응용통계연구 제 14권 2호, 2001년, pp. 415-427

주성분점수를 이용한 이변량 공간자료에 대한 감도분석

최승배 [1] 강창완 [2]

요 약

공간통계학에서는 다변량 공간자료에 대한 예측방법으로서 코크리깅 기법을 이용한다. 본 논문에서는 코크리깅을 위한 첫 번째 단계인 교차베리오그램의 추정에 대한 감도분석 대신에 일반통계학적 측면에서 주성분점수를 이용한 감도분석방법을 제안한다. 변수가 2개인 경우, 교차베리오그램에 대한 감도분석의 결과와 제안된 주성분점수를 이용한 감도분석의 결과를 비교해 본다. 모의실험을 통하여 제안한 방법의 타당성을 검증하고, 실제 자료를 이용한 사례분석의 결과로써 재확인해 본다.

주요용어: 베리오그램, 교차베리오그램, 코크리깅, 감도분석.

1. 서 론

공간자료는 확률과정(stochastic process)의 실현치로서 $\{Z(s) : s \in D \subset R^d\}$와 같이 표현된다. 여기서 s는 D의 임의의 위치를 나타내고, R^d는 d차원의 유클리드 공간이다. 통상적으로 공간차원인 d는 1,2,3이고, $Z(s)$는 정상성(stationarity)을 만족한다고 가정한다. 이러한 공간자료는 $(z_i, s_i) : i = 1, \cdots, n$과 같이 표현 가능하다. 여기서 z_i는 위치 s_i에서 얻어진 관측치이다. 통상적으로 공간상에서 얻어진 관측치들은 서로 멀리 떨어져 있을수록 낮은 상관을, 가까이 있을수록 높은 상관을 갖는다. 이러한 공간자료는 1) 베리오그램(variogram)의 추정, 2) 추정된 베리오그램을 근거로 한 모델 적합, 3) 적합된 베리오그램 모델을 이용한 예측식의 추정, 의 3가지 단계로 분석된다. Choi and Tanaka(2000)는 이러한 3가지 단계 각각에 대해서 영향을 미치는 관측치를 찾기 위한 영향함수(influence function)를 유도하고, 그에 대한 유용성을 보였다.

베리오그램은 공간분석에서 자주 이용되는 공간 종속성의 척도이고, 베리오그램을 구하는 것은 공간분석을 위한 초기 작업 단계이다. 베리오그램은 확률과정이 본질적인 정상(intrinsically stationarity)이라는 가정 하에 다음과 같이 정의된다.

$$\begin{aligned} 2\gamma(d) &= E(Z(s+d) - Z(s))^2 \\ &= Var(Z(s+d) - Z(s)), \end{aligned} \tag{1.1}$$

1) (614-714) 부산광역시 부산진구 가야동 산 24번지, 동의대학교 통계상담실, 책임연구원
E-mail: statcst@hyomin.dongeui.ac.kr
2) (614-714) 부산시 부산진구 가야동 산24, 동의대학교 수학컴퓨터통계학부, 조교수
E-mail: cwkang@hyomin.dongeui.ac.kr

여기에서 $\gamma(d)$는 세미베리오그램(semivariogram)이다. 세미베리오그램의 추정량은 다음에 의해서 계산된다.

$$\hat{\gamma}(d) = \frac{1}{2N_d} \sum_{N(d)} (z(s_i) - z(s_j))^2, \tag{1.2}$$

여기에서 $N(d)$는 유클리드 거리 $|s_i - s_j| = d$를 갖는 전체 쌍의 집합, N_d는 $N(d)$에 속하는 쌍의 수, $z(s_i)$는 공간위치 s_i에서의 관측치이다.

일반적으로 임의의 위치에서 하나 이상의 측정치가 얻어진 자료는 다변량 공간자료로서 간주된다. Choi et al.(2000)은, 다변량 공간자료의 경우(변수가 2개인 경우), 코크리깅(cokriging)을 위한 첫 번째 단계인 교차베리오그램(cross-variogram)의 추정단계에 대한 영향함수를 유도하고, 실제 예를 통하여 유용성을 보였다. 코크리깅은 크리깅(kriging)의 자연스러운 확장으로서 임의의 한 위치에서의 수치를 예측하는데 있어서 한 개 이상의 공변량을 사용하는 방법이다. 따라서 공간통계학에서는 미지의 위치의 값을 예측하는데 있어서 코크리깅이 크리깅보다 예측력이 높은 것으로 알려져 있다. 교차베리오그램 $\gamma_{ij}(d)$는, 임의의 $s, s+d \in D$와 $i, j = 1, , p$에 대해서, 확률과정이 결합적으로 본질적인 정상성을 만족한다는 가정 하에서 다음을 만족한다.

$$\begin{aligned} E\left[Z_i(s+d) - Z_i(s)\right] &= 0, \\ Cov\left[(Z_i(s+d) - Z_i(s))(Z_j(s+d) - Z_j(s))\right] &= 2\gamma_{ij}(d). \end{aligned} \tag{1.3}$$

따라서 교차베리오그램은 다음과 같이 정의되어진다.

$$\gamma_{ij}(d) = E\left[(Z_i(s+d) - Z_i(s))(Z_j(s+d) - Z_j(s))\right]/2, \quad i \neq j, \tag{1.4}$$

여기서 $p = 2$인 경우(주변수인 V와 보조변수인 U)를 고려하면, 교차베리오그램의 추정량은 다음에 의해서 계산된다.

$$\hat{\gamma}_{12}(d) = \hat{\gamma}_{21}(d) = \frac{1}{2N_d} \sum_{N(d)} (v(s_i) - v(s_j))(u(s_i) - u(s_j)). \tag{1.5}$$

본 논문은 Choi et al.(2000)에서 사용한 공간통계학적인 분석방법인 교차베리오그램을 적용하지 않고, 일반통계학적인 접근 방법으로서 주성분점수(PCS: Principal Component Score)를 이용한 이변량 공간자료에 대한 감도분석을 제안한다. 그리고 이변량 공간자료의 경우, 시뮬레이션과 실제 적용사례를 통하여 제안한 방법의 타당성을 입증한다. 감도분석은 개체를 한 개씩 빼는 경우에 있어서 가중치의 변화를 조사해 보는 것에 의해서 각 개체에 대한 영향력을 평가하는 것이다. 이러한 영향력 평가를 위한 방법으로서 본 논문에서는 유도된 영향함수의 적절한 놈(norm)이 고려된다. 또한 편의상 세미베리오그램은 거리 d 만의 함수이고, 공간과정(spatial process)은 등방(isotropy: 방향에 의존하지 않는다)이라고 가정한다. 그리고 혼동이 없는 한 세미베리오그램을 교차베리오그램과 구별하기 위해서

자기베리오그램의 용어를 사용하고, 표기의 간략화를 위해서 교차베리오그램에 대한 감도분석법을 "C방법", PCS를 사용한 감도분석법을 "P방법"이라는 용어를 사용한다. 그리고 분석을 위해서 공간분석을 위한 모듈인 S+SpatialStat를 가지고 있는 S-Plus를 사용한다.

본 논문은 다음과 같이 구성되어 진다. 자기베리오그램과 교차베리오그램의 추정단계에서 감도분석을 위한 영향함수가 2장에서 유도된다. 3장에서는 일반통계학적 관점인 PCS에 의한 감도분석 방법을 소개하고 모의실험을 통하여 제안된 방법의 타당성을 검증한다. 4장 에서는 제안한 PCS를 이용한 감도분석결과와 공간통계학적 관점인 교차베리오그램에 의한 감도분석 방법을 실제 적용사례를 통하여 비교해 본다. 결론으로 마지막장을 맺는다.

2. C방법에 의한 영향력 평가

2.1. 표본 자기베리오그램 추정에 대한 영향함수

표본 자기베리오그램에 대한 영향함수를 유도하기 위하여 식 (1.2)에 다음과 같이 개체에 대해서 가중치를 도입한다.

$$v_\epsilon(d) = \frac{1}{2}\sum_{N(d)} w_{ij}\{z(s_i) - z(s_j)\}^2. \tag{2.1}$$

$N(d, i^*)$를 유클리드 거리 d를 갖는 전체의 쌍들 중에서 i^*를 포함하는 쌍의 집합이고, w_{ij}는 $\sum w_{ij} = 1$의 조건을 만족하는 쌍 (i, j)에 부여되는 가중치로서 다음과 같이 표현된다.

$$w_{ij} = \begin{cases} \tilde{w}_{ij}/\sum_{N(d)} \tilde{w}_{ij}, & (i,j) \in N(d, i^*) \\ 1/\sum_{N(d)} \tilde{w}_{ij}, & (i,j) \notin N(d, i^*), \end{cases} \tag{2.2}$$

여기서, $\tilde{w}_{ij}$는 섭동(perturbation)을 도입한 가중치로서 $(i, j) \in N(d, i^*)$에서는 $1 + \epsilon$이고, $(i, j) \notin N(d, i^*)$에서는 1이다. 비 섭동상태에서는 $1/|N(d)|$이다. 이 때, 섭동 후의 자기베리오그램은 다음과 같이 전개될 수 있다.

$$v_\epsilon(d) = v(d) + \epsilon v^{(1)}(d; s_{i^*}) + \cdots, \tag{2.3}$$

여기서, $\epsilon' = \epsilon(N_{d,i^*}/N_d)$이고, $v^{(1)}(d; s_{i^*}) = \left[\sum_{N(d,i^*)}\{z(s_i) - z(s_j)\}^2/(2N_{d,i^*}) - v(d)\right]$ 이다. 여기서 N_{d,i^*}는 $N(d, i^*)$의 쌍들의 수를 나타낸다. 식 (2.3)의 $v^{(1)}(d; s_{i^*})$는 $v(d)$의 $\epsilon = 0$에 대한 1차 미분계수(경험영향함수)이고, $\epsilon' = -N_{d,i^*}/N_d$ (즉, $\epsilon = -1$)은 i^*번째의 개체를 제거하는 것을 의미한다. $v^{(1)}(d; s_{i^*})$에서 $\sum_{N(d,i^*)}\{z(s_i) - z(s_j)\}^2/2N_{d,i^*}$는 i^*에 관계하는 쌍에 근거한 차의 분산이고, $v(d)$는 전체 쌍에 근거한 차의 분산이다. 따라서, i^*에 관계하는 부분으로부터 계산된 자기베리오그램이 전체로부터 구해진 자기베리오그램과 큰 차이가 있다면, 영향함수의 절대치는 크게되고, 개체의 영향이 크다고 할 수 있다.

2.2. 표본 교차베리오그램 추정에 대한 영향함수

일변량 공간분석에서의 베리오그램은 단 하나의 변수의 표본 수치들에 의해서 추정된다. 2개 이상의 변수에 대한 표본 수치들에 의해서 추정되는 경우는 다변량 공간자료분석을 고려해야 한다. 다변량 공간자료분석은 일변량 경우의 자연스러운 확장으로서 앞에서 언급된 3가지 분석단계를 거친다. 이 장에서는 다변량 공간자료분석에 대한 첫 번째 단계로서 교차베리오그램의 추정에 있어서의 영향함수를 유도한다. 본 논문에서는 이변량의 경우에 대해서만 고려되어 지고, 두 변수들 중 V는 주변수이고, U는 보조변수라고 가정한다. 교차베리오그램에 대한 영향함수를 유도하기 위해서 식 (1.5)에 다음과 같이 가중치를 도입한다.

$$v_\epsilon^{UV}(d) = \frac{1}{2}\sum_{N(d)} w_{ij}\{(v(s_i)-v(s_j))(u(s_i)-u(s_j))\}. \tag{2.4}$$

w_{ij}는 2.1절에서 정의된 가중치이고, 섭동상태의 교차베리오그램은 다음과 같다.

$$v_\epsilon^{UV}(d) = \sum_{N(d)}\{\tilde{w}_{ij}/\sum_{N(d)}\tilde{w}_{ij}\}\{(v(s_i)-v(s_j))(u(s_i)-u(s_j))\}/2. \tag{2.5}$$

또한 식 (2.5)는 테일러 급수에 의해서 다음과 같이 표현된다.

$$\begin{aligned} v_\epsilon^{UV}(d) = v_{UV}(d) &+ \epsilon\frac{N_{d,i^*}}{N_d}[\sum_{N(d,i^*)}\frac{1}{2N_{d,i^*}}(v(s_i)-v(s_j))(u(s_i)-u(s_j)) - v_{UV}(d)] \\ &-\epsilon^2(\frac{N_{d,i^*}}{N_d})^2[\sum_{N(d,i^*)}\frac{1}{2N_{d,i^*}}\{(v(s_i)-v(s_j))(u(s_i)-u(s_j))\} - v_{UV}(d)] + \cdots . \end{aligned} \tag{2.6}$$

교차베리오그램에 대한 영향함수는 다음과 같이 정의된다.

$$v_{UV}^{(1)}(d;s_i^*) = \frac{N_{d,i^*}}{N_d}[\sum_{N(d,i^*)}\frac{1}{2N_{d,i^*}}\{(v(s_i)-v(s_j))(u(s_i)-u(s_j))\} - v_{UV}(d)]. \tag{2.7}$$

식 (2.7)에서 $v_{UV}(d)$는 모든 쌍에 근거한 교차베리오그램이고,

$$\sum_{N(d,i^*)}\frac{1}{2N_{d,i^*}}\{(v(s_i)-v(s_j))(u(s_i)-u(s_j))\}$$

은 i^*번째에 관계하는 쌍에 근거한 교차베리오그램이다.

3. P방법에 의한 영향력 평가

이 절에서는 2.2절의 코크리깅을 위한 첫 번째 단계인 교차베리오그램 추정에서의 감도분석을 주성분분석에 의해 구현 가능함을 제안한다. 주성분분석의 결과를 이용한 감도분

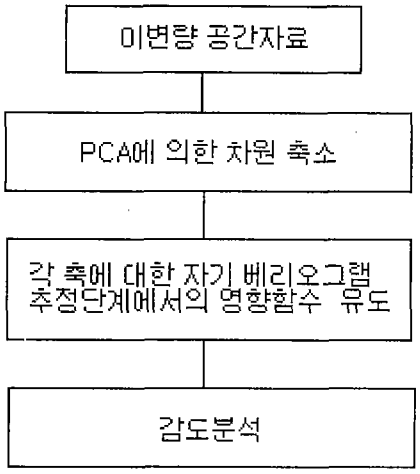

그림 3.1: PCA에 의한 감도분석 흐름도

석의 절차는 그림 3.1에 주어져 있다. PCA에 의한 영향력 평가방법의 제 1단계는 주어진 이변량 공간자료에 대해서 주성분분석을 행하고 설명력 있는 축까지 주성분 점수를 얻는다. 제 2단계는 1단계에서 얻어진 각 축의 주성분점수를 일변량 공간자료로서 상정해서 자기베리오그램 추정에서의 영향함수를 구하고, 이를 이용하여 감도분석을 행한다.

이 절에서 제안한 방법의 타당성 검토를 위하여 다음과 같이 모의실험을 하였다. 이변량 공간자료(주 변수와 보조변수)를 생성하기 위하여 공간자료분석에서 일반적으로 널리 사용되는 세 가지 모델, 구형모델(Spherical Model), 가우시안 모델(Gaussian Model), 그리고 지수모델(Exponential Model), 을 고려하였다.

$$\gamma_1(d) = \begin{cases} \theta_1 + \theta_2[(3/2)(d/\theta_3) - (1/2)(d/\theta_3)^3], & d <= \theta_3 \\ \theta_1 + \theta_2, & d > \theta_3 \end{cases} \tag{3.1}$$

$$\gamma_2(d) = \begin{cases} 0, & d = 0 \\ \theta_1 + \theta_2[1 - exp(-d^2/\theta_3^2)], & d \neq 0 \end{cases} \tag{3.2}$$

$$\gamma_3(d) = \begin{cases} 0, & d = 0 \\ \theta_1 + \theta_2[1 - exp(-d/\theta_3)], & d \neq 0 \end{cases} \tag{3.3}$$

여기서 각각의 모델의 모수로서 Nugget Effect θ_1은 1, Sill $\theta_1 + \theta_2$는 7, 그리고 Range θ_3는 2를 사용하였다. 모의실험은 다음과 같은 순서로 진행되었다.

(1) 지리통계학적 자료를 생성하기 위해서, 0에서 10을 횡과 종으로 10×10의 격자(Lattice)로 구성된 100개의 위치좌표를 발생시킨다.

(2) 단계 (1)의 격자내의 또 다른 격자를 25개 발생시킨다. 이것은 단계 (1)에서 얻어진 자료들과 다른 공분산 구조를 갖는 오염된 자료를 얻기 위한 것이다. 분석을 위한 자료는 총 125개이다.

(3) 앞에서 소개한 세 가지 분포를 따르는 자료를 단계 (2)에서 발생시킨 125개의 격자 위치에 대응하도록 모의실험을 통하여 각 분포를 따르는 지리 통계학적 자료를 생성시킨다. 여기서 첫 번째 개체부터 100번째 개체까지는 모수 값이 $\theta_1 = 1$, $\theta_2 = 6$, $\theta_3 = 2$인 공분산을 갖는 각 모델로부터 얻어진 자료이고, 101번째부터 125번째 자료는 임의로 부여한 모수 값 $\theta_1 = 50$, $\theta_2 = 100$, $\theta_3 = 3$을 갖는 공분산을 갖도록 얻어진 오염된 자료이다. 각 모분포에서 주변수와 보조변수에 해당하는 난수를 독립적으로 생성시킨다.

(4) 단계 (3)에서 얻어진 두 변수는 독립이기 때문에 생성된 자료 자체는 이변량 공간자료로 간주할 수 없다. 따라서, 변수간에 서로 상관을 갖도록 변수변환을 시도한다. 변수변환 방법은 두 변수간의 상관 정도를 ρ로 부여하여 변수변환시키는 다음과 같은 방법을 고려했다 (Bickel and Doksum, 1976, p. 25).

$$U = X$$

$$V = \rho X + \sqrt{1-\rho^2}Y$$

변수 X, Y는 모의실험에 의해서 얻어진 주변수와 보조변수이고, U, V는 두 변수간에 상관이 ρ를 갖도록 변환된 변수이다. 여기서 ρ의 변화에 따른 결과를 비교해 보기 위하여 ρ가 -0.3, 0.3, -0.7, 0.7인 경우에 대해서 모의실험을 행한다.

(5) 단계 (4)에서 얻어진 자료(즉 세 가지 모델의 각각에 대해서 네 가지 상관계수를 고려해서 얻어진 자료)를 이용하여 C방법과 P방법을 적절한 기준에 의해서 비교한다. 또한 두 방법에 대한 오염된 자료 25개중에서 탐지된 개체수를 제시함으로써 두 방법에서의 탐지비율을 비교한다.

(6) 위의 단계 (1)에서 단계 (5)까지의 과정을 독립적으로 총 100회 반복 시행한다. 즉, 125개의 공간자료로 구성된 100개의 자료집합이 모의실험에 이용된다.

두 방법의 비교를 위해서 놈을 기준으로 큰 값을 갖는 상위 25개의 개체들 중에서 인위적으로 오염시킨 25개 자료(101-125번째 개체)가 탐지되는 개수들을 100개의 자료집합에 대해서 평균을 구했다. 이에 대한 모의실험 결과는 표 3.1에 주어져 있다. 이 결과에 의하면 P방법이 몇몇의 경우에 있어서 C방법보다 탐지력면에서 좋은 결과를 보이고 있으나, 두 방법에서 탐지된 비율은 그다지 큰 차이를 보이고 있지 않다. 또한 모든 경우(12가지 경우)에 대한 평균 탐지개수는 C방법에서는 16이고, P방법에서는 17의 결과임을 알 수 있다. 오염자료 25개를 기준으로 했을 때, C방법의 탐지비율은 64%, P방법은 68%로서 두 방법 모두에 있어서 그다지 높은 탐지력을 보이지 않고 있다고 판단할 수 있다. 그러나 이것은 방법의 성능에 기인하는 것이 아니라 오염시킨 자료 25개가 교차베리오그램 추정 시 모두 영향력 있는 개체로 나타나는 것은 아니기 때문이다.

표 3.1: P방법에 의해서 얻어진 영향함수의 유클리드 놈을 이용한 영향분석결과

구분			오염자료 25개 중 탐지된 갯수			C방법과 P방법의 일치 비율				
			C방법	P방법		91 - 100%	81 - 90%	71 - 80%	61 - 70%	60% 이하
모델	구형	$\rho = -0.3$	15	17	빈도	0	0	1	45	54
					누적	0	0	1	46	100
		$\rho = 0.3$	15	17	빈도	0	0	10	75	15
					누적	0	0	10	85	100
		$\rho = -0.7$	16	14	빈도	0	27	59	14	0
					누적	0	27	86	100	100
		$\rho = 0.7$	17	18	빈도	0	57	40	3	0
					누적	0	57	97	100	100
	가우시안	$\rho = -0.3$	16	16	빈도	0	0	3	60	37
					누적	0	0	3	63	100
		$\rho = 0.3$	16	17	빈도	0	0	10	73	17
					누적	0	0	10	83	100
		$\rho = -0.7$	18	18	빈도	1	36	53	10	0
					누적	1	37	90	100	100
		$\rho = 0.7$	16	19	빈도	0	57	40	3	0
					누적	0	57	97	100	100
	지수	$\rho = -0.3$	16	17	빈도	0	1	20	75	4
					누적	0	1	21	96	100
		$\rho = 0.3$	14	16	빈도	0	0	3	59	38
					누적	0	0	3	62	100
		$\rho = -0.7$	16	19	빈도	0	13	55	32	0
					누적	0	13	68	100	100
		$\rho = 0.7$	17	16	빈도	0	66	34	0	0
					누적	0	66	100	100	100

그리고 모든 경우(세 가지 모델과 네 가지 상관계수의 조합)에 있어서 영향력이 제일 큰 관측치는 두 방법에서 동일하게 얻어졌다. 제안한 방법의 타당성 검토를 위한 기준은 C방법과 P방법에 의한 결과에서 놈을 기준으로 큰 값을 갖는 상위 25개중 몇 개가 일치하는가에 따라서 두 방법의 유사성에 대한 기준으로 설정했다. 상관계수가 0.7인 구형 모델인 경우를 예로 들면, 25개중에서 81-90%에 해당하는 개체 수는 20-22이므로, 모의실험에 의해서 얻어진 100개의 자료집합 중에서 두 방법이 일치되는 개체의 수가 20-22개를 만족하는 자료집합의 수를 빈도로 나타냈다. 즉, 100개의 자료집합중에서 일치하는 개체수가 20-22개를 만족하는 자료집합이 57개임을 나타낸다.

표 3.1의 결과에서 구형, 가우시안, 지수 모델과 상관계수ρ= -0.7, 0.7의 각각에 대하여

두 방법이 70%이상 일치하는 정도가 100개의 자료집합 중에서 86, 97, 90, 97, 68, 100개가 일치함을 알 수 있다. 지수 모델에서 ρ=-0.7 인 경우의 일치하는 자료집합이 다른 경우보다 적은 수치를 보이고 있으나 거의 70%에 가깝기 때문에 두 방법은 거의 비슷한 결과를 보이고 있다고 할 수 있다. 한편, 세 모델 모두에 있어서 상관계수가 낮은 경우는 상관계수가 높은 경우보다 일치비율이 낮음을 확인할 수 있다.

4. 적용사례

4.1. 데이타

여기에서 사용된 데이터는 National Cartographic Information Center(NCIC)의 Digital Elevation Model(DEM)으로부터 얻어진 고도자료이다(Isaaks and Srivastava(1989), pp 115-119). 이 자료는 470개의 표본위치에서 주변수인 V와 보조변수인 U로 측정된 다변량 공간자료이다. 470개의 표본위치 중에서 첫 195개의 위치에서의 U는 결측값을 갖는다. 그리고 결측값을 갖는 위치를 제외한 275개의 관측값에 대한 두 변수의 상관계수는 0.55이다.

4.2. C방법에 의한 감도분석

교차베리오그램의 추정에 있어서 영향력있는 관측치를 탐지하기 위해서 2.2절에서 소개된 이론을 통하여 영향함수를 유도할 수 있고, 유도된 영향함수의 유용성은 Choi et al.(2000)에서 입증되었다. C방법에 의해 얻어진 영향함수를 이용한 lag별 영향분석결과가 그림 4.1에 주어져 있다.

거의 모든 lag에서 24번째 관측치가 교차베리오그램을 추정하는데 있어서 영향을 미치는 관측치임을 알 수 있다. 그리고 몇몇의 lag에서 4, 36, 37 그리고 174번째 관측치가 영향력 있는 관측치임을 알 수 있다. 이 그림은 PCA를 이용한 방법과 비교하기 위해서 그림 4.3과 비교되어 질 것이다. 그림 4.2는 C방법에 의해서 얻어진 영향함수의 유클리드 놈을 이용한 영향분석결과이다. 이 그림은 그림 4.1의 각 lag에 대한 영향력을 놈의 지표를 통하여 영향력평가를 위한 것이다. 그림 4.1에서 탐지된 영향력 있는 관측치들이 그림 4.2에서도 탐지되고 있음을 볼 수 있다.

4.3. P방법에 의한 감도분석

이 절에서는 4.2절에서의 C방법에 의한 감도분석 결과와 비교되어 진다. 그림 4.3은 교차베리오그램을 이용하지 않고, 변수 V와 U에 대해서 PCA를 행해서 얻어진 제 1 주성분 점수를 이용하여 감도분석한 결과이다. 즉, 주어진 다변량 공간데이터에 대해서 PCA를 행하고, 얻어진 제 1 주성분 점수를 일변량 공간데이터로 상정해서 자기베리오그램에 대한 감도분석 절차를 적용하여 얻어진 결과이다.

P방법에 의한 결과인 그림 4.3은 C방법에 의한 결과인 그림 4.1과 거의 동일한 결과를

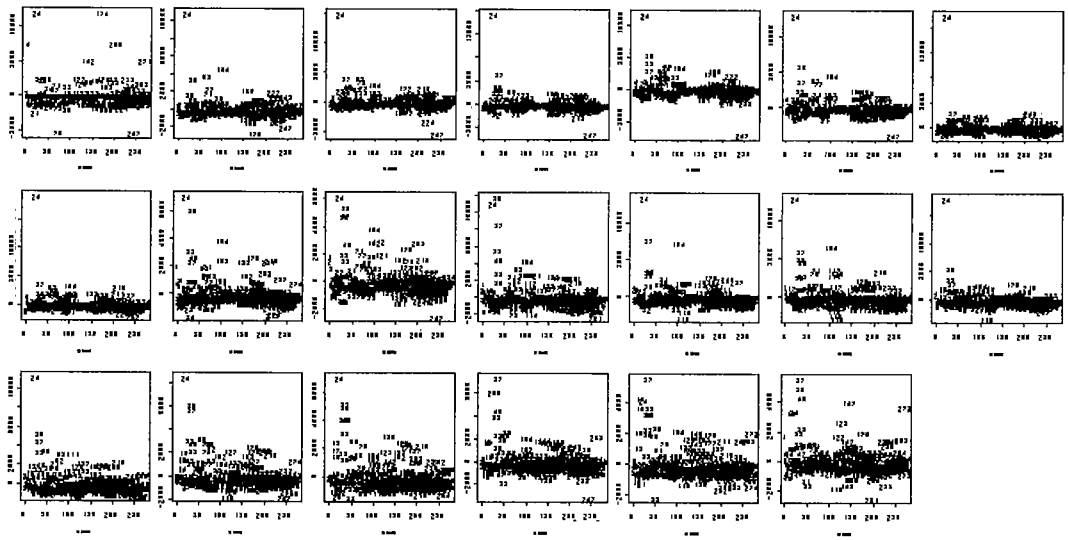

그림 4.1: C방법에 의해 얻어진 영향함수를 이용한 lag별 영향분석결과

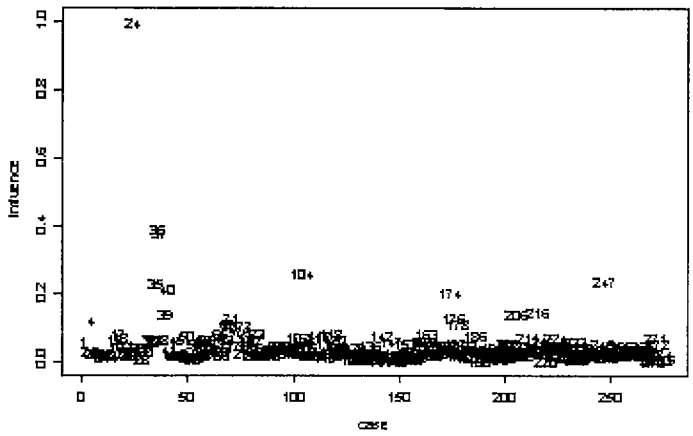

그림 4.2: C방법에 의해서 얻어진 영향함수의 유클리드 놈을 이용한 영향분석결과

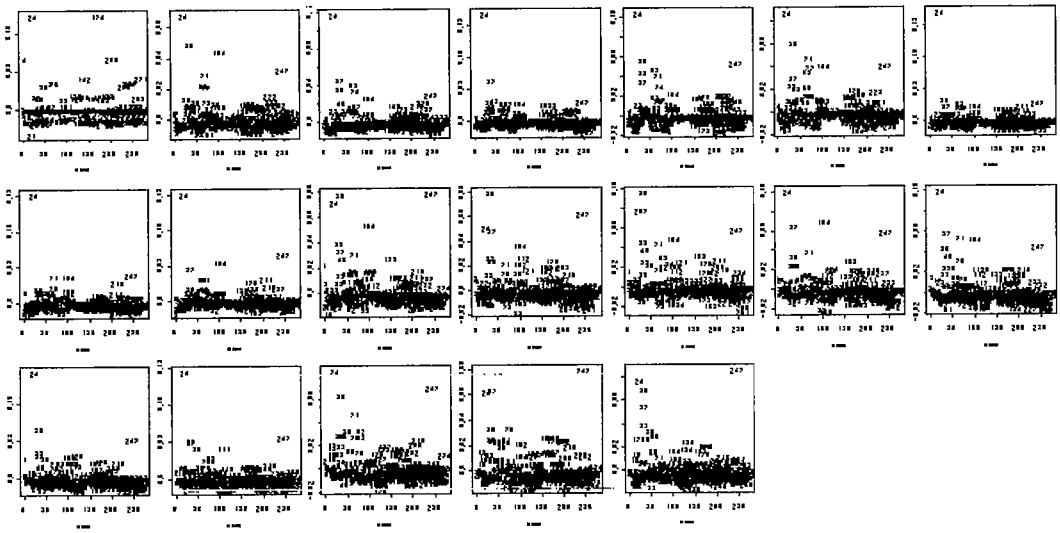

그림 4.3: P방법에 의해 얻어진 영향함수를 이용한 lag별 영향분석결과

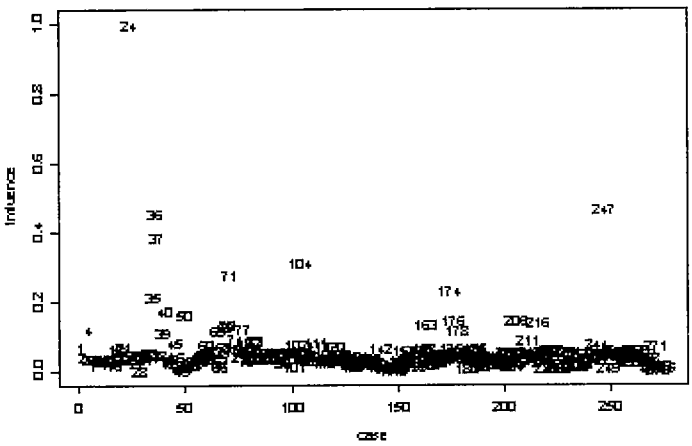

그림 4.4: P방법에 의해서 얻어진 영향함수의 유클리드 놈을 이용한 영향분석결과

보이고 있음을 알 수 있다. 놈에 의한 영향평가의 결과인 그림 4.4 또한 그림 4.2와 비슷한 결과를 보이고 있다. 따라서 이변량 공간자료가 주어졌을 때, 코크리깅을 위한 교차베리오그램의 추정단계에서의 영향분석은 PCA에 의한 결과를 가지고 일변량 공간자료에 대한 영향분석을 통하여 가능하다고 할 수 있다.

