응용통계연구 제 17권 2호, 2004년, pp. 327-336

공간통계분석에서 이상점 수정방법의 효율성비교

이진희 [1] 신기일 [2]

요 약

이상점이 존재하는 공간자료(spatial data) 분석에서 이상점(outlier)의 영향력를 줄이기 위한 방법으로 로버스트 변이도(robust variogram)를 사용한다. 최근 이상점을 먼저 수정한 후 변이도를 추정하는 방법을 사용하면 더 좋은 분석결과를 얻을 수 있다는 것이 알려졌다. 본 논문에서는 이상점이 존재하는 공간자료 분석에서 Mugglestone 등(2000)이 제안한 이상점 수정법과 본 논문에서 제안한 새로운 이상점 수정법의 효율성을 비교하였다.

주요용어: 로버스트 변이도, 패치 이상점, 고립된 이상점, 중위수 수정법, 크리깅 수정법

1. 서론

통계적 자료분석에서 이상점이 존재할 경우 그 영향력을 줄이기 위한 많은 연구들이 진행되었다. Huber(1979)는 이상점이 있는 시계열 자료 분석에서 로버스트 시계열 평활법(robust time series smoothing)을 사용하여 이상점의 효과를 축소하였다. Denby와 Martin (1979)도 이상점이 있는 시계열 자료에서 자기회귀 모수의 로버스트 추정에 대하여 토론하였으며 Martin과 Yohai(1986)는 ARMA 모형에서 ACF를 추정할 때 이상점이 있을 경우 심각한 편의(bias)로 효율적이지 못한 추정량을 줄 수도 있음을 지적하였다. 공간자료 분석에서 변이도(variogram)의 추정은 크리깅(kriging) 가중 값을 결정하기 때문에 매우 중요한 분석 단계이다. 변이도의 추정은 일반적으로 Matherorn(1964)이 제안한 전통적인 방법을 이용하나 이 방법은 이상점에 상당히 민감하다고 알려져 있다. 공간통계 분석에서 이상점의 효과를 줄이기 위한 많은 로버스트 통계량이 연구되었다. Cressie와 Hawkins(1980)는 이상점이 있는 자료의 분석은 두 위치간의 차에 대한 절대치를 이용하여 변이도의 척도(scale)를 바꾸어주는 로버스트 통계량을 제안하였다. Genton(1998)은 전체 자료에 포함된 이상점의 수와 분산의 크기에 따라 기존의 로버스트 통계량과 자신이 제안한 로버스트 통계량의 효율성을 비교하였다. Lark(2000)는 토양 자료를 이용하여 기존에 발표된 여러가지 로버스트 통계량을 비교하였으며 이를 통하여 자료의 특징에 따라 통계량들이 차이가 있음을 보였다. 이러한 이유로 이상점을 먼저 수정한 후 분석을 실시하는 방법들이 연구되었다. Hawkins와 Cressie(1984)는 고립된 이상점이 존재하는 공간자료 분석에서 이상점을 수정한 후 분석하는 방법을 제안하였다. 최근 Mugglestone 등(2000)은 공간격자(spatial lattice) 자료에서

1) (411-764) 경기도 고양시 일산구 마두 1동 809번지 국립암센터 연구소 암등록 통계연구과, 연구원
E-mail: jhlee@ncc.re.kr
2) (449-791) 경기도 용인시 모현면 왕산리 산 79 한국외국어대학교 통계학과, 교수
E-mail: keyshin@stat.hufs.ac.kr

이상점이 존재할 경우 Nirel 등(1998)에 의해 제안된 척도에 대한 편의의 비율을 최소화하는 값을 이용하여 이상점을 탐지하였다. 또한 이들은 모형에 기초하지 않는 데이터 수정 방법인 중위수(median)를 이용하여 이상점을 수정한 후 ACF와 스펙트라를 추정하였으며 그 효율성도 비교 하였다. 본 논문에서는 Mugglestone 등(2000)의 방법을 지질통계자료(Geostatistic data)에 응용할 수 있는 새로운 방법을 제안하였다. 이는 격자자료(Lattice data)가 아닌 지질자료(Geostatistics)인 경우에는 이웃한 자료만을 이용하여 이상점을 수정하는 Mugglestone 등(2000)의 방법은 효율성이 떨어질 수 있기 때문이다. 따라서 본 논문에서는 공간통계분석에서 일반적으로 사용하는 크리깅 방법을 이용하여 이상점을 수정하는 방법을 제안하였다. 2장에서는 공간 가법 이상점 모형과 기존의 이상점 수정 방법들을 살펴보았고 3장에서는 본 논문에서 제안한 방법인 크리깅 수정법을 살펴보았다. 4장에서는 실제자료와 모의실험을 통하여 Mugglestone 등(2000)이 제안한 방법과 본 논문에서 제안한 방법을 비교하였다. 마지막으로 결론은 5장에 있다.

2. 공간 가법 이상점 모형

최근 Nilel 등(1998)과 Muggleston 등(2000)은 공간 격자자료에서 이상점이 존재할 경우 중위수를 이용하여 이상점을 수정하는 방법을 제안하였다. 이상점이 존재하는 자료 분석에서 일반적으로 사용되는 모형은 Hawkins와 Cressie(1984) 그리고 Martin과 Yohai(1986)가 사용했던 모형으로 다음과 같다.

$$Y_{u,v} = X_{u,v} + Z_{u,v}\nu_{u,v} \tag{2.1}$$

$\{Y_{u,v}\}$는 오염과정(contaminated process)이고 $\{X_{u,v}\}$는 오염되지 않은 과정 (uncontaminated process)이다. $\{\nu_{u,v}\}$는 $X_{u,v}$에 비하여 큰 분산과 상수평균을 갖는 오염과정이며, $\{Z_{u,v}\}$는 관측값이 이상점이면 "1" 을 관측값이 이상점이 아니면 "0"값을 갖는 지시함수이다. 또한 $X_{u,v}$, $Z_{u,v}$, $\nu_{u,v}$는 이차 정상성을 만족하고 서로 독립이라 가정하며, 만일 $Z_{u,v}$가 독립적으로 분포되어 있으면 고립된 이상점이고 공간적으로 상관되어 있으면 패치 이상점이 된다.

2.1. 모형에 기초한 이상점 탐지와 수정방법

Hawkins와 Cressie(1984)는 고립된 이상점이 있는 자료 분석에서 Huber(1979)의 로버스트 시계열 평활을 공간자료 형태로 바꾸었을 때 (2.2)식과 같이 표현됨을 보였고 예측하고자 하는 지점의 이웃 관측 값들을 사용하여 크리깅 가중값을 계산하였으며 얻어진 가중값을 이용하여 로버스트 예측 값을 얻고 이를 이상점 수정에 이용하였다.

$$\hat{Z}_{-j}(s_j) = \sum_{i=1, i\neq j}^{n} \lambda_{ji} Z(s_i) \tag{2.2}$$

여기서 $\hat{Z}_{-j}(s_j)$는 이상점으로 판명되어진 한 지점을 그 지점을 뺀 나머지 자료를 이용하여 예측한 값이고, λ_{ji}는 크리깅 가중값, $Z(s_i)$는 각 지점에서의 관측 값이다. 이 방법은 크리

깅 가중값을 이용한 가중된 중위수와 크리깅 분산을 사용하여 수정하는 방법이다. 모형에 기초하는 방법과 모형에 기초하지 않는 방법을 함께 이용한 이 방법은 고립된 이상점에 효과적으로 사용될 수 있는 방법이며 이상점이 패치로 존재할 경우도 중위수를 사용하였으므로 이상점의 영향력을 줄일 수 있다. 그러나 이상점이라 생각되는 하나의 자료만을 제거하는 방법을 사용하기 때문에 패치 이상점인 경우 이상점의 효과가 지속될 수도 있다.

2.2. 모형에 기초하지 않은 이상점 탐지와 수정방법

Mugglestone 등(2000)은 Nilel 등(1998)의 이론을 바탕으로 이상점을 탐지하는 방법과 수정방법을 제시하였다. 이 방법은 자료에 고립된 이상점이 존재할 경우뿐만 아니라 패치 이상점이 존재할 경우에도 적용할 수 있다. 먼저

$$\psi(y_{u,v}; M, g^0, g^1) = \begin{cases} y_{u,v}, & \text{if } |y_{u,v} - g^0(y_{u,v})| \le M \\ g^1(y_{u,v}), & \text{otherwise} \end{cases} \tag{2.3}$$

라 하자. 위 (2.3)식에서 $g^0(y_{u,v})$는 표본 중앙값으로, $|y_{u,v} - g^0(y_{u,v})| \le M$이면 $y_{u,v}$는 이상점이 아니고 $|y_{u,v} - g^0(y_{u,v})| > M$이면 $y_{u,v}$는 이상점으로 판단하게 된다. (2.3)식에서 이상점인지 아닌지를 판단하기 위한 M 값은 상대 편의를 최소로 하는 값으로 이를 구하기 위해서는 $x_{u,v}$의 분산을 구하여야 한다. 그러나 공간 가법 이상점 모형에서 실제로 관측한 자료는 $x_{u,v}$가 아닌 $y_{u,v}$이므로 정확한 $x_{u,v}$의 분산을 구하기가 어렵다. 이에 대한 대안으로 평균대신 이상점에 로버스트한 중앙값을 이용한 척도 추정량 (Hampel $p.105$)인 (2.4)식을 사용한다.

$$S_n = 1.483MAD(y_i) = 1.483med_i\{|y_i - med_j(y_j)|\} \tag{2.4}$$

여기서 상수 1.483은 불편추정치(unbias estimator)를 위한 상수이다. 또한 (2.3)식으로부터 탐지된 이상점들은 $g^1(y_{u,v})$로 대치하며, 대치 값인 $g^1(y_{u,v})$는 이상점이 고립되어 있을 때와 패치일 경우 그 계산 방법이 다르다. 만일 이상점이 고립된 이상점이면 이상점에 이웃하는 관측 값들의 중앙값이고, 패치 이상점이면 0도 90도 180도 270도로 각도를 바꾸어 가며 구한 중앙값들의 평균을 사용하게 된다. 본 논문에서는 앞으로 이 방법을 중위수 수정법(median corrected)이라 할 것이다. 전술한 바와 같이 지질자료(Geostatistic data)에 위의 방법을 그대로 적용하면 효율성이 떨어질 수 있다. 또한 2.1에서 언급한 Hawkins와 Cressie(1994) 방법 또한 패치 자료에서 발생할 수 있는 문제를 해결하지 못하고 있다. 이러한 기존의 방법을 개선한 크리깅 수정법에 관하여 다음절에 설명하였다.

3. 새로 제안된 방법

패치 이상점이 있는 자료분석에서 모형에 기초하는 방법으로 이상점을 수정한다면 그 수정 값들에 또 다른 이상점들이 영향을 주어 이상점의 영향이 지속될 수 있다. 이렇게 지속될 수 있는 이상점의 영향력을 줄이기 위해 본 논문에서는 이상점으로 탐지된 값들을 제외한 나머지 관측 값들만을 이상점 수정에 사용하였으며 앞으로는 이를 크리깅 수정법(kriging corrected)이라 부르겠다. 크리깅 수정법을 좀 더 자세히 설명하면 다음과 같다.

1단계 Mugglestone 등(2000)의 이상점 탐지방법을 이용하여 이상점을 탐지한다.

2단계 탐지된 이상점을 제외한 나머지 자료를 이용하여 변이도를 추정한다.

3단계 2단계에서 구한 변이도를 이용하여 이상점으로 탐지된 지점의 자료를 크리깅한다.

4단계 이상점으로 관측된 각 지점의 자료를 크리깅한 값들을 이용하여 수정한다.

5단계 수정된 자료를 이용하여 다시 변이도를 추정한다.

6단계 추정된 변이도를 이용하여 크리깅을 실시하고 또한 예측 오차를 구한다.

