공간통계분석에서 이상점 수정을 위한 방법비교

이 진희[1] 신 기일[2]

요약

공간 자료에서 이상점이 존재할 경우 변이도(Variogram)를 추정함에 있어 그 효과를 줄이기 위한 방법으로 로버스트(robust) 변이도를 이용한다. 그러나 이상점이 존재하는 자료분석에서 로버스트 변이도를 사용하기에 앞서 이상점을 수정한 자료를 사용하였을 경우 그 효율성 또한 좋다고 알려져 있다. 본 논문에서는 이상점이 존재하는 자료를 분석함에 있어 기존의 이상점 수정법 및 새로운 이상점 수정법의 효율성을 비교하였다.

주요용어 : 변이도, 패치 이상점, 고립된 이상점 , 중위수 수정법, kriging 수정법

1. 서론

공간통계분석에 있어 변이도의 추정은 kriging 가중 값을 결정하기 때문에 중요한 분석 단계이다. 일반적 자료에서 변이도의 추정은 전통적인 방법(Matherorn,1964)을 이용한다. 그러나 이상점이 있는 자료의 경우 전통적인 방법은 이상점에 상당히 민감하다고 알려져 있어 이에 대한 극복방안으로 이상점에 민감하지 않은 로버스트 추정량들이 연구되었다.(Cressie와 Hawkins, 1980; Dowd's, 1984; Rousseeuw와 Croux, 1992; 1993; Genton, 1998) 그러나 이러한 로버스트 통계량들은 효율성에 있어 자료의 특징에 따라 많은 차이를 보임도 알려져 있다.(Lark, 2000) 이러한 이유 등으로 Hawkins와 Cressie(1984)는 Huber(1976)의 로버스트 시계열 평활법(robust time series smoothing)을 공간통계에 적용하여 이상점이 존재하는 자료 분석에서 로버스트 추정법과 함께 이상점을 수정하는 방법을 제안하였다. 그러나 이러한 kriging을 이용한 수정법은 고립된(isolated) 이상점일 경우는 별 무리가 없으나 만일 이상점이 패치(patchy)로 존재할 경우 이상점의 효과가 계속 남아 있게 된다.

최근 Mugglestone등(2000)은 공간 격자 자료에서 이상점이 존재할 경우 Nirel등(1998)에 의해 제안된 척도에 대한 편의의 비율을 최소화하는 값을 이용하여 이상점을 탐지하였다. 또한 이들은 모형에 기초하지 않는 데이터 수정 방법을 이용하여 ACF와 스펙트라를 추정한 후 그 효율성을 비교분석 하였다. 본 논문에서는 Mugglestone등(2000)의 방법을 지질통계 데이터 분석으로 확장하여 응용하여 보았다. 그러나 이 경우는 모형에 기초하지 않은 방법으로 만일 자료에 적합한 모형이 존재할 경우 그 효율성이 떨어지게 되는데 이의 보완 방법으로 본 논문에서는 공간통계에서 일반적으로 사용하는 kriging 방법을 이용하여 이상점을 수정하였다. 통계적 자료분석에서 이상점 탐지를 위한 일반적인 방법은 히스토그램이나 상자그림 또는 잔차를 이용한다. 그러나 본 논문에서는 이러한 일반적인 탐지방법대신 Mugglestone등(2000)이 제안한 방법을 이용하여 이상점을 탐지하였다. 그리고 탐지된 이상점을 kriging을 이용하여 수정한 후 기존의 로버스트 통계량을 사용하였을 경우와 중위수 수정법을 이용하였을 경우 그리고 제안된 방법의 효율성을 kriging 오차를 통하여 확인하여 보았다. 2장에서는 먼저 공간 가법 이상점 모형과 기존의 이상점 수정 방법들을 살펴보았고 3장에서는 본 논문에서 제안한 방법인

1 경기도 용인시 모현면 왕산리 한국외국어 대학교 일반대학원 통계학과, 박사과정
2 경기도 용인시 모현면 왕산리 한국외국어 대학교 자연과학대학 정보 통계학과, 교수

kriging 수정법을 살펴보았다. 이에 대한 효율성 확인을 위하여 4장에서는 실제 자료 분석과 함께 모의실험을 실시하였고 결론은 5장에 있다.

2. 공간 가법 이상점 모형

통계적 분석에서 이상점이 존재할 경우 그 영향력을 줄이기 위하여 많은 로버스트 통계량에 대한 연구(Huber, 1981; Barnett과 Sewis. 1994; Hample등,1986; Simonoff등,1984)가 진행되었다. 공간통계에서는 이상점의 효과를 줄이기 위해 로버스트 추정량을 이용하여 변이도를 추정(Matherorn,1964; Cressie와 Hawkins,1980; Dowd's,1984; Rousseeuw와 Croux,1992; 1993; Genton,1998)하거나 이상점을 수정하여 그 효과를 줄이는 방법들(Hawkins와 Cressie,1984; Nilel등,1998; Muggleston, 2000)이 연구되었다.
본 논문에서 사용된 모형은 이상점이 존재할 경우 주로 사용되는 공간 가법 이상점 모형(Hawking와 Cressie, 1984; Martin과 Yohai, 1986; 1991)으로 다음과 같다.

$$Y_{u,v} = X_{u,v} + Z_{u,v}\nu_{u,v} \tag{1}$$

여기서 $\{Y_{u,v}\}$ 는 오염과정 이고 $\{X_{u,v}\}$는 오염되지 않은 과정이다. $\{\nu_{u,v}\}$는 X에 비교하여 큰 분산을 가지며 평균이 상수인 오염과정, $\{Z_{u,v}\}$는 관측값이 이상점이면 "1" 을 관측값이 이상점이 아니면 "0" 값을 갖는 지시함수이다. 또한 X, Z 그리고 ν는 이차 정상성을 만족하고 서로 독립이라 가정하며, Z 와 ν는 서로 독립이면 고립된 이상점이고 공간적으로 상관되어 있으면 패치 이상점이 된다.

2.1. Hawkins 와 Cressie(1984)의 이상점 탐지와 수정방법

Cressie와 Hawkins(1980)는 분석할 자료에 이상점이 존재할 경우 전통적인 방법을 사용하여 변이도를 추정한 후 예측할 경우 그 효율성이 떨어지므로 이 경우 이상점에 덜 민감한 방법으로 변이도를 추정할 것을 제안하였다. 또한 Hawkins 와 Cressie(1984)는 이 로버스트 추정치를 사용하여 예측을 하는데 아래와 같은 kriging 방법을 이용하여 이상점을 수정하여 분석을 실시한다면 더 좋은 예측결과를 줄 수 있다고 제안하였다.

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

여기서 $\hat{Z}_{-j}(s_j)$는 이상점으로 판명되어진 한 지점을 뺀 나머지 자료를 이용하여 예측된 값이고, λ_{ij}는 가중값, $Z(s_i)$는 각 지역에서의 관측 값이다. 그러나 이 방법은 이상점이 패치로 존재할 경우 이상점에 가장 가까이 있는 관측 값들에 큰 가중치를 주게 되는 모형의 특징으로 인하여 예측결과가 예측지점의 바로 이웃에 존재하는 다른 이상점의 영향을 많이 받게되어 이상점의 효과를 줄이기에 부적절하게 된다. 모형에 기초할 경우 발생하는 이러한 단점을 보완한 방법이 모형에 기초하지 않는 Mugglestone등의 중앙값을 이용한 이상점 수정 방법이다.

2.2 중앙값 수정방법

Nilel등(1998)과 Mugglestone등(2000)의 이상점을 탐지하는 방법과 수정방법은 다음과 같다.

$$\psi(y_{u,v}; M, g^0, g^1) = \begin{cases} y_{u,v.} & \text{if } |y_{u,v} - g^0(y_{u,v})| \le M, \\ g^1(y_{u,v}), & otherwise. \end{cases} \tag{2}$$

위 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)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 우리가 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 (3)식을 사용한다.

$$S_n = 1.483\, MAD(y_i) = 1.483\, med_i\{|y_i - med_j(y_j)|\} \quad (3)$$

여기서 상수 1.438은 불편추정량을 위한 상수이다. (1)식으로부터 탐지된 이상점들에 대하여는 $g^1(y_{u,v})$로 대치시키게 된다. 대치 값인 $g^1(y_{u,v})$는 만일 이상점이 고립된 이상점이면 이상점이라고 판단된 자료들에 이웃하는 관측 값들의 중앙값이고, 패치 이상점이면 0도 90도 180도 270도로 각도를 바꾸어 가며 구한 중앙값들의 평균을 사용하게 된다.
그러나 위 방법은 모형에 기초하지 않은 방법이기 때문에 만일 자료에 알맞은 모형이 존재할 경우 모형을 이용하여 추정한 결과를 이용하여 수정된 값들에 비하여 그 효율성이 떨어지게 된다. 이러한 모형에 기초하지 않은 방법을 사용하였을 경우의 보완 방법인 kriging 수정법을 다음에서 살펴볼 것이다,

3. 새로 제안된 방법

Hawkins 와 Cressie(1984)의 kriging 방법으로 이상점을 수정할 경우 이상점들에 이웃하는 관측 값들도 이상점일 경우 그 수정 값들에 또 다른 이상점이 영향을 주어 이상점의 영향이 지속되게 된다. 이러한 지속되는 이상점의 영향력을 줄이기 위해 본 논문에서는 이상점으로 탐지된 값들을 제외한 나머지 관측치들만을 이용하여 모수를 추정한 후 이 추정값을 이용하여 제외된 이상점들을 예측하여 수정하였다. 여기서 이상점 탐지 방법은 Muggleston(2000)등의 방법을 이용하였다.

3.1 알고리듬

본 논문에서 사용된 이상점 수정 방법은 이상점을 수정함에 있어 먼저 모형에 기초하지 않은 방법으로 이상점을 수정한 자료를 이용하여 가장 적절한 모형을 찾아낸 다음 이 모형을 이용하여 다시 한번 이상점을 수정하는 방법으로 다음과 같다.

1단계 : Muggleston(2000)등의 이상점 탐지방법을 이용하여 이상점을 탐지한다.
2단계 : 탐지된 이상점을 제외한 나머지 자료를 이용하여 변이도를 추정한다.
3단계 : 2단계에서 구한 변이도를 이용하여 이상점으로 탐지된 지점을 kriging한다..
4단계 : kriging된 값들을 이용하여 이상점을 수정한다.
5단계 : 수정된 자료를 이용하여 다시 변이도를 추정한다.
6단계 : 추정된 변이도를 이용하여 kriging을 실시한 후 예측 오차를 구한다.

위 단계들을 이용하여 수정된 방법의 효율성을 살펴보기 위하여 다음 장에서는 실제 자료분석과 모의실험을 실시하여 보았다.

4. 실제 자료 분석과 모의실험

이 장에서는 2장에서 소개했던 이상점 수정 방법과 본 논문에서 제안한 kriging 수정법의 효

율성 비교를 위하여 실제 자료분석을 통하여 kriging 오차를 비교하여 보았고, 모의실험을 통하여 실제 모수에 잘 적합하는 방법과 효율성도 함께 살펴보았다. 본 논문에서는 자료 분석을 함에 있어 패치 이상점인 경우와 고립된 이상점만 존재하는 데이터 두 경우 모두 분석하여 보았다. 자료분석은 S-Plus 프로시져를 이용하였다. 얻어진 자료는 패치 이상점이 있는 자료의 경우 꽃가루 딱정벌레의 수 자료(Muggleston등,2000)이고, 고립된 이상점만 존재하는 경우는 바구미 종자 수에 대한 자료(Muggleston등,2000)이다.

4.1 패치 이상점자료 분석

이 절에서 사용된 패치 이상점을 위한 자료는 Muggleston등(2000)이 사용한 자료로 7×4 격자에서 꽃가루 딱정벌레의 수를 조사한 자료로 이 자료를 이용하여 각각의 이상점 수정방법으로 이상점을 수정한 후 추정한 변이도를 이용하여 예측을 실시하였다. 예측방법은 예측하는 지점의 관측치만 제외한 모든 관측치를 이용하였다.

모형	이름	range	sill	nugget	MSE
gaussian	N-C	186.8222029	1562.0440232	0.4517292	26.87753
	M-C	98.3499858	117.8334959	0.2170121	30.08649
	K-C	94.3202036	114.3900389	0.2087342	29.84894

<표 1> 패치 이상점의 모수추정 결과

먼저 본 논문에서 사용된 용어를 살펴보면 N-C(Non Corrected)는 이상점을 수정하지 않고 로버스트 방법을 이용하여 모수를 추정한 경우이고 K-C(Kriging Corrected)는 본 논문에서 제안한 방법인 kriging 수정법, 그리고 M-C(Median Corrected)는 중위수 수정법을 의미한다. 각각의 방법을 이용하여 구한 변이도를 살펴보면 이상점을 수정하지 않은 경우의 범위와 sill등이 큰 값을 가짐을 알 수 있으면 예측오차를 살펴보면 M-C방법에 비하여 K-C방법으로 수정한 경우 더 좋은 결과를 줌을 알 수 있다.

4.2 고립된 이상점자료 분석

고립된 이상점에 대한 자료는 7×4격자에서 얻어진 바구미 종자 수 자료이다. 이 경우에 있어서도 패치 이상점과 같은 방법으로 분석을 실시하였다.