5. 결론

본 논문의 3장에서 제안한 방법의 타당성을 보이기 위하여 세 모델과 네 경우의 상관계수를 고려하여 모의실험을 행했다. 시뮬레이션의 결과에서 오염된 25개의 자료에 대해서 두 방법의 탐지비율은 거의 비슷한 결과를 보였다. 그리고 상관계수를 고려한 구형, 가우시안, 그리고 지수모델의 각각에 대해서 두 방법이 70%이상 일치정도가 86, 97, 90, 97, 68, 100개로서, C방법과 P방법을 이용한 감도분석의 결과가 거의 일치함을 보았다. 또한 4장에서는 실제 자료를 이용한 사례분석을 통해서도, 이변량의 경우, 두 방법에 의한 감도분석은 거의 비슷한 결과를 얻을 수 있었다. 일반적으로 C방법에 의한 감도분석은 1) 변수가 3개 이상인 경우에 대한 기존연구가 없으며, 2) 복잡한 과정을 거쳐야하고, 3) 각 단계에서 탐지된 영향력 있는 관측치에 대한 원인 규명이 어렵고, 4) 자료 해석적인 측면에서 설명하기에 어려움이 있다. 본 논문에서, 이변량인 경우에 대해서, 제안한 P방법에 의한 감도분석은 이러한 단점들을 보완할 수 있다는 측면에서 큰 의미를 갖는다고 할 수 있다. 따라서 본 논문은 다음의 두 가지 특징을 갖는다. 1) 조금의 정보에 대한 손실을 감수할 수 있다면, PCA를 사용하여 분석함으로써 C방법의 복잡한 과정을 피할 수 있다는 점과, 2) PCA에 의한 해석적인 차원에서 의미 있는 자료 해석이 가능하다는 장점을 갖는다. 만약 변수가 3개 이상일 때, P방법에 의한 결과에서 2개 이상의 주성분이 유의한 경우에는 각 주성분을 이용하여 일변량 감도분석을 수행한다. 이러한 경우, 감도분석의 결과가 주성분별로 존재하게 되며, 이를 각 주성분에 대해서 해석한다. 그러나 공간자료분석에서 변수가 세 개 이상인 경우에 대한 감도분석이 연구된 바 없기 때문에, 세 개 이상의 변수에 대해서 두 방법론의 비교는 불가능하다. 즉, 변수의 수가 p의 경우, 교차베리오그램은 $\binom{p}{2}$개 만큼 존재하며, 이들 각각에 대한 감도분석은 행할 수 있지만, 코크리깅을 위한 전체 공분산구조에 대한 감도분석은 행할 수 없다. 이를 위해서는 변수가 세 개 이상인 경우에 대한 감도분석의 연구가 필요하다. 따라서 변수가 세 개 이상인 경우는 차후 연구과제로 남기고, 본 연구에서는 변수가 두 개인 경우로 제한한다.

참고문헌

[1] Basu, S., Gunst, R.F., Guertal, E.A. and Hartfield, M.I. (1995). The effects of influential observations on sample semivariograms, *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. **2**. No. 4, 490-512.

[2] Bickel, P.J. and Doksum, K.A. (1989). *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, Inc., San Francisco.

[3] Choi, S.B., Kim, K.K., and Tanaka, Y. (2000). *Sensitivity analysis in Auto- and Cross-variogram estimation*, Journal of the Korean Data Analysis Society, Vol. **2**, No. 1, 91-107.

[4] Choi, S.B. and Tanaka, Y. (2000). Sensitivity analysis in spatial statistics: Detecting influential observations in spatial prediction, *Journal of Japanese Society computational Statistics*, Vol. **13**. 25-39.

[5] Cressie, N.A.C. (1991). *Statistics for Spatial Data*, John Wiley and Sons, New York.

[6] Gunst, R,F. and Hartfield, M.I. (1997). *Robust Semivariogram Estimation in the Presence of Influential Spatial Data Values, In Modelling Longitudinal and Spatially Correlated Data*, Gregoire, T.G. et al.(eds), 265-274. Springer-Verlag, New York.

[7] Isaaks, E.H. and Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics*, Oxford Universal Press, New York.

[8] Kaluzny, S.P., Vega, S.C., Cardoso, T.P. and Shelly, A.A. (1996). *S+SpatialStats User's Manual*, Mathsoft, Inc., Seattle.

[9] Tanaka, Y. (1994). Recent Advance in Sensitivity Analysis in Multivariate Statistical Methods, *Journal of Japanese Society computational Statistics*, Vol **7**, 1-25.

[2001년 6월 접수, 2001년 8월 채택]

Sensitivity Analysis for Bivariate Spatial Data Using Principal Component Score

Seung-Bae Choi [1] Chang-Wan Kang [2]

ABSTRACT

In spatial statistics, multivariate spatial data is analyzed in following three stages; 1) estimating of auto- and cross-variogram, 2) fitting of the auto- and cross-variogram model, and 3) cokriging. This paper proposes sensitivity analysis used the principal component score(PCS) instead of spatial analysis method for multivariate spatial data. In case of multivariate spatial data with two variables, influence function in estimating the cross-variogram as the first stage of cokriging is derived. And also it is derived by using the PCS. The results of sensitivity analysis for two methods are compared.

Keywords: Variogram; Cross-variogram; Cokriging; Sensitivity Analysis.

1) Researcher, Statistical Consulting Center, Dongeui University.
E-mail: statcst@hyomin.dongeui.ac.kr
2) Assistant Professor, Department of Computer Science and Statistics, Dongeui University.
E-mail: cwkang@hyomin.dongeui.ac.kr