The Korean Communications
in Statistics Vol. 11 No. 2, 2004
pp. 355-368

The Estimation of Theoretical Semivariogram Adapting Genetic Algorithm for Kriging

JeSeon Ryu[1], YoungSun Park[2], and KyungJoon Cha[3]

Abstract

In order to use Kriging, one has to estimate three parameters(nugