강우량 추정에서 유전자 알고리즘을 활용한 크리깅 방법의 적용 *

류제선 [1] 박영선 [2] 차경준 [3]

요 약

공간적으로 영향을 받는 위치에서의 상호 연관성을 고려한 예측모형 중에서 크리깅(kriging) 방법은 관측된 데이터를 보간(interpolation)하고, 부드럽게 연결(smoothing)하며, 새로운 데이터를 예측(prediction)하는 통계적 모형으로서 많이 활용되고 있다. 크리깅 모형을 적용하기 위해서는 먼저 주어진 두 위치에서의 비연관성을 나타내는 세미베리오그램 (semivariogram)의 3가지 모수(nugget, sill, range)를 추정해야 한다. 본 연구에서는 전역적 최적화 방법인 유전자 알고리즘(genetic algorithm)을 도입하여 세미베리오그램 모수들을 추정하였고, 이를 통해 강우량(rainfall)에 대한 크리깅 추정량을 산출하고 효과성을 판단하였다.

주요용어: 강우량, 크리깅, 세미베리오그램, 유전자 알고리즘

1. 서론

최근 들어, 홍수 등과 같은 기후변동에 의한 피해를 줄이기 위해 기후자료에 대한 적절한 분포를 찾으려는 연구가 진행되고 있다. 실제로, 계획되어 있는 홍수량을 초과하는 강우량이 빈번하게 발생하고 있으며, 기존의 홍수방어 시설물에 대한 안전도를 저해하고 있는 형편이다(이동률, 2002). 이에, 강우관측망의 설계(이재형 등, 2002)와 강우모형의 연구(오은선, 2002)가 활발하게 진행되고 있다.

본 연구에서는 강우량에 대한 적절한 추정량을 세워보고자, 강우관측 지점에서 채취한 13년간의 강우량 데이터에 대하여 크리깅 모형을 적용하여 그 추정량을 제시하였다. 이를 위해, 유전자 알고리즘(genetic algorithm)을 적용하여 세미베리오그램의 3가지 모수를 추정하는데 있어 자동화할 수 있도록 하였다. 최근의 세미베리오그램 모수 추정 프로그램은 시각적으로 결정하거나, MLE(Maximum Likelihood Estimator) 방법을 적용하고 있다(Stephen 등, 1996). 그러나, 시각적으로 결정하는 것은 부정확한 결론을 얻을 수 있으며, MLE 방법을 적용하는 것은 수렴하지 않을 때, 잘못된 정보를 제공하는 원인이 될 수도 있

* 이 논문은 2002년 한양대학교 일반 연구비 지원으로 연구되었음.

1) (133-701) 서울시 성동구 행당동 17, 한양대학교 자연과학대학 수학과, 시간강사
E-mail: fbwptjs@daum.net
2) (133-701) 서울시 성동구 행당동 17, 한양대학교 자연과학 연구소, 연구교수
E-mail: pppppys@hanyang.ac.kr
3) (133-701) 서울시 성동구 행당동 17, 한양대학교 자연과학대학 수학과, 교수
E-mail: kjcha@hanyang.ac.kr

다. 따라서, 본 연구에서는 세미베리오그램의 모수를 추정하는 데 있어, 어느 정도의 오차를 고려하면서 자동화되고 유한한 시간에 결정할 수 있는 유전자 알고리즘을 이용하였다.

2장에서는 크리깅 중에 가장 보편적으로 사용되고 있는 보통 크리깅과 세미베리오그램에 대하여 설명하였다. 3장에서는 실측된 강우량 자료에 대하여 유전자 알고리즘을 통한 세미베리오그램 모수를 추정하였고, 이를 이용하여 강우량에 대한 크리깅 추정치를 산출하였다. 4장에서는 본 연구의 결론을 제시하였다.

2. 크리깅 모형 (Kriging Model)

공간적 위치 $\{ s_1, \ldots, s_n\} \in D \subset R^d$에서 관측된 n개의 자료 $Z = (Z(s_1), \cdots, Z(s_n))'$는 확률공간 $D \subset R^d$에서의 점들과 반응값으로 표현되며, 이는 확률과정

$$\{Z(s) : s \in D \subset R^d\}$$

의 실현값으로서 모델링된다.

지역 안에 있는 임의의 두 지점 s_1, s_2에서 확률과정 $Z(s)$에 의해 가정된 값들 사이에서의 종속성은 다음과 같이 정의된 세미베리오그램 함수를 사용한다.

$$\Gamma(s_1, s_2) = \frac{1}{2}V[Z(s_1) - Z(s_2)] = \frac{1}{2}E[\{(Z(s_1) - Z(s_2)) - (m(s_1) - m(s_2))\}^2],$$

여기서, $m(\cdot)$은 모든 가능한 확률과정 $Z(s)$의 실현값의 평균이다. 확률변수 Z의 효율적인 분석을 위해 이차 정상성을 가정하면, 세미베리오그램을

$$\Gamma(s_1, s_2) = \frac{1}{2}E[\{(Z(s_1) - Z(s_2))^2] = \Gamma(s_1 - s_2)$$

와 같이 두 점 s_1과 s_2의 거리에 의한 함수로서 나타낼 수 있다.

2.1. 보통 크리깅(Ordinary Kriging)

보통 크리깅은 주어진 점 $s_i, i = 1, 2, \cdots, n$에서 확률과정 $Z(s)$의 n개의 관측값을 알고 있다고 가정할 때, 임의의 점 s_0에서의 추정값 $\hat{Z}(s_0)$는 선형결합

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

으로 추정되며, 여기서 $\sum_{i=1}^{n} \lambda_i = 1$이고 $\hat{Z}(s_0)$는 최소분산을 갖는 최량선형비편향예측량(Best Linear Unbiased Predictor)이다.

위의 수식을 행렬 형식으로 표현하면, 주어진 점 s에서의 실현값 $Z(s)$에 대하여

$$\hat{Z} = \lambda Z(s)$$

이다. 이때, $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_n)$를 크리깅 계수라 하며 보통 크리깅은 크리깅 계수를 찾는 데에 목적이 있다.

간단한 계산 과정을 거쳐 크리깅 계수를 산출하면

$$\lambda = \Gamma^{-1}\left(\gamma + 1_n \frac{1 - 1_n'\Gamma^{-1}\gamma}{1_n'\Gamma^{-1}\gamma}\right)$$

이 된다. 여기서, 1_n은 길이가 n이고 모든 원소가 1인 벡터이며, Γ는 주어진 점들 사이에서의 세미베리오그램이고, γ는 주어진 점들과 추정하고자 하는 점 사이에서의 세미베리오그램이다.

2.2. 세미베리오그램의 산출

위에서 정의된 세미베리오그램의 적률 추정량은, 임의의 두 점 s_i와 s_j에서

$$\hat{\Gamma}(h) \equiv \frac{1}{2} \sum_{\|N(h)\|} (Z(s_i) - Z(s_j))^2$$

이며, 여기서 $N(h) \equiv \{(s_i, s_j); s_i - s_j = h \in R; i, j = 1, 2, \cdots, n\}$이고 $\|N(h)\|$는 $N(h)$에 있는 거리쌍들의 개수이다.

세미베리오그램 모형은 공식화하는 것이 가능하며, 본 연구에서는 다음과 같은

$$v_0 + v_1 \left[1 - \exp\left\{-\left(\frac{h}{v_2}\right)^2\right\}\right]$$

gaussian 모형을 적용하였다. 여기에서 v_0, v_1, v_2은 각각 nugget, sill, range라 한다.

3. 실증분석

본 연구에 사용된 강우량 데이터는 통계청 홈페이지 (http://www.nso.go.kr/)의 통계 DB인 KOSIS에서 제공하고 있는 강우량에 관한 자료와 기상청에서 발간하고 있는 기상연보를 활용하였다. KOSIS에서 제공하고 있는 수량 데이터는 강우량 관측 지점명과 월별 강우량에 대하여 제공하고 있으며, 기상연보에서 관측지점별 경도와 위도를 제공하고 있다. 본 연구에서는 과거 13년(1990 ~ 2002) 동안 강우량이 가장 많았던 8월의 데이터에 대한 평균값을 적용하였다. 여기에서, 울진 데이터는 결측치로 인하여 2000년 이후에 측정한 결과를 이용하였고, 울릉도 데이터는 이상치로서 제외하였다. 본 연구에서의 크리깅 방법 및 유전자 알고리즘에 관한 프로그램은 S-Plus 2000의 사용자 함수를 이용하였다.

3.1. 실험적 세미베리오그램의 산출

세미베리오그램은 관측값을 이용하여 추정하였는데, 우선 24개의 관측된 지점에서 모든 가능한 조합 ${}_{24}C_2 = 276$개의 표본 세미베리오그램을 산출하였다. 다음으로, 주어진 lag에 따라 관측된 거리의 평균과 산출된 세미베리오그램의 평균값을 구하였다.

[표 1] lag=7에서 강우량 측정치의 세미베리오그램

meandist	meangamma	npoint
0.3963208	1414.683	9
0.7913371	1495.487	31
1.1863366	1503.542	27
1.5464780	1874.956	38
1.9462568	2570.954	38
2.3088817	1722.566	36

본 연구에서는 lag별 추정선을 고려해 볼 때, 오차가 적을 것으로 판단된 lag=7인 결과를 이용하여 세미베리오그램의 모수를 결정하기로 하였으며, 그에 대한 값은 [표 1]에서와 같다. 여기에서, meandist는 평균 거리이며 meangamma는 세미베리오그램의 평균값 그리고 npoint는 점들의 갯수이다.

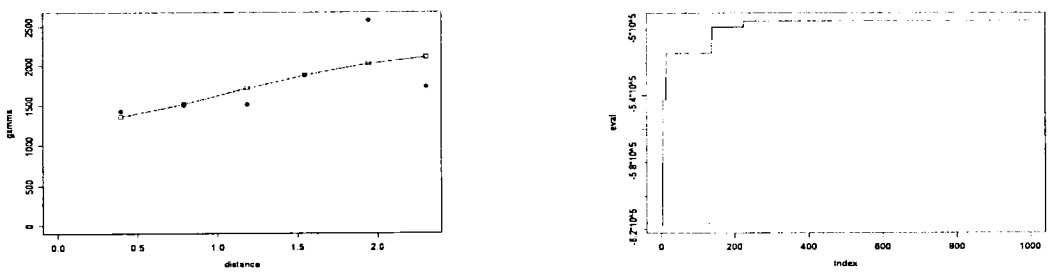
3.2. 유전자 알고리즘을 이용한 세미베리오그램 모수의 결정

유전자 알고리즘에서 사용할 내부적인 숫자의 표현방법은 정밀도 문제에서 실수가 이진수보다 우수하기 때문에 실수를 취하여 분석하였다. 또한, 컴퓨팅 시간을 고려하여 염색체의 수는 $m = 20$으로 하고, 세대수를 1000으로 하였다.

교배연산자는 산술적 교배를 적용하였다. 이는 두 부모 g_i^t와 g_j^t가 교배될 때, 그 자손세대는 1차 선형 결합(linear combination)

$$g_k^{t+1} = ag_i^t + (1-a)g_j^t, if R \leq p_c, g_i^t > g_j^t$$

으로 정의된다. 여기서 g_k^{t+1}는 $t+1$세대의 임의의 k번째 염색체이고, a는 구간 [0.5,1]에서의 가중치, g_i^t와 g_j^t는 각각 t세대에서의 i, j번째 염색체이다. R은 구간 [0,1]에서의 난수이며 p_c는 임의의 주어지는 교배율(crossover rate)이다. 본 연구에서는 20개의 염색체 중 임의의



(a) 최소제곱 추정선 (b) 세대별 진화과정

[그림 1] 유전자 알고리즘의 추정

[표 2] 추정된 세미베리오그램의 모수와 최소제곱 추정값

nugget	sill	range	root MSE
1255.016	880.7061	1.35454	116.7271

개수는 좋은 형질을 가진 유전자를, 나머지 유전자에 대하여 가중치 $a = 0.7$ 그리고 교배율은 $p_c = 0.3$으로 하였다. 또한, 돌연변이 연산자는 균등 돌연변이 연산자를 적용하였으며, 이때, t세대의 임의의 염색체

$$g_i^t =< g_1, \ldots, g_m >$$

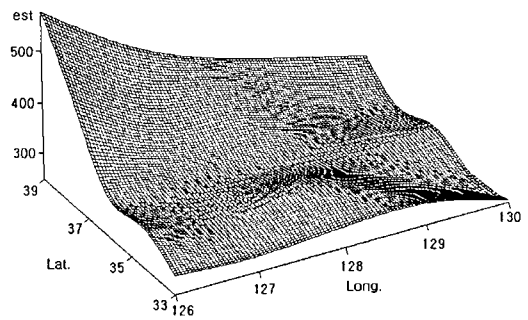
내의 각 원소 $g_i, i = 1, \ldots, m$는 같은 돌연변이 확률 p_e를 갖는다. 이 연산자가 한 번 작용하면, 그 결과로 $t+1$세대의 염색체 $< g_1, \ldots, g_k', \ldots, g_m >$가 생성되며, 이때 g_k'는 k번째 원소가 속하는 영역내의 난수이다. 본 연구에서는 돌연변이 비율 p_e을 0.1로 분석하였다.

분석 결과의 평가를 위하여 산출된 평균거리에서의 평균 세미베리오그램과 추정량의 차의 제곱값을 최소화되도록 하는 모수를 산출하였다. 최적의 세미베리오그램의 3가지 모수 nugget, sill, range와 최소제곱 추정값(root MSE)은 [표 2]와 같다. 또한, [그림 1]에서는 유전자 알고리즘을 이용한 최소제곱 추정선(a)과 세대별 진화과정(b)을 보여주고 있다.

3.3. 크리깅 추정량

[표 2]에서 추정된 세미베리오그램의 3가지 모수를 이용하여 관측된 점들 사이에서의 비연관성과 더불어 관측되어야 할 점과 관측된 점들 사이에서의 비연관성을 결정할 수 있다.

또한, 크리깅 계수는 세미베리오그램에 의해 결정되므로, 본 연구에서는 크리깅 추정의 효과성을 살펴보기 위해 관측지점 하나를 제외하고 나머지 관측지점으로 제외된 하나의 지점을 추정하는 cross-validation 방법을 이용하였다. 이에 대한 결과로서, 전국적으로 추정된 강우량 분포를 시각화하기 위하여 [그림 2]와 같이 추정값에 대한 3차원 그림을 그려 보았다. 전반적으로 서울, 경기 부근에 많은 강우량을, 부산, 광주 부근에서 적은 강우량을 나타내고 있는 것으로 보인다. 서울에서 가장 큰 편차가 있었는데, 이는 데이터에 있어 다른 지역에 비해 큰 강우량 수치를 보이고 있기 때문으로 사료된다.



[그림 2] 크리깅 추정량에 의한 강우량 분포도

4. 결론

본 연구는 유전자 알고리즘을 적용하여 추정된 세미베리오그램의 모수를 이용한 결과, 강우량 추정에서의 크리깅 방법을 이용한 결과에서 매우 효과적인 결과를 얻을 수 있었다(root MSE = 8.41). 이에 대한 장점은 다음과 같았다.

첫째, 기존의 세미베리오그램의 추정은 시각적인 확인 과정을 필요로 하며, 이는 중간 단계에서의 검증절차가 요구될 뿐만 아니라, 그 결과에 대한 객관적 신뢰성이 부족하다. 이에 비해, 유전자 알고리즘은 세대가 증가함에 따라 최적화된 결과를 제공함과 동시에 크리깅 방법을 적용하는 데 있어 자동화 프로세스를 구축할 수 있다는 장점이 있다.

둘째, 일부 세미베리오그램 추정 패키지의 경우에는 MLE 방법을 이용하고 있으나, 이 방법은 종종 무한 loop에 빠져 사용자가 원하는 결과를 제공하지 못하는 경우가 발생하기도 한다. 이에 비해, 유전자 알고리즘은 사용자가 원하는 계산 시간을 미리 지정할 수 있기 때문에 유한한 시간에서의 전역적으로 최적화된 값을 찾을 수 있다고 하겠다.

참고문헌

[1] 통계청 홈페이지, http://www.nso.go.kr/

[2] 기상연보, 2001, 기상청.

[3] 오은선, 2002, 강수량분포에 적용되는 Kappa분포의 모수추정, 전남대학교석사학위논문.

[4] 이동률, 2002, 기후변동과 확률강우량의 변화, 건설기술정보.

[5] 이재형, 유양규, 정재성, 2002, 강우관측망 최적설계 기법 개선에 관한 연구, 대한토목학회논문집, 22;5;B, pp.671-677.

[6] Chambers, L. 1994, Practical handbook of genetic algorithm, CRC Press.

[7] Cressie, N., 1991, Statistics for spatial data, John Wiley & Sons, New York.

[8] Matheron, G., 1971, The theory of regionalized variables and its applications, Cahiers du centre de morphologie mathematique, 5, Fontainebleau, France

[9] Ryu, J.S. et al., 2002, Kriging interpolation methods in geostatistics and DACE Model, KSME international journal, 16(5), pp.619-632.

[10] Sacks, J. et al., 1989, Design and analysis of computer experiments, Statistical Science, 4, No.4, pp.409-435.

[11] Stephen, P.K. et al., 1996, S+ spatialstats User's Manual, MathSoft, Inc.