4. 실제 자료 분석과 모의실험

이 장에서는 2장에서 소개했던 중위수 수정법과 본 논문에서 제안한 크리깅 수정법의 효율성 비교를 위하여 실제 자료 분석을 실시하였다. 분석에 이용한 자료는 Mugglestone 등(2000)에서 사용한 패치 이상점 자료와 고립된 이상점만 있는 자료로 표 4.1과 표 4.2이다.

표 4.1: 고립

15	16	16	7
10	9	11	11
2	15	13	10
10	13	10	10
14	10	11	9
11	19	11	7
10	1	15	30

+

표 4.2: 패치

32	11	57	10
100	8	5	6
49	6	11	6
13	5	3	10
16	9	4	3
11	13	4	9
1	5	6	9

+

또한 자료분석과 모의 실험에서는 약한 의미의 정상성을 만족하는 모형인 구형모형, 지수모형, 가우시안모형의 3가지 모형 중 가장 알맞은 모형을 선택하여 사용하였으며 자세한 모형 식은 다음과 같다. 식 (4.1)은 구형모형(spherical model)이고 식 (4.2)는 지수모형(exponential model)이며 식 (4.3)은 가우시안모형(gaussian model)이다.

$$\gamma_s(h,\theta) = \begin{cases} 0 & h = 0 \\ \theta_0 + \theta_1\{\frac{3}{2}\frac{h}{\theta_2} - \frac{1}{2}(\frac{h}{\theta_2})^3\} & 0 < h \le \theta_2 \\ \theta_0 + \theta_1 & h \ge \theta_2 \end{cases} \tag{4.1}$$

$$\gamma_e(h,\theta) = \begin{cases} 0 & h = 0 \\ \theta_0 + \theta_1\{1 - \exp(-\frac{h}{\theta_2})\} & h > 0 \end{cases} \tag{4.2}$$

$$\gamma_g(h,\theta) = \begin{cases} 0 & h = 0 \\ \theta_0 + \theta_1\{1 - \exp(-\frac{h^2}{\theta_2^2})\} & h > 0, \end{cases} \tag{4.3}$$

여기서 θ_0는 뭉치(nugget)를 θ_1은 부분문턱(partial sill)을 그리고 범위(range)에 있어서는 각 모형에 따라 약간씩 다르게 나타난다. 일반적으로 θ_2를 이용하여 범위(range)를 구하는데 구형모형의 경우는 θ_2가 범위를 나타내지만 지수모형과 가우시안모형에 있어서는 θ_2에 상수 배를 곱하여 근사적으로 구하게 된다. 즉 지수 모형의 경우 $3\theta_2$가 실제 범위가 되고 가우시안모형의 경우 실제 범위는 $\sqrt{3}\theta_2$가 사용된다(Goovaerts, 1997). 위 모형들에서 사용되는 모수인 뭉치, 문턱, 범위를 좀 더 자세히 설명하면, 일반적으로 거리 $h = 0$이면 두 변수의 차의 분산인 변이도(variogram)의 추정치, 즉, $\hat{\gamma}(0) = 0$이다. 그러나 자료에 따라서는 h가 0에 가까워 지는데도 변이도 값이 0으로 수렴하지 않은 경우가 있는데 이러한 불연속의 크기를 뭉치(nugget)라 한다. 또한 공간 자료가 약한 의미의 정상성을 만족하면 변이도는 거리 h가 증가해도 무한히 증가하지 않고 특정한 값에 수렴하게 되는데 이때 가장 큰 값, 즉 $\max_h 2\hat{\gamma}(h)$를 문턱이라 하며 문턱에서 뭉치를 뺀값을 부분문턱이라 한다. 마지막으로 범위는 변이도의 문턱에 해당되는 거리 h를 말한다. 즉 범위는 상관관계가 있는 거리의 끝이 된다. 본 논문에서 모의실험을 위한 자료의 생성은 S-plus의 rfsim 함수를 사용하였으며, 이상점의 생성은 일반적으로 계통 추출로 선택된 값에 충분히 큰값을 더해 주거나 작은 값을 빼 주어 이상점을 생성하나 본 논문에서 이 방법을 사용하면 패치 이상점을 생성시키기가 어려워 먼저 자료를 크기순으로 나열(ordering) 한 후 가장 큰 값부터 차례로 6개의 지점을 선택하였고 가장 작은 값에서부터 차례로 4개 지점을 선택하여 선택된 값에 큰 값에는 5σ를 더하고 작은 값에서는 5σ를 빼 주어 이상점을 생성하였다. 분석은 S-plus의 SpatialStat 모듈을 이용하였고, 예측은 예측하는 지점의 관측 값만을 제외한 모든 관측치를 이용(leave one out)하였으며 실제값과 예측값을 이용하여 각 방법을 비교하였다. 본 논문에서 사용된 용어들은 편의상 다음과 같이 정의하였다. 본 논문에서는 이상점 수정방법의 효율성 비교를 위하여 이상점을 수정하지 않고 단지 로버스트 통계량만을 사용한 경우의 분석도 실시하였는데 이를 N-C(Non Corrected)라 하였다. 또한 Mugglestone 등의 중위수 수정법은 M-C(Median Corrected), 본 논문에서 제안한 크리깅 수정법은 K-C(Kriging-Corrected)로 표기하였다. 각 표에 제시한 예측오차(MSE)는 $\sigma_0^2 = \frac{1}{n}\sum_{i=1}^{n}(Z(s_i) - \hat{Z}(s_i))^2$를 이용하여 구하였으며 여기서 $\hat{Z}(s_i)$는 i 번째 지점을 제외하고 구한 예측 값이다.

4.1. 고립된 이상점

4장에서 소개한 Mugglestone 등의 두 자료중 고립된 이상점 자료는 표 4.1의 바구미 종자 수 자료이다. 이 자료에서 이상점으로 탐지된 각각의 지점을 위에서 언급한 두 가지 수정방법으로 수정한 후 추정한 변이도를 이용하여 예측한 후 얻은 결과이다.

고립된 이상점의 분석결과인 표 4.3을 살펴보면 N-C 방법과 M-C방법은 가우시안모형이, K-C 방법은 구형모형이 가장 적절한 모형으로 나타났으며 이상점을 수정하지 않은 경우는 이상점들의 영향으로 발생할 수 있는 과 추정 현상으로 굉장히 큰 범위와 문턱(sill)을

표 4.3: 고립된 이상점의 모수추정 결과와 예측오차

추정모형	방법	범위(range)	문턱(sill)	조각(nugget)	예측오차(MSE)
가우시안	N-C	226.29	305.137	0.23327	23.6260
가우시안	M-C	0.8529	0.0612	0.000	14.1117
구형	K-C	1.5945	0.00810	0.48704	14.0686

주고 있다. 새로운 분석 기법의 효과를 보기위해 구한 예측오차(MSE)를 살펴보면 그 차이가 작기는 하지만 K−C 방법이 M−C 방법에 비하여 더 좋은 결과를 주고 있다.

4.2. 패치 이상점

패치 이상점에 대한 자료는 표 4.2로 꽃가루 딱정벌레의 수를 조사한 자료이다. 이 경우에 있어서도 고립된 이상점과 같은 방법으로 분석하였으며 그 결과가 표 4.4에 나와 있다. 결과를 살펴보면 세 가지 경우 모두 가우시안 모형이 가장 적절한 모형으로 나왔으며 상당히 큰 모수추정 결과를 주고 있다. 예측오차는 고립된 이상점의 결과에서와 같이 K−C 방법이 M−C 방법에 비하여 더 좋은 결과를 준다

표 4.4: 패치 이상점의 모수추정 결과와 예측오차

추정모형	방법	범위(range)	문턱(sill)	조각(nugget)	예측오차(MSE)
가우시안	N-C	186.8222	1562.044	0.45172	26.8775
가우시안	M-C	98.3500	117.8334	0.21701	30.0864
가우시안	K-C	94.3202	114.3900	0.20873	29.8489

