$\hat{Z}(\mathbf{s}_0)$의 크리깅 분산은 다음과 같다.

$$Var[\hat{Z}(\mathbf{s}_0)] = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} + \sigma^2 - 2\mathbf{a}^T\boldsymbol{\sigma}.$$

2.2.2 지시크리깅

지시크리깅은 임계값(threshold value)을 넘을 확률을 추정하는 방법으로, 연속형인 자료값을 0, 1의 지시값으로 변환한 이항자료를 연속형 자료라 가정하여 일반크리깅 기법을 적용한다. 이를 위하여 공간자료는 강한 정상성(strict stationarity)을 가정하고, $Z(\mathbf{s})$는 다음과 같이 지시 변환을 실시한다.

$$I_{(\mathbf{s},z)} = \begin{cases} 1, & if \quad Z(\mathbf{s}) \leq c \\ 0, & otherwise. \end{cases}$$

지시크리깅의 예측결과는 임계값을 넘을 확률로 계산된다.

2.2.3 절단 가우시안 크리깅

절단 가우시안 크리깅은 일반크리깅 방법을 응용하여 지시크리깅의 대안으로 고안된 방법으로, C*á*ceres *et al.* (2010)가 제안하였다. 절단 가우시안 크리깅은 연속형으로 관찰된 공간자료를 특정한 값으로 자료를 여러 개의 범주로 나눌 수 있을 때, 특정 범주에 속할 확률을 계산하는 방법이다. 각 범주에 속할 모확률(population probability)은 식

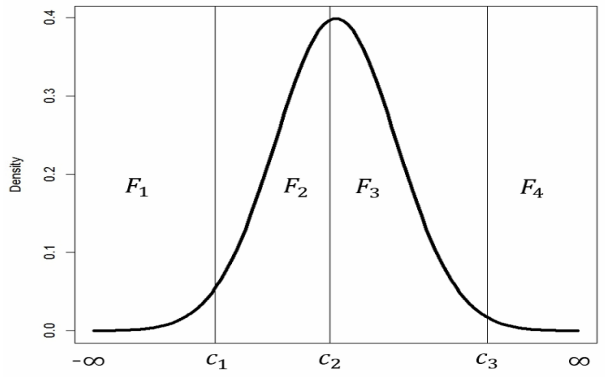


그림 3: 표준정규분포 확률밀도

(2.6)과 같다.

$$Pr(\mathbf{s}_0 \in F_q) = \Phi\left(\frac{c_q - Z(\mathbf{s}_0)}{\sigma(\mathbf{s}_0)}\right) - \Phi\left(\frac{c_{q-1} - Z(\mathbf{s}_0)}{\sigma(\mathbf{s}_0)}\right). \qquad (2.6)$$

여기서, c_q는 임계값, q는 임계값의 갯수이다. F_q는 특정 범주 q를 의미한다. Φ는 표준정규분포의 누적분포 함수이다. 식 (2.6)의 모확률값을 추정하기 위한 식은 다음과 같다. $q = 1,\ldots,Q,\ \ j = 1,\ldots,R$에 대해서

$$\widehat{Pr}(\mathbf{s}_0 \in F_q) = \frac{1}{R}\sum_{j=1}^{R}\left[\Phi\left(\frac{c_q - \hat{Z}^j(\mathbf{s}_0)}{\hat{\sigma}(\mathbf{s}_0)}\right) - \Phi\left(\frac{c_{q-1} - \hat{Z}^j(\mathbf{s}_0)}{\hat{\sigma}(\mathbf{s}_0)}\right)\right].$$

여기서, j은 재생성한 표본의 갯수, $\hat{Z}^j(\mathbf{s}_0)$은 j번째 표본에서 계산된 $\mathbf{s}_0$의 일반크리깅 예측값, $\hat{\sigma}(\mathbf{s}_0)$은 원자료의 일반크리깅 분산에 제곱근을 취한 값이다. 확률 추정값을 구하기 위해 R개로 재생성한 표본들을 일반크리깅에서 구한 예측값과 크리깅 분산값을 이용하여 표준화를 한다. 마지막으로, 각 표본에서 추정된 확률값들의 산술평균을 통하여 각 범주에 속할 확률값을 추정한다. 예를들어, 그림 3의 F_2의 속할

확률을 절단 가우시안 통하여 구하는 식은 다음과 같다.

$$\widehat{Pr}(\mathbf{s}_0 \in F_2) = \frac{1}{R}\sum_{j=1}^{R}\left[\Phi\left(\frac{c_2 - \hat{Z}^j(\mathbf{s}_0)}{\hat{\sigma}(\mathbf{s}_0)}\right) - \Phi\left(\frac{c_1 - \hat{Z}^j(\mathbf{s}_0)}{\hat{\sigma}(\mathbf{s}_0)}\right)\right].$$

여기서, 그림 3은 표준정규분포(standard normal distribution)로부터 생성한 자료를 임계값 c_1, c_2, c_3를 이용하여 4개의 범주로 나눈 것이다. 여기서 각 범주를 F_1, F_2, F_3, F_4라 한다.

2.3 t-분포 하에서의 절단크리깅을 이용한 공간자료의 예측

2.3.1 t-분포의 확률밀도

t-분포의 정의는 다음과 같다. 확률변수 W와 V는 서로 독립이며 W는 $N(0,1)$을 따르는 확률변수이고, V는 자유도 r인 카이제곱분포(chi-squared distribution)를 따르는 확률변수라고 하자. W와 V의 결합확률밀도함수(joint probability density function), $h(w,v)$는 다음 식과 같다. $-\infty < w < \infty, 0 < v < \infty$에 대하여

$$h(w;v) = \frac{1}{\sqrt{2\pi}}e^{-w^2/2}\frac{1}{\Gamma(r/2)2^{r/2}}v^{r/2-1}e^{-v/2}.$$

새로운 확률변수 T를 다음과 같이 정의할 때, 확률변수 T는 t-분포를 따른다.

$$T = \frac{W}{\sqrt{V/r}}.$$

t-분포는 좌우대칭의 종모양인 분포로서 평균, 중위수, 최빈값 모두 0이다. t-분포의 모양은 자유도로 결정되며 자유도가 커질수록 표준정

규분포에 근사하게 된다. t-분포의 자유도를 k라 하였을 때 t-분포의 분산은 다음과 같다. 임의의 $k > 2$에 대하여

$$Var(T) = \frac{k}{k-2}.$$

2.3.2 절단 t-분포 크리깅

절단 t-분포 크리깅은 기존의 절단 가우시안 크리깅 방법을 t-분포로 확장한 방법이다. 확률변수의 분포가 좌우대칭이면서 정규분포보다 두터운 꼬리를 가지는 형태일 때, 표준정규분포의 누적분포함수가 아닌 t-분포의 누적분포함수를 이용하여 각 범주에 속할 확률을 예측하는 방법에 대하여 연구하고자 한다. 본 연구에서는 이러한 방법을 절단 t-분포 크리깅이라 정의하겠다. 또한 절단 가우시안 크리깅과 절단 t-분포 크리깅 두 방법을 절단 크리깅이라 통칭한다. 절단 t-분포 크리깅 방법의 모확률을 구하는 식 (2.7)와 같다.

$$Pr(\mathbf{s}_0 \in F_q) = G_k\left(\frac{c_q - Z(\mathbf{s}_0)}{\sigma(\mathbf{s}_0)}\right) - G_k\left(\frac{c_{q-1} - Z(\mathbf{s}_0)}{\sigma(\mathbf{s}_0)}\right) \qquad (2.7)$$

여기서 G_k는 자유도 k인 t-분포의 누적분포함수를 의미한다. 각 범주에 속할 확률 추정값은 다음 식과 같다. $q = 1, \ldots, Q, \;\; j = 1, \ldots, R$에 대하여

$$\widehat{Pr}(\mathbf{s}_0 \in F_q) = \frac{1}{R}\sum_{j=1}^{R}\left[\hat{G}_k\left(\frac{c_q - \hat{Z}^j(\mathbf{s}_0)}{\hat{\sigma}(\mathbf{s}_0)}\right) - \hat{G}_k\left(\frac{c_{q-1} - \hat{Z}^j(\mathbf{s}_0)}{\hat{\sigma}(\mathbf{s}_0)}\right)\right]. \quad (2.8)$$

여기서, j은 재생성한 표본의 갯수, $\hat{Z}^j(\mathbf{s}_0)$은 j번째 표본에서 계산된 $\mathbf{s}_0$의 일반크리깅 예측값, $\hat{\sigma}(\mathbf{s}_0)$은 원자료의 일반크리깅 분산에 제곱근을 취한 값이다. 절단 t-분포 크리깅 확률 추정값을 구하기 위하여 절단

가우시안 크리깅과 동일한 절차를 거친다. 먼저, R개의 표본을 일반 크리깅 결과로 표준화한다. 그리고 표준화된 값을 자유도 k를 따르는 t-분포의 누적분포함수를 이용하여 확률값을 추정한다. 마지막으로, R개의 표본에서 추정된 확률값들의 산술평균을 구하여, 특정 범주에 속할 확률 추정값을 구한다.

절단 t-분포 크리깅을 적용하기 위해서는 t-분포의 누적분포함수를 이용하는데, 그 값을 계산하려면 분포의 자유도에 대한 정보가 있어야 한다. 따라서 절단 가우시안 크리깅에서와 달리 자유도를 추정해야 하는 단계가 필요하다. 자유도를 추정하기 위한 방법은 다음과 같다. 먼저, 공간자료를 식 (2.2)과 같이 가정하고, 여기서 $\boldsymbol{\epsilon} = \Gamma^T \boldsymbol{\delta}$이라 하면 $\hat{\boldsymbol{\epsilon}}$은 다음 식과 같이 구할 수 있다.

$$\hat{\boldsymbol{\epsilon}} = \mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}_O = \hat{\Gamma}^T \hat{\boldsymbol{\delta}}$$

여기서, $\hat{\boldsymbol{\beta}}_O$은 최소제곱추정량(ordinary least-squares estimator)이고, $\hat{\Gamma}^T$은 추정된 공간모수 $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$을 이용하여 구한 분산-공분산행렬을 촐레스키분해(cholesky decomposition)한 것이다. 또한, $\hat{\boldsymbol{\delta}} = [\hat{\delta}_1, \hat{\delta}_2, ..., \hat{\delta}_n]^T$는 (n×1) 크기를 가지는 벡터이다. 따라서, $\hat{\boldsymbol{\delta}}$는 다음과 같다.

$$\hat{\boldsymbol{\delta}} = (\hat{\Gamma}^T)^{-1} \hat{\boldsymbol{\epsilon}}$$

자유도를 추정하는 방법인 최대우도추정법에 대한 자세한 내용은 다음장에 서술한다.

2.3.3 최대우도추정법

최대우도추정법(maximum likelihood estimation method)이란 모수를 추정할 때 사용하는 방법으로, 우도함수(likelihood function) 혹은 로그우도함수를 이용하여 모수를 추정하는 방법이다. 본 연구에서 $\boldsymbol{\delta}$의 자유도를 추정하기 위하여 최대우도추정법이 사용되며 이 때, $\hat{\boldsymbol{\delta}}$는 서로 독립이다. 결합확률밀도함수가 k의 함수일 때 이를 우도함수라 한다. $i = 1, 2, \cdots, n$에 대하여

$$L(k;\hat{\boldsymbol{\delta}}) = f(\hat{\delta}_1;k)f(\hat{\delta}_2;k)\ldots f(\hat{\delta}_n;k) = \prod_{i=1}^{n} f(\hat{\delta}_i;k).$$

최대우도추정법이란 우도함수 $L(k;\hat{\boldsymbol{\delta}})$를 최대로 하는 k를 찾는 것이고, 그 때의 추정량 $\hat{k}$를 최대우도추정량(maximum likelihood estimatior)이라 한다. 즉, 실제로 관측된 자료가 얻어질 확률이 가장 높은 k을 찾는 것을 최대우도추정법이라 한다. 크리깅에 대한 보다 더 자세한 내용은 Cressie(1993)를 참고하길 바란다.

