크리깅방법에 의한 오존도 예측

장지희[1], 남궁 평[2]

요 약

공간자료에 대한 통계적 모형과 상관관계, 거리모형 등을 고려하여 크리깅 방법에 의한 미 측정지역의 오존도를 예측한다. 서울시의 오존자료를 이용하여 예측한 결과 보통 크리깅방법이 효율적이다.

주요용어 : 크리깅, 차의 분산, 공간예측

1. 서 론

지리학, 생태학, 환경, 농업, 수문학 등의 분야에서 공간 자료를 이용하여 특정지역의 정보를 분석하고 예측하는 학문을 지구통계학(geostatistic)이라고 한다. 공간 자료에 대한 통계적 접근은 본래 실험계획에서 비롯되었으나 실험계획에서의 자료처럼 공간 자료를 임의화, 블록화, 반복화 할 수 없는 문제가 발생하게 되면서 확률과정모형으로 접근하기 시작하였다.

크리깅은 남아프리카의 광산 기술자였던 D. G. Krige의 이름에서 유래되었으며, 1950년대에 추출된 광물질 등급에 기초한 분포로부터 최적의 광물질 등급분포를 결정하기 위한 경험적 방법을 개발하였다. 크리깅에 의한 예측방법은 원래 지질통계학 분야에서 널리 이용되었던 기법으로 현재는 많은 분야에서 활용되고 있으며, 이 방법은 관측점이 불규칙한 경우에 등고선이나 곡면의 보간 등에 유용하다.

확률과정모형을 바탕으로 한 공간 자료에 대한 통계적 모형은 다음 식과 같으며 공간에 대해서 연속이라는 가정 하에 예측하게 된다.

$$Z(s)=\{\, s \mid s \in D \,\}, \quad D \subset R^d$$

여기서 $Z(s)$: 거리 s에 대한 함수, s : 각 표본들 간의 거리, D : 표본이 위치하고 있는 공간, R^d : d 차원의 공간

2. 크리깅 (Kriging)

2.1 단순 크리깅 (Simple Kriging)

단순크리깅은 $E(Z(s))=\mu(s)$, $s \in D$ 를 안다는 가정 하에 사용할 수 있는 가장 간단한 방

1) ADN 연구원, 서울 강남구 논현동 91-3
2) 성균관대학교 통계학과 교수, 서울 종로구 명륜동 3가 53

법으로 추정값의 형태를 관측값들의 선형결합으로 가정한다.

$$\hat{Z}(s_0) = \sum_{i=1}^{n} l_i Z(s_i) + k$$

여기서 $E(Z(s_0) - \hat{Z}(s_0))^2$이 최소가 되기 위한 $l_1, \ldots, l_n, k$ 를 구하면 된다.

2.2 보통 크리깅 (Ordinary Kriging)

보통 크리깅은 크리지(Krige)에 의해 개념이 제안되었고 후에 마테론에 의해 발전되었다. 보통 크리깅은 널리 이용되는 방법으로 모형의 평균은 사전에 알 수 없고 일정하다는 전제 하에 예측하는 방법이다. 모형가정과 예측값에 대한 가정은 다음과 같다.

$$Z(s) = \mu + \delta(s), \quad s \in D, \ \mu \in R, \ \mu = \text{unknown}$$
$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i), \quad \sum_{i=1}^{n} \lambda_i = 1$$

여기서 μ는 미지이고 $\sum_{i=1}^{n} \lambda_i = 1$ 이라는 조건은 불편성을 만족하는 조건이다.

2.3 보편 크리깅 (Universal Kriging)

보편 크리깅은 $\mu(s)$가 고정된 상수가 아니라는 가정 하에서 공간을 예측하는데 $\mu(s)$를 $\{f_0(s), \ldots, f_n(s)\}$, $s \in D$의 선형결합으로 가정한다. 이 경우 $f_i(s)$는 이미 주어진 s에 관한 함수이며 모형가정은 다음과 같다.

$$Z(s) = \sum_{j=1}^{p+1} f_{j-1}(s)\beta_{j-1} + \delta(s), \quad s \in D$$

$$\equiv Z = X\beta + \delta$$

여기서 $\beta \equiv (\beta_0, \ldots, \beta_p)' \in R^{p+1}$, $X \equiv (f_{j-1}(s_i)) : n \times (p+1)$ 행렬

보편 크리깅 방법은 보통 크리깅 방법과 마찬가지로 정규분포 하에서 최적이다.

2.4 거리역수 크리깅 (Inverse Distance Kriging)

거리역수 크리깅은 주변표본들의 $Z(s_i)$이 미 측정지점의 $Z(s_0)$사이가 거리의 영향을 받는 방법을 말한다. 즉, 거리에 역을 취한 값들을 가중값으로 환산하는 방법이다. 예측값 가정은 다음과 같다.

$$\hat{Z}(s_0) = \sum_{i=1}^{n} w_i Z(s_i), \quad (\sum w_i = 1, \ w_i = f(h_i))$$

여기서 $\hat{Z}(s_0)$는 미 측정지점에서의 예측값이고 $Z(s_i)$는 표본들의 관측값이며 w_i는 표본들의 가중값이다.

3. 사례연구

서울의 대기오염(오존)분포도를 그려보기 위한 준비 단계로 차의 분산을 추정하여 이를 크리깅 방법에 적용시킨다. 그리고 미 측정지역의 오존도를 예측해보고, 임의의 한 측정소가 측정기의 고장으로 오존도를 측정할 수 없다는 가정하에 크리깅 방법들을 비교한다. 분석에 사용된 자료는 대기환경월보(2000. 1~12)에 자료이며 현재 서울의 27개 지역 측정소에서 매 8시간마다 1번으로 하루에 17번 측정되고 있다(부록참조).

공간예측을 하기 위해서 차의 분산 분석을 시행하기로 한다. 351개의 원 자료를 이용한 방법의 차의 분산 모형적합은 모형적합의 정확도면에서는 신뢰성이 있는 방법이나 시간이 너무 많이 걸린다는 단점을 지니고 있다. 원 자료를 이용한 적합 방법에 대한 단점인 시간문제를 해결하기 위해 마테론(1962)이 제안한 lag방법을 적용하였다.

<표 1> lag를 이용한 오존자료의 차의 분산 모형적합

	$C(0)$	a_s
lag=6	0.0000165	2.0
lag=7	0.0000138	3.5
lag=8	0.0000195	3.5

<표 2>와 같이 구간 수가 달라짐에 따라서 추정된 차의 분산의 형태에 차이가 있음을 확인 할 수 있다. 따라서 구간을 나누는 방법을 사용할 때 원 자료를 잘 설명해주는 구간 수의 선택 기준에 대한 보다 집중적인 연구가 필요하다. 분석 시 원 자료와 비슷한 차의 분산 모형을 갖는 lag=7을 선택하여 결과를 살펴보았다.

<표 2> 각 쌍들의 거리 구간과 빈도

구간	변수	LAG	하한	상한	빈도	%
1	ozone	0	0	2.50	3	0.009
2		1	2.50	7.50	81	0.231
3		2	7.50	12.50	102	0.291
4		3	12.50	17.50	90	0.256
5		4	17.50	22.50	50	0.142
6		5	22.50	27.50	24	0.068
7		6	27.50	32.47	1	0.003
8		7	32.47	37.47	0	0

<표 3>을 보면 351개의 거리 쌍들의 점이 거리구간별로 해당 거리에 분포되어 있음을 알 수 있다. 올바른 예측을 하기 위해서는 많은 거리 쌍들의 점을 가지고 차의 분산을 계산해야 한다.

여기서 구간 폭이 작다는 것은 가능한 많은 거리구간을 갖는다는 것과 동일한 의미이다. 그러나 차의 분산을 계산하기 위해선 각 구간 내에 적어도 30개의 쌍들의 점을 사용해야 한다. 만약 구간 폭이 매우 작다면 하나 또는 그 이상의 구간에서는 매우 작은 쌍들의 점을 갖게 될 것이다. 반면에 구간 폭이 매우 크다면 거리구간 내에 포함된 쌍들의 점의 수는 예측을 위해 필요이상이 될 것이다.