4.3. 모의 실험

본 논문에서 실시한 모의실험은 먼저 10×10 격자를 만든 후 식 (4.1)부터 식 (4.3)까지 공간 정상성을 만족하는 3가지 모형인 구형모형(spherical model)과 지수모형(exponential model) 그리고 가우시안모형(gaussian model) 으로부터 자료를 생성하여 분석을 실시하였다. 생성된 자료들의 실제 모수는 지수모형은 범위=2, 문턱=2, 조각=0, 가우시안모형은 범위=2, 문턱=2.5, 조각=0, 구형모형은 범위=1, 문턱=3, 조각=0이다. 또한 각 모형에서 생성된 자료에 약 10%의 이상점을 생성한 후 실제 자료분석에서 실시하였던 것과 같은 절차로 분석을 실시하였으며 각 모형에 대하여 55번 반복하였다. 효율성 비교를 위한 크리깅 오차는 55번 반복하여 얻어진 크리깅 오차의 평균을 이용하였다. 그러나 모수 추정에서 대부분 실제 모수 근처의 값들도 추정되었으나 실제모수와 아주 큰 차이를 보이는 몇개의 값들도 추정 되었다. 이렇게 실제 모수와 아주 큰 차이를 보이는 몇몇의 추정 값들로 인해 평균을 사용하기 어려워 평균대신 중위수를 사용하여 비교하였다.

표 4.5: 지수모형(범위=2, 문턱=2,조각=0)으로 생성한 자료

추정모형	방법	범위(range)	문턱(sill)	조각(nugget)
구형모형	N - C	3.3563	4.0389	0.3450
	M - C	3.4079	1.0980	0.0214
	K - C	3.1822	0.8894	0.0675
지수모형	N - C	2.2755	5.3186	0.0075
	M - C	2.8153	1.8504	0
	K - C	2.2703	1.2658	0
가우시안모형	N - C	1.6506	3.1026	0.8955
	M - C	1.6488	0.2204	0.2204
	K - C	1.5372	0.7164	0.2649

표 4.6: 가우시안모형(범위=2, 문턱=2.5, 조각=0)으로 생성한 자료

추정모형	방법	범위(range)	문턱(sill)	조각(nugget)
구형모형	N - C	3.5363	4.0389	0.3450
	M - C	2.5229	2.2836	0
	K - C	2.5379	2.7549	0
지수모형	N - C	1.3260	6.2291	0
	M - C	1.6447	2.9105	0
	K - C	1.6522	3.5107	0
가우시안모형	N - C	0.9913	5.0783	0
	M - C	1.0834	2.1983	0.0014
	K - C	1.1045	2.6287	0.0381

표 4.7: 구형모형(범위=1, 문턱=3, 조각=0)으로 생성한 자료

추정모형	방법	범위(range)	문턱(sill)	조각(nugget)
구형모형	N - C	1.9665	4.2640	0.0124
	M - C	2.2612	1.7126	0
	K - C	2.2310	1.7018	0.0418
지수모형	N - C	0.9319	5.0133	0
	M - C	1.2530	0	0
	K - C	1.2094	2.0045	0
가우시안모형	N - C	0.9939	3.8631	0.8214
	M - C	1.0749	1.4224	0.2935
	K - C	1.5372	0.7164	0.2649

표 4.8: 각 추정모형에 대한 예측오차

실제 모형	사용모형	예측오차(MSE)		
		N - C	M - C	K - C
구형모형	구형모형	1.5752	0.7452	0.5262
	지수모형	1.3924	1.7900	1.7705
	가우시안모형	1.7503	0.7943	0.7315
지수모형	구형모형	1.0597	0.3604	0.3576
	지수모형	1.0238	0.3552	0.3575
	가우시안모형	1.1998	0.4145	0.3846
가우시안 모형	구형 모형	0.2991	0.2984	0.2462
	지수모형	0.3012	0.3469	0.2681
	가우시안모형	0.3157	0.2943	0.1464

모의실험결과 모수추정에 있어서는 M–C와 K–C방법은 서로 차이를 보이지 않고 있으나 두 방법 모두 로버스트 추정법에 비하여 우수한 결과를 주고 있다. 예측 오차에서도 각 수정 방법에 따라 큰 차이를 보이지는 않으나 M–C방법에 비하여 K–C방법으로 수정한 경우가 대체적으로 더 작은 MSE를 주고 있다. 특히 가우시안 모형으로 자료를 생성하였을 경우 K–C방법이 M–C방법에 비하여 상당히 좋은 결과를 주고 있다.

5. 결론

자료 분석에서 이상점에 관한 연구는 매우 오래전부터 진행되어 왔다. 이는 이상점이 자료분석에 미치는 영향력의 중요성을 말해주고 있으며, 특히 자료의 수가 적은 공간통계학의 경우 더욱더 중요할 수밖에 없다. 위에서 살펴본 세 가지 이상점을 다루는 방법의 결과들은 많은 차이는 아니지만 지질자료(Geostatistic data)의 경우 본 논문에서 제안한 K–C 방법이 예측면에서 더 우수한 결과를 준다. 또한 이상점이 패치로 존재할 경우 중위수 수정법에 비하여 크리깅 수정법은 그 사용방법이 간편하다는 장점도 가진다.

참고문헌

Cressie, N. (1993). *Statistics for Spatial Data*, John Wiley and Sons, Inc.

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

Denby, L. and Martin, R. D. (1979). Robust estimation of the first-order autoregressive parameter, *Journal of the American Statistical Association*, **74**, 140-146.

Genton, M. C. (1998). Highly robust variogram estimation, *Mathematical Geology*, **30**, 213-221.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics*, Wiley, New York.

Hawkins D. M. Cressie, N. (1984). Robust kriging-A proposal, *Journal of the International Association of Mathematical Geologists*, **16**, 3-18.

Huber, P. J. (1979). Robust smoothing In Robustness in Statistics, R. L. Launer and G. N. Wilkinson, eds. Academic Press, New York, 33-47.

Lark, R. M. (2000). A comparison of some robust estimators of the variogram for use in soil survey, *European Journal of soil science*, **51**, 137-157.

Martin, R. D. and Yohai, V. J. (1986). Influence curves for time series, *Annals of Statistics*, **11**, 1608-1630

Matheron, G. (1962). Traite de Geostatistique appliquee, *Tome I. Memoires du Bureau de Recherches Geologiques et Minieres, No. Editions Technip*, Paris.

Mugglestone, M. A, Barnett, V., Nirel, R. and Murray, D. A. (2000). Modeling and analysing outliers in spatial lattice data, *Mathematical and computer modelling*, **32** 1-10.

Nirel, R., Moira A. and Mugglestone, M. A. (1998). Outlier-robust spectral estimation for spatial lattice processes, *Communications and Statistics: Theory and Methods*, **27**, 3095-3111.

Stephen, P. K., Silvia C. V., Tamre, P. C. and Alice, A. S. (1998). *S+SpatialStats User's Manual*, Springer.

[2003년 8월 접수, 2004년 3월 채택]

On the Efficiency of Outlier Cleaners in Spatial Data Analysis

Jin-Hee Lee[1] Key-Il Shin[2]

ABSTRACT

Many researchers have used the robust variogram to reduce the effect of outliers in spatial data analysis. Recently it is known that estimating the variogram after replacing outliers is more efficient. In this paper, we suggest a new data cleaner for geostatistic data analysis and compare the efficiency of outlier cleaners.

Keywords: Robust variogram, Pathcy outlier, Isolated outlier, Median corrected, Kriging corrected

1) Researcher, Cancer Registration & Biostatistics Branch, National Cancer Center Research Institute, 809 Madu-dong, Ilsan-gu Goyang-si, Gyeonggi-do, 411-764, Korea
E-mail: jhlee@nccre.re.kr

2) Professor, Department of Statistics, Hankuk University of Foreign Studies, San 79, Wangsan, Mohyun, Yongin, Kyonggi, 449-791, Korea
E-mail:keyshin@stat.hufs.ac.kr