제 3 장

모의실험

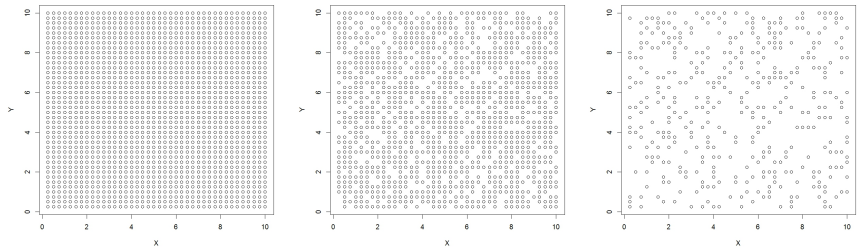
3.1 모의실험 방법

본 장에서는 각각의 크리깅 예측 성능을 비교하기 위하여 표준정규분포와 자유도 3, 5, 10, 20을 따르는 t-분포를 이용하여 총 5가지의 분포에서 자료를 생성하여 모의실험을 실시하였다. 본 모의실험에서의 공간자료는 2차 정상성을 만족한다고 가정하였고, 등방성 모형을 가정하였다. 모의실험자료의 위치지점은 x축, y축 각각 0.25부터 10까지 0.25의 등간격을 가지는 총 1600개의 격자자료로 생성하였다. 각 위치지점에서의 자료값은 공분산모형(세미베리오그램)을 이용하여 생성하였다.

$$\mathbf{z} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}) \tag{3.1}$$

여기서 $\mathbf{z} = \Gamma^T \boldsymbol{\delta}$이고, $\boldsymbol{\delta}$는 표준정규분포와 자유도 3, 5, 10, 20을 가지는 t-분포에서 임의 생성한 값이다. 또한 Γ는 분산-공분산행렬을 촐레스키 분해한 값이다. 모의실험자료의 공간모수 참값은 θ_1=0.1, θ_2=0.9, θ_3=3이고, 임계값은 0이다. 임계값을 기준으로 자료를 이항자료로 변환하여 각 크리깅 방법에 적용하였다.

절단 t-분포 크리깅 방법을 적용하기 위해서는 각 분포마다 자유도 추정이 필요한데 이는 R에 내장된 함수 `fitdistr`를 이용하였다. 모의실험의 절차는 다음과 같다.



(a) 1600개 지점 (b) 1200개 지점 (c) 400개 지점

그림 4: 모의실험자료 위치지점

(1) 0.25부터 10까지 등간격을 가지는 1600개의 격자자료를 생성한다.

(2) (1)에서 구한 위치자료를 이용하여 구형모형 하에서 분산-공분산 모형(세미베리오그램)을 구한 뒤 분산-공분산행렬을 계산한다.

(3) 표준정규분포와 각각 자유도 3, 5, 10, 20을 가지는 t-분포 하에서 각각 100개 생성한다.

(4) (3)에서 각 분포에서 생성한 자료와 (2)에서 구한 분산-공분산행렬을 결합하여 자료값을 생성한다.

$$\mathbf{z} = \Gamma^T \boldsymbol{\delta}.$$

(5) (1)에서 생성한 1600개 위치지점을 그림 4와 같이 임의로 1200개 지점과 400개 지점으로 나눈다.

(6) (5)의 400개 지점에서 200개 지점을 완전임의추출로 100개 표본을 재생성한다. 또한 400개 지점으로 공간모수($\theta_1, \theta_2, \theta_3$)를 추정

하고, 크리깅 분산을 구한다.

(7) (6)에서 재생성한 100개 표본 중 99개 표본을 이용하여 절단 가우시안 크리깅, 절단 t-분포 크리깅을 예측값을 구한다.

(8) (6)과 (7)에서 재생성한 100개 표본 중 나머지 1개 표본을 이용하여 일반크리깅과 지시크리깅 방법을 이용하여 예측값을 구하고, 일반크리깅 결과값은 연속형 값으로 예측되므로, 임계값 0을 이용하여 지시변환을 실시한다.

(9) (7)과 (8)에서 구한 크리깅 예측값들을 이용하여 크리깅 방법의 성능을 정분류율, 민감도, 특이도를 이용하여 비교한다.

본 모의실험에서는 재표집된 200개 지점으로 각 크리깅 방법을 적용하여 1200개 지점의 예측값을 구하였다. 일반크리깅, 지시크리깅, 절단 가우시안 크리깅, 절단 t-분포 크리깅의 예측 성능을 비교하기 위해서 정분류율, 민감도, 특이도 등의 척도를 이용하였다.

각 분포에서 200개 지점을 가지는 자료를 100개씩 생성하여, 생성된 100개의 표본을 각 크리깅 방법을 적용하여 예측값을 구하였다. 이에 대하여, 크리깅 방법별로 분할표를 작성하여 정분류율, 민감도, 특이도를 구할 수 있다. 모의실험에서는 각 크리깅 방법의 정분류율, 민감도, 특이도를 비교하여 가장 높은 값에 일순위를 부여하여, 각 크리깅방법마다 100개 자료 중 일순위를 가지는 자료의 갯수를 통하여 각 크리깅 방법의 성능의 차이를 비교하고자 한다.

(a) 원자료 (b) TGK 예측결과 (c) TTK 예측결과

그림 5: t(10) 예측 결과

1200개 지점의 값이 임계값보다 높게 나타난 관측값(예측값)은 음영처리함.; TGK, 절단 가우시안 크리깅.; TTK, 절단 t-분포 크리깅.

3.2 모의실험 연구 결과

그림 5은 자유도 10을 따르는 t-분포에서 첫 번째로 생성한 자료의 예측 결과를 그림으로 표현한 것이다. 그림 5 (a)는 원자료 값이 자료의 0보다 큰 경우를 1로 주어 1인 경우를 표시, (b)는 절단 가우시안 크리깅 예측값이 0보다 크다고 예측된 지점을 1로 주어 1인 경우를 표시, (c)는 절단 t-분포 크리깅 결과 0보다 크다고 예측된 지점을 1로 주어 1인 경우를 표시한 그림이다. 따라서, 생성된 자료마다 각 크리깅별로 분할표를 작성하여 정분류율, 민감도, 특이도를 구할 수 있고, 생성된 각 자료마다 분할표는 일반크리깅, 지시크리깅, 절단 가우시안 크리깅, 절단 t-분포 크리깅의 4개 분할표를 생성할 수 있다. 4개의 분할표에서 각각 정분류율, 민감도, 특이도를 계산할 수 있고, 그 중 가장 높은 값을 가지는 크리깅방법에 각 척도마다 일순위를 부여하여 각 방법을 비교할 수 있다.

본 연구에서 제안한 절단 t-분포 크리깅을 적용하기 위해서는 t-분

표 1: 모수추정결과

Distribution		θ_1	θ_2	θ_3	k
Normal	mean ± sd	0.103±0.03	0.887±0.171	3.008±0.406	57.431±29.897
	IQR	(0.083, 0.124)	(0.778, 1.000)	(2.871, 3.135)	(35.367, 79.000)
t(3)	mean ± sd	0.290±0.165	2.676±0.843	3.065±0.482	3,300±0.360
	IQR	(0.178, 0.366)	(2.214, 2.950)	(2.834, 3.111)	(3.015, 3.488)
t(5)	mean ± sd	0.165±0.058	1.521±0.373	3.069±0.460	5.528±0.782
	IQR	(0.125, 0.194)	(1.267, 1.688)	(2.894, 3.180)	(4.998, 6.033)
t(10)	mean ± sd	0.125±0.034	1.147±0.244	3.092±0.469	13.870±8.215
	IQR	(0.094, 0.152)	(0.997, 1.315)	(2.918, 3.191)	(9.179, 16.052)
t(20)	mean ± sd	0.110±0.025	1.002±0.228	2.988±0.448	25.022±14.579
	IQR	(0.900, 0.128)	(0.879, 1.083)	(2.845, 3.083)	(17.019, 27.944)

t(k), t-distribution.

포의 자유도를 추정해야 한다. 자유도를 추정하기 위해서는 먼저 공간모수 너겟(θ_1), 부분문턱(θ_2), 범위(θ_3)을 추정해야한다. 추정된 결과는 표 1 같다.

표 1를 보면, t-분포의 자유도가 커짐에 따라 공간모수 너겟, 부분문턱, 범위의 추정값의 평균이 참값과 유사한 값을 가지는 것을 알 수 있다. 또한, 표준 정규분포에서 생성한 자료를 절단 t-분포 크리깅에 적용시키기 위해서는 표준 정규분포에서 생성된 자료를 t-분포라 가정하여 자유도를 구한 결과, 추정된 자유도의 평균은 57.431으로 추정되었다.

모의실험자료를 각 크리깅에 적용한 뒤 각각 정분류율, 민감도, 특이도를 계산하였다. 표 2는 각각의 크리깅별 정분류율, 민감도, 특이도의 일순위를 가지는 자료의 갯수를 표기한 표이다. 먼저 모집단이 정규분포인 경우를 살펴보면, 모든 척도에서 절단 t-분포 크리깅(TTK)와 절단 가우시안 방법(TGK)의 차이가 거의 없는 것을 알 수 있다. 자유

표 2: 크리깅방법별 일순위 비교

		TTK	TGK	UK	IK	TTK=TGK
Normal	정분류율	99	100	0	0	99
	민감도	61	61	19	13	61
	특이도	64	65	16	13	64
t(3)	정분류율	85	81	5	4	78
	민감도	70	70	11	17	67
	특이도	53	52	28	16	51
t(5)	정분류율	91	92	5	18	88
	민감도	63	64	15	18	63
	특이도	51	52	21	23	50
t(10)	정분류율	94	93	5	0	92
	민감도	54	54	25	20	53
	특이도	70	70	16	14	70
t(20)	정분류율	96	95	3	0	94
	민감도	60	61	23	11	60
	특이도	66	66	19	15	66

TTK, 절단t-분포크리깅; TGK, 절단가우시안크리깅; UK, 일반크리깅; IK, 지시크리깅.

도 3을 갖는 t-분포의 경우를 살펴보자. 정분류율의 경우 TTK가 85개, TGK가 81개로 다소 차이를 보이고, 민감도는 일순위를 갖는 자료값이 70개로 동일한 것을 알 수 있다. 자유도 5인 경우를 살펴보면 정분류율의 경우 TTK가 91개, TGK가 92개였고, 민감도, 특이도 또한 일순위를 가지는 자료값이 1개씩 차이나는 것을 알 수 있다. 자유도 10인 경우의 정분류율은 TTK가 94개, TGK가 93개이고, 민감도와 특이도는 동일한 값을 가졌다. 마지막으로 자유도 20인 경우를 살펴보자. 정분류율의 경우 TTK가 96개, TGK가 95개이고, 민감도는 TTK가 60개, TGK가 61개이고 특이도는 동일한 값을 가진다. 결과적으로 자유도 3의 자료의

표 3: 크리깅방법별 일순위 확률 비교

		TTK	TGK	UK	IK
Normal	정분류율	0.495	0.5	0	0
	민감도	0.305	0.305	0.19	0.13
	특이도	0.32	0.325	0.16	0.13
t(3)	정분류율	0.425	0.405	0.05	0.04
	민감도	0.35	0.35	0.11	0.17
	특이도	0.265	0.26	0.28	0.16
t(5)	정분류율	0.455	0.46	0.05	0.18
	민감도	0.315	0.32	0.15	0.18
	특이도	0.255	0.26	0.21	0.23
t(10)	정분류율	0.47	0.465	0.05	0
	민감도	0.27	0.27	0.25	0.2
	특이도	0.35	0.35	0.16	0.14
t(20)	정분류율	0.48	0.475	0.03	0
	민감도	0.3	0.305	0.23	0.11
	특이도	0.33	0.33	0.19	0.15

TTK, 절단t-분포크리깅; TGK, 절단가우시안크리깅; UK, 일반크리깅; IK, 지시크리깅

정분류율은 TTK와 TGK가 다소 차이를 보이나, 나머지 분포에서는 큰 차이를 보이지 않았다. 그러나 자유도가 커짐에 따라, 정분류율의 경우에 TTK와 TGK가 동일한 값을 가지는 자료의 갯수가 증가하는 것을 알 수 있다. 이를 통하여 자유도가 커짐에 따라 절단 가우시안 크리깅 방법과 절단 t-분포 크리깅 방법이 차이가 감소하는 것을 알 수 있었다.

표 3은 표 2의 결과를 확률로 표현한 것이다. 확률을 계산할 때, 절단 가우시안 크리깅과 절단 t-분포 크리깅이 결과값이 동일한 경우가 많으므로, 두 방법에 한하여 표 2의 결과에 각각 0.5씩 곱해주어 확률값을 계산하였다. 표 3에서 정분류율의 경우 TTK가 가장 확률값이

높은 분포는 자유도 3, 자유도 10, 자유도 20이다. 민감도의 경우, 정규분포와 자유도 3, 자유도 10의 경우 TTK와 TGK가 동일하였고, 그 외 자유도 5와 20은 TGK의 확률값이 더 높았다. 특이도의 경우, 자유도 5, 10, 20은 TTK와 TGK가 동일한 값을 가졌고, 표준정규분포와 자유도 3인 분포는 TGK가 더 높은 값을 가졌다. 정분류율의 경우, TTK와 TGK 두 방법 모두 자유도가 커짐에 따라 점점 0.5에 근사해져가는 것을 알 수 있었다. 이는 자유도가 커짐에 따라 두 방법의 차이가 감소하는 것을 의미하고, 이는 자유도가 작아질수록 두 방법론의 차이가 존재함을 의미한다. 따라서 자료가 t-분포를 따르는 경우에는 절단 가우시안 크리깅 방법보다 자유도를 추정하여 절단 t-분포 크리깅을 적용해야 하는 것을 의미한다.

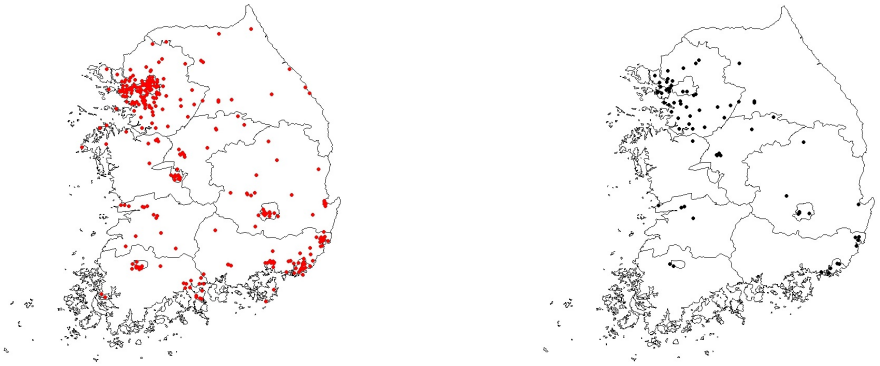
표 2와 표 3을 통하여, TTK방법과 TGK방법의 차이를 정분류율, 민감도, 특이도 관점으로 비교한 결과 t-분포의 자유도가 작은 경우, 두 방법론의 차이가 있음을 보였고, t-분포의 자유도가 커짐에 따라 두 방법론의 차이가 적어지는 것을 확인할 수 있다. 따라서, 자료가 t-분포를 따르는 경우에는 절단 가우시안 크리깅 방법보다 자유도를 추정하여 절단 t-분포 크리깅을 적용해야 할 것이다.

제 4 장

실증자료

4.1 연구 방법

모의실험을 통하여 일반크리깅과 지시크리깅, 절단 가우시안 크리깅, 절단 t-분포 크리깅 등 4가지 방법의 예측결과를 비교해보았다. 본 장에서는 실제 관측 자료를 이용하여 일반크리깅, 지시크리깅, 절단 가우시안 크리깅, 절단 t-분포 크리깅 방법을 적용, 비교할 것이다. 본 연구에서 사용한 실증 자료는 한국환경공단 홈페이지에서 제공하는 2012년도 미세먼지(PM_{10}) 연평균 관측 자료이다. 도시대기 측정망 자료, 도로변대기 측정망 자료, 국가배경농도 측정망 자료, 교외대기 농도 측정망 자료 등을 이용하였고 그 중 울릉도, 제주도에 속해 있는 지점을 제외한 내륙지방의 측정망 303개 지점을 이용하여 6365개의 격자지점을 예측하였다. 또한, 위치자료 뿐 아니라 도시규모변수를 추가하여, 서울특별시와 6개의 광역시에 속해있는 지점의 경우에는 1로, 그 외의 지점은 0으로 부여하였다. 그림 6은 미세먼지 관측지점을 나타낸 지도로서, (a)는 본 연구에서 이용한 303개 지점을 표시한 것이다. 실제 관측 자료를 이용한 분석은 지수모형, 가우시안모형, 구형모형 등 세 가지 공분산 모형 중 가장 잘 적합한 모형인 가우시안모형을 선택하여 각 크리깅 방법을 적용하였다. 일반크리깅 결과는 환경부에서 정한 대기환경 기준인 연평균치 50$\mu g/m^3$을 임계값으로 하여 예측값이 대기환경기준보다 높으면 1로, 낮으면 0으로 지시변환하여 지시크리깅, 절



(a) 전체 측정망 관측지점　　(b) 연평균대기기준 초과 지점

그림 6: 미세먼지 관측지점

단 가우시안 크리깅, 절단 t-분포 크리깅 방법과 비교하였다. 실증자료에서 표본을 생성하기 위해서 교차 타당법(cross-validation)을 이용하였다. 303개 위치지점의 실제 관측자료를 1개씩 제외하여 크기가 302인 표본을 총 303개 생성하였다. 이를 이용하여 6365개의 격자지점에 대해 절단 가우시안 크리강과 절단 t-분포 크리깅을 이용하여 예측하였다.

(a) TGK 예측지도 (b) TGK 대기기준 초과지점

그림 7: 절단 가우시안 크리깅 예측지도

4.2 실증자료 연구 결과

미세먼지 자료를 분석한 결과 원자료의 303개의 관측지점 중 대기기준을 넘는 위치지점은 73개 지점이다. (그림 6 (b)). 먼저 절단 가우시안 크리깅과 결과를 보면, 그림 7는 절단 가우시안 크리깅 예측결과를 지도에 표시한 것이다. (a)는 추정된 확률값을 예측지도로 표현한 것이고, (b)는 미지의 예측지점 6365개 지점 중 예측된 값이 대기기준 50μg/㎥보다 크다고 예측된 지점을 표시한 그림이다. 6365개 지점 중 연평균 대기기준보다 높은 지점은 552개 지점으로 예측되었다. 그림 7을 통하여 대기기준보다 높다고 예측될 가능성이 높은 지역은 서울을 둘러싼 경기 북부와 남부지방, 그리고 충청북도 북쪽 지역이다.

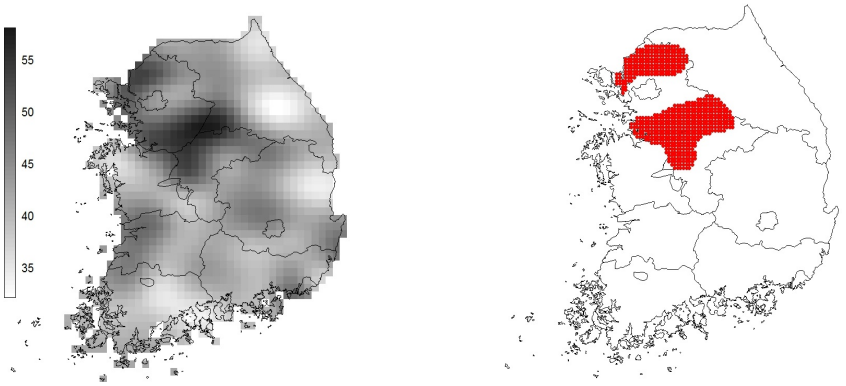
절단 가우시안 크리깅을 절단 t-분포 크리깅에 적용시키기 위하여, 실제 관측 자료를 t-분포라 가정하고 자유도를 추정하였다. 추정된 자유도 값은 37.302으로 이를 이용하여 절단 t-분포 크리깅을 적용한 결과는 그림 8과 같다.

(a) TTK 예측지도 (b) TTK 대기기준 초과지점

그림 8: 절단 t-분포 크리깅 예측지도

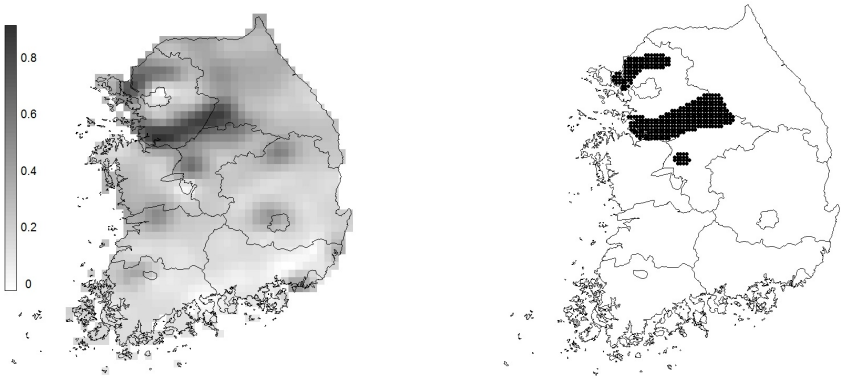
절단 t-분포 크리깅 결과 예측값이 연평균 대기기준보다 큰 지점은 전체 6365개 지점 중 552개 지점으로 절단 가우시안 크리깅의 결과와 동일하다. 이는 추정된 자유도가 37.302으로 큰 값을 가지므로, 표준정규분포와 t-분포는 유사해지기 때문이다. 다음으로 실증 자료를 일반 크리깅과 지시크리깅에도 적용하였는데, 일반 크리깅 결과는 아래의 그림 9과 같다. 그림 9의 (a)는 일반크리깅 예측값의 지도이고, (b)는 예측된 값이 대기기준 50$\mu g/m^3$보다 넘는 지역을 지도에 표시한 것이다. 전체 6365개 지점 중 545개 지점으로, 그 지점은 서울을 기준으로 경기도 북부지방과, 경기도 남쪽지역이 연평균 기준보다 높을 것으로 예측되었고, 충북지역 중 일부분이 연평균 기준보다 높을 것으로 예측되었다.

그림 10은 지시크리깅의 결과를 지도에 표시한 것으로 (a)는 예측된 확률값의 예측지도이고 (b)는 연평균 대기기준보다 크다고 예측된 지점을 지도에 표시한 것이다. 그 지점은 420개 지점으로 세 크리깅 방법 중 가장 적은 지점을 예측하였다. 그림 10의 (b)를 통하여 그 지점은



(a) UK 예측지도 (b) UK 대기기준 초과지점

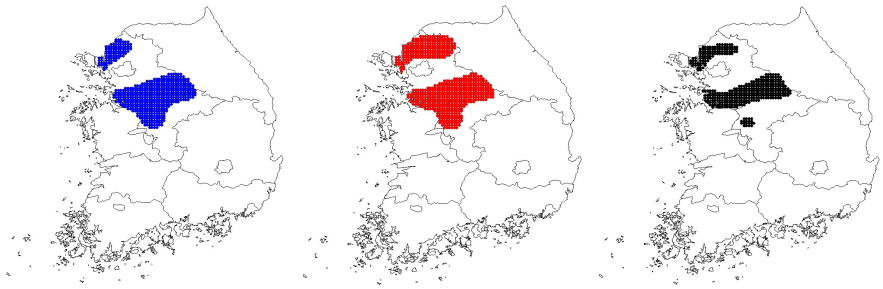
그림 9: 일반크리깅 예측지도



(a) IK 예측지도 (b) IK 대기기준 초과지점

그림 10: 지시크리깅 예측지도

서울을 둘러싼 경기 북부지역과 남부지역 그리고 충청북도 일부 지역에 해당하는 것을 알 수있다. 그림 11은 각 크리깅별 예측값이 대기기준을 넘는 지역의 예측지도이다. 그림 11의 (a)를 살펴보면 원자료의 경우 연평균 기준을 넘는 지역이 전국적으로 다양한 반면에 일반크리깅과 지시크리깅, 절단 가우시안 크리깅 그리고 절단 t-분포 크리깅 결과는 서울과 인접한 일부 지역으로 한정된 결과를 보였다.



(a) TTK 대기기준 초과지점 (b) UK 대기기준 초과지점 (c) IK 대기기준 초과지점

그림 11: 각 크리깅별 예측값 대기기준 초과지점

각 크리깅 방법을 실증 자료에 적용한 결과, 절단 가우시안 크리깅과 절단 t-분포 크리깅 결과가 동일하다는 결과를 얻었다. 이는 실증자료를 t-분포라 가정하고 자유도를 추정한 결과 37.302으로 추정되어, t-분포의 누적분포함수와 정규분포의 누적분포함수가 매우 근사한 값을 가지기 때문에 얻어진 결과이다. 이는 모의실험에서 t-분포의 자유도가 커질수록 두 방법이 동일해지는 결과와 같다. 또한 각 크리깅 방법의 예측값으로 대기기준 50$\mu g/m^3$보다 높을 지점을 지도에 표기하였는데, 4가지 방법 모두 서울을 둘러싼 경기북부, 경기남부, 충청북도 일부 지역인 것을 확인 할 수 있었다. 절단 가우시안 크리깅, 절단 t-분포 크리깅, 일반크리깅 지역은 경기 북부, 남부, 충청북도 북부 지역으로 예측하여 거의 유사한 결과를 얻었다. 지시크리깅의 경우 서울을 둘러싼 경기북부, 경기남부 지역과 충북 일부 지역으로 예측되었다.

제 5 장

결론

본 논문에서는 기존의 절단 가우시안 크리깅방법을 확장하여 t-분포 자료에도 적용할 수 있도록 t-분포의 누적분포함수를 이용하여 공간자료를 예측하는 방법인 절단 t-분포 크리깅에 대하여 연구하였다. 이를 위하여 t-분포와 표준정규분포를 따르는 자료를 생성하여 모의실험 연구를 진행하였다. 모의실험에서는 표준정규분포와 다양한 자유도를 따르는 t-분포에서 자료를 생성하여 자유도가 변화함에 따라 어떤 변화가 있는지 살펴보았다. 이를 위하여 일반크리깅, 지시크리깅, 절단 가우시안 크리깅, 절단 t-분포 크리깅 4가지 방법의 분할표를 작성하여 정분류율, 민감도, 특이도를 이용하여 비교하였다.

먼저 모의실험을 통하여 표준정규분포를 따르는 자료를 생성하여 이를 각각 절단 가우시안 크리깅, 절단 t-분포 크리깅, 일반크리깅, 지시크리깅을 적용하였다. 그 결과 절단 가우시안 크리깅과 절단크리깅의 경우 100개의 자료중 정분류의 경우에 100개의 자료 중 99개가 동일하다는 결과을 얻었다. 절단 t-분포 크리깅, 일반크리깅, 지시크리깅의 예측결과를 순위값을 가지고 비교하였을 때 절단 t-분포 크리깅의 결과가 정분류율, 민감도, 특이도 모든 척도에서 일순위 값을 가지는 자료가 가장 많다는 결과를 얻었다.

다음으로 각각 자유도를 따르는 t-분포에서 자료를 생성하여 모의실험을 한 결과, 자유도가 작아질수록 절단 t-분포 크리깅과 절단 가우

시안 크리깅 방법이 차이가 있음을 확인하였고, 자유도가 커짐에 따라 본 연구에서 제안한 절단 t-분포 크리깅과 절단 가우시안 크리깅 결과와 유사한 결과를 얻는 것을 확인 할 수 있었다.

모의실험 결과를 토대로 환경관측공단의 홈페이지 에어코리아에서 제공하는 2012년도 연평균 미세먼지 관측자료를 가지고 절단 가우시안 크리깅과, 절단 t-분포 크리깅, 일반크리깅 지시크리깅을 적용해 보았다. 그 결과, 절단 가우시안 크리깅과 절단 t-분포 크리깅 두 방법의 예측값의 차이가 없음을 확인하였다. 이는 자료를 t-분포라 가정하고 자유도를 추정하였을 때, 37.302으로 추정되어 표준정규분포와 t-분포가 근사함에 따라 생기는 결과이다. 따라서 실증자료를 통하여 절단 가우시안 크리깅과 절단 t-분포 크리깅을 비교한 결과, 모의실험에서 자유도가 커질수록 절단 가우시안 크리깅과 절단 t-분포 크리깅의 차이가 감소하는 것과 같은 결과를 얻었다.

실증자료를 절단 t-분포 크리깅과 일반크리깅에 적용하여 비교하였을 때, 지점의 수의 차이는 있지만 대기기준을 넘길 것이라고 예측된 지역은 경기 남부, 북부, 충북 북쪽 지역으로 동일한 지역을 예측하였다. 지시크리깅 예측결과 대기기준을 넘긴 지역은 경기북부, 경기남부 지역과 충북 일부 지역으로 예측되었다. 절단 t-분포 크리깅, 일반크리깅 결과와 충청북도 지역에서 다소 다른 지역을 예측하였다.

향후에는 좌우대칭분포 중 t-분포 뿐 아니라 라플라스분포에도 절단 크리깅 기법을 적용할 수 있는 방법에 대한 연구가 필요하다. 더 나아가 좌우대칭 분포가 아닌 한 쪽으로 치우친 분포인 카이제곱분포나 감마분포 등에도 적용할 수 있는 방법에 대해서 연구할 필요성이 있다. 또한, 연속형 자료를 이산형 자료로 변환하여 포아송분포에도 적용시킬 수 있는 방법에 대하여 연구하고자 한다.

참고 문헌

[1] 고혜지 (2014). 비등방성 공간자료의 연관성 연구, 성신여자대학교 대학원, 석사학위논문.

[2] 고혜지, 박만식 (2014). 예측성능 제고를 위한 범용크리깅과 지시크리깅의 결합, 한국자료분석학회, 제16권, 제4호, 1871-1884.

[3] 김동휘, 류동우, 이주형, 최인걸, 이우진 (2010). 인천 송도국제도시 지층분포추정을 위한 크리깅 방법의 비교연구, 한국지반공학회, 제26권, 제5호, 57-64.

[4] 김선우, 정애란, 이성덕 (2005). 공간자료에 대한 지리적 가중회귀 모형과 크리깅의 비교, 한국통계학회, 제18권, 제2 호, 271-280.

[5] 이상일 (2002). 지하수 및 토양 오염농도의 지구통계학적 추정-||. 적용, 대한환경공학회, 제24권, 제2 호, 303-307.

[6] 정승환, 박만식, 김기환 (2010). 풍속자료의 공간예측, 응용통계연구, 제23권, 제2호, 345-356.

[7] 조홍래, 정종철 (2006). 강우자료에 대한 공간보간 기법의 적용. 한국GIS학회, 제14권, 제1호, 29-41.

[8] 최승배, 문승호, 강창완, 조장식, 이정형 (2008). SAS/STAT 을 이용한 공간예측, 자유아카데미.

[9] 최종근 (2007). 지구통계학, 시그마프레스.

[10] 최지은, 박만식 (2013). 다양한 관측네트워크에서 얻은 공간자료들을 활용한 계층모형 구축, 응용통계연구, 제26권, 294-305.

[11] 허태영, 박만식, 엄진기, 오주삼 (2007). 최단경로 기반 교통량 공간 예측에 관한 연구, 한국통계학회, 제20권, 제3호, 459-473.

[12] 한국환경관측공단. http://www.airkorea.or.kr/

[13] C*á*ceres, A., Emery, X., and Riquelme, R. (2010). Truncated gaussian kriging as an alternative to indicator kriging. MININ 2010, IV International Conference on Mining Innovation, Santiago, Chile, in prep [4].

[14] Cressie, N. A. C. (1993). Statistics for Spatial Data, New York: John Wiley & Sons.

[15] Elbasiouny, H., Abowaly, M., Abu_Alkheir, A., Gad, A. (2014). Spatial variation of soil carbon and nitrogen pools by using ordinary kriging method in an area of north Nile Delta, Egypt, Catena, Vol. 113, 70–78.

[16] Gundogdu, K.S., Guney, I. (2007). Spatial analyses of groundwater levels using universal kriging, J. Earth Syst Sci., Vol. 116, No. 1, 49–55.

[17] Joseph, V.R., Hung, Y., Sudjianto, A. (2008). Blind kriging: a new method for developing metamodels, ASME Journal of Mechanical Design, Vol. 130, 3, 031102.1-031102.8.

[18] Schabenberger, O., Gotway., C.A. (2005). Statistical methods for spatial data analysis, Boca Raton: Champman & Hall/CRC.

[19] Selby, B., Kockelman, K.M. (2013). Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression, Journal of Transport Geography, Vol. 29, 24–32.

[20] Wang, X., Kockelman, K. M. (2009). Forecasting network data: Spatial interpolation of traffic counts using texas data, data. Transportation Research Record 2105, 100–108.

Abstract

Spatial Prediction of Truncated Kriging based on the t-distributions.

Bichna Choi
Department of Statistics
The Graduate School
Sungshin Women's University

Spatial statistics is a field of statistics, which analyzes the geo-statistical spatial data for the reliable prediction at unmeasured locations. This method is called kriging. In truncated Gaussian kriging, the continuous spatial data are divided into several categories and then the cumulative distribution function of standard normal distribution is used in estimating the probability of each category.

In this paper, we proposed a new method to estimate the probability of each category by utilizing t-distribution to the truncated kriging. In order to apply the cumulative distribution function of t-distribution, we estimated the degree of freedom in t-distribution via the maximum likelihood estimation method. In the simulation, we considered different degrees of freedom in t-distribution. We also examined the performances of our proposal compared

to common kiriging methods in terms of positive predictive value, sensitivity and specificity. Finally, we employed them to the annual fine particulate matter(PM_{10}) in 2012.

Keywords : t-distribution, Spatial correlation, Covariance model, Truncated Gaussian kriging, Truncated t-distribution kriging, Universal kriging, Indicator kriging