모형	이름	range	sill	nugget	MSE
gaussian	N-C	226.29256	305.13767	0.2332786	23.62602
gaussian	M-C	0.8529093	0.0612102	0.0000000	14.11177
spherical	K-C	1.5945492	0.0081048	0.4870424	14.06865

<표 2> 고립된 이상점의 모수추정 결과와 예측오차

이 경우도 패치 이상점과 같은 모수추정 결과를 얻었으며 예측오차 또한 K-C방법이 M-C방법에 비하여 더 좋음을 알 수 있다.

4.3 모의실험 결과

Exponential 모형(range=2.5, sill=3, nugget=0)으로 생성된 자료이다. 자료에 대한

모형 적합과 모수추정 결과를 살펴보면 먼저 exponential 모형으로 적합하였을 경우 이상점을 수정하지 않은 경우가 실제 모수와 가장 큰 차이를 보이고 있으며 본 논문에서 제안한 K-C 방법이 가장 근접함을 알 수 있다.

모형	이름	range	sill	nugget
spherical	N-C	4.4230516	3.6867739	0.4799875
	K-C	5.0945533	1.6007912	0.1449169
	M-C	5.1551869	1.5915047	0.1562869
exponential	N-C	1.9722350	4.5181760	0.0000000
	K-C	2.6169430	2.0068950	0.0000000
	M-C	2.6242122	2.0028469	0.0028478
gaussian	N-C	2.1544810	3.0783480	1.0816550
	K-C	2.4920670	1.3615190	0.386800
	M-C	2.5303316	1.3504865	0.4007223

<표 3> 모의실험에서 각 모형에 대한 모수추정 결과

이름	MSE		
	spherical	exponential	gaussian
N-C	0.7701639	0.8073376	1.0097480
M-C	0.4357031	0.4238518	0.5547025
K-C	0.4351512	0.4272464	0.5546402

<표 4> 모의실험에서 각 모형에 따른 예측오차

5. 결론

통계적 자료를 분석함에 있어 이상점에 관한 연구가 많다는 것은 그 만큼 이상점이 자료분석에 미치는 영향이 크다고 해석할 수 도 있다. 특히 자료의 수가 적을 경우 이는 더욱더 중요할 수밖에 없다. 위에서 살펴본 세 가지 이상점을 다루는 방법을 살펴본 결과 모수 추정에 있어 K-C방법이 가장 근접한 추정 값을 주고 있으며 예측 오차도 K-C 방법이 비교적 좋은 결과를 준다.

참고문헌

1. Beran, J. (1994), On a class of M-estimators for Gaussian long-memory models, Biometrika, 81, 755-766.
2. Chang, I., Tiao, G. C. and Chen, C.(1988), Estimation of time series parameters in the presence of outliers, Technometrics, 30, 193-204
3. Cressie, N. (1993), Statistics for spatial data, John Wiley & Sons, Inc.
4. Cressie, N. and Hawkins D. M.(1980), Robust estimation of the variogram, I.

Mathematical Geology, Vol. 12, No. 2, 115-125.
5. Hawkins D. M.. Cressie, N.(1980), Robust kriging-A proposal, Journal of the International Association of Mathematical Geologists 16, 3-18.3. 76.
6. Genton, M. C, (1998), Highly robust variogram estimation, Mathematical Geology, Vol. 30, No. 2, 213-221.
7. Huber, P. J.(1981) Robust statistics, Wiley, New York.
8.. Lark R. M.(2000) A comparison of some robust estimators of the variogram for use in soil survey, European Journal of soil science, V. 51, 137-157.
9. Martin, R. D. and Yohai, V. J. (1986). Influence curves for time series, Ann. of Statist., 11, 1608-1630
10. Matheron, G. (1962). Traite de Geostatistique appliquee, Tome I. Memoires du Bureau de Recherches Geologiques et Minieres, No. 14. Editions Technip, Paris.
11. 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
12. Nirel R., Moira A., Mugglestone, M. A.(1998). Outlier-Robust Spectral Estimation for Spatial Lattice Processes, Commun. and Statistics-Theory and Methodolagy, 27(12), 3095-3111.
13. Simonoff, J. S., A. (1984) Comparison of Robust Methods and Detection of Outliers Techniques When Estimating a Location parameter, Communications in Statistics-Theory and Methods, 13, 813-842.
14. Stephen, P. Kaluzny, Silvia C. Vega, Tamre P. Cardoso and Alice, A. Shelly(1998), S+SpatialStats User's Manual, Springer
15. Barnett V. and T. Sewis. (1994) Outliers in statistical data, John Wiley, New York 3rd edition.