<표 3>에서 lag0은 전형적으로 제한된 구간으로 나머지 구간들의 너비(4.9958km)에 반(2.4979km)이기 때문에 이 구간에 대해선 차의 분산을 계산하기 위한 약속은 무시된다. 이유는 구간 폭에 따라 lag0에 포함된 쌍들의 수가 30개 보다 작을 수도 많을 수도 있기 때문이다. 여기서는 lag0에 포함되는 쌍들의 점이 3개이나 lag1에서 차의 분산을 구하기 위한 충분한 수를 갖고 있으므로 경험적 차의 분산을 구할 수 있다. 그러나 구간을 나누는 방법에서의 문제점은 독립인 것처럼 보이는 멀리 떨어져있는 점들의 쌍까지도 포함한다는 것이다. 이를 이용해서 경험적 차의 분산의 모형을 그릴 순 없을 것이다. 그러므로 상관관계 정도를 예측한다면 지리학적으로 비슷한 장소를 우선으로 열거할 수 있을 것이다. 따라서 가장 많은 쌍들의 점을 갖는 구간은 둘 이상의 구간을 초과하여 상관관계 정도를 예측하는 것은 의미가 없으므로 그 다음의 하나의 구간만을 고려해야 할 것이다.

<표 3> lag3까지의 차의 분산

구간	각 lag 단위 값	각 lag에 포함된 쌍들의 수	각 lag의 평균거리	차의 분산
1	-1	27	.	.
2	0	3	2.2246	0.0000043
3	1	81	5.3693	0.0000126
4	2	102	10.1315	0.0000118
5	3	90	14.7701	0.0000175

이를 통해 빈도가 가장 높은 lag2에서 한 단계 추가한 구간, 즉 분석에 필요한 구간 수가 3이므로 lag3까지의 차의 분산을 보면 <표 4>와 같다.

1

<표 4> 세 지역의 크리깅 예측값과 분산, 95% 예측구간

지역명	예측값 $\hat{Z}(s_0)$	분산 $\sigma_k^2(s_0)$	95% 예측구간	
			하한 $\hat{Z}(s_0)-1.96\sigma_k(s_0)$	상한 $\hat{Z}(s_0)+1.96\sigma_k(s_0)$
1 지역	0.0179	1.45E-05	0.0105	0.0254
2 지역	0.0172	1.46E-05	0.0097	0.0247
3 지역	0.0179	1.23E-05	0.0110	0.0247

lag=7을 선택한 결과가 $C(0)=0.0000138$, $a_s=3.5$이므로 적합된 모형식은 다음과 같다.

$$r(h)=0.0000138[(1-\exp(-\frac{\|h\|^2}{3.5^2})]$$

따라서 위의 모형식으로 세 지역의 크리깅 예측값과 분산 그리고 95% 예측구간을 구한 결과는 〈표 4〉 와 같다. 3 지역의 분산이 가장 작게 나타났으며, 이 예측값으로 미 측정지역의 오존오염도를 파악할 수 있다.

4.결 론

미 측정 또는 측정의 신뢰도가 떨어지는 측정소의 오존을 주변 측정소의 관측값을 이용하여 예측하였다. 보통 크리깅 방법은 추정위치의 물리량(오존도)를 계산하기 위해 주변 기지점과의 거리 및 물리량을 고려하며, 거리역수 크리깅 방법은 거리만이 고려 대상이다. 따라서 두 크리깅 방법의 차이는 물리량의 상대적 크기 차를 어떻게 반영하는가에 있다. 두 가지 방법은 각각 물리량의 크기와 거리, 단순거리 등이 가중값을 구성하게 되는 것이다. 미 측정으로 가정한 구의동의 오존도를 100% 신뢰한다고 가정한다면 결과를 살펴볼 때 두 크리깅 방법 중 보통 크리깅 방법이 미미하게 좋은 결과를 보여준다. 두 크리깅 방법 간의 차이는 그다지 크지 않은데 이 사실은 물리량, 즉 오존도의 공간적 변동성은 그 거리의 차이에 비하여 미미하다는 사실을 의미한다. 다시 말해 가중값은 오존도보다 측정소 간의 거리에 의존함을 알 수 있다.

한편, 경제적인 면을 고려하여 측정소 개수를 줄이는 방법을 생각해보았다. 그 결과 측정소 개수를 줄이더라도 두 방법의 크리깅 예측값은 미미한 차이로 보통 크리깅 예측값이 실제값에 가까운 결과가 나왔다. 거리를 고려한 경우(구의동과 거리가 먼 18개 측정소의 자료만 이용했을 경우)도 마찬가지 결과가 나왔다. 여기서 예측오차가 무작위로 선택한 18개 측정소의 자료만을 이용했을 경우보다 큰 것으로 보아 공간 상 거리에 영향을 받는다는 것을 알 수 있었다. 예측하고자 하는 구의동 오존도가 거리가 먼 경우에도 어느 정도 실제값을 예측할 수 있었는데 이것은 공간적 변동성 자체가 작기 때문에 이러한 결과가 나온다고 할 수 있다.

다만 보통 크리깅의 가정인 정규성과 lag문제가 발생되지 않는다면 같은 결과를 얻을 수 있겠다. 반면 정규성을 만족하나 lag문제가 발생한다면 적절한 보통 크리깅 예측값을 얻을 수 없으므로 거리만을 가중값으로 두는 거리역수 크리깅 방법을 이용하는 것이 바람직하다.

참고문헌

[1] 박성현 (1996), 우리나라 환경통계의 현황, 응용통계연구, 제9권 1호, pp. 179-202.

[2] 유성모, 엄익현 (1997), 강우강도자료를 이용한 semi-variogram의 추정과 예측, Proceedings of the Spring Conference, Korean Statistical Society.

[4] 한국수자원공사 (1996), Kriging기법을 이용한 강우 공간분포에 관한 연구.

[5] Cressie, N. (1989). Spatial Prediction and Ordinary Kriging, *Mathematical Geology*, 21, pp. 493-494.

[6] Cressie, N. (1991), *Statistical for Spatial Data*, New York, Wiley.

[7] Isobel Clack. and Paker, H. (1980), *Geostatistics*, Mackay School of Mines.

[8] Jean-Paul Chiles Pierre Delfiner (1999), *Geostatistics Modeling Spatial Uncertainty*, John Wiley & Sons, Inc.

[9] Matheron, G.(1963), *Principles of Geostatistics*, Economic Geology, 58, pp.1246-1266

[10] Sacks, et. al. (1989), DACE(*Design and Analysis of Computer Experiments*).

[11] Steven K. Thompson (1992), *Sampling*, John Wiley & Sons, Inc.

[12] William, V. H. and Isobel Clack. (2000), *Practical Geostatistics 2000*, Ecosse North.

<부록> 서울의 측정소 위치와 2000년 오존오염도 자료

측정소명	TM좌표(m)		기준점(시청앞) 과의 거리(km)		2000년 오존오염도 (ppm)
	가로	세로	가로	세로	
시청앞	197,868	451,606	0	0	0.014
이화동	199,850	452,750	1.982	1.144	0.018
면목동	206,660	454,115	8.792	2.509	0.018
신설동	202,078	452,262	4.21	0.656	0.015
불광동	193,987	456,566	-3.881	4.96	0.022
마포	195,328	449,353	-2.54	-2.253	0.018
문래동	190,337	445,935	-7.531	-5.671	0.018
사당동	197,670	443,050	-0.198	-8.556	0.010
관악산	195,954	439,387	-1.914	-12.219	0.027
대치동	204,930	443,540	7.062	-8.066	0.017
잠실동	207,338	444,941	9.47	-6.665	0.016
시흥동	191,970	439,000	-5.898	-12.606	0.010
천호동	210,850	448,800	12.982	-2.806	0.016
번동	202,540	459,210	4.672	7.604	0.014
길음동	202,338	456,083	4.47	4.477	0.020
한남동	200,490	448,720	2.622	-2.886	0.018
구의동	208,150	449,280	10.282	-2.326	0.017
성수동	204,870	449,810	7.002	-1.796	0.013
방학동	203,450	461,840	5.582	10.234	0.022
남가좌동	192,320	452,400	-5.548	0.794	0.013
구로동	190,720	442,200	-7.148	-9.406	0.016
반포동	199,530	444,790	1.662	-6.816	0.017
화곡동	184,560	449,570	-13.308	-2.036	0.015
방이동	210,990	446,700	13.122	-4.906	0.017
신정동	187,400	447,000	-10.468	-4.606	0.015
상계동	206,180	461,900	8.312	10.294	0.024
궁동	184,950	444,350	-12.918	-7.256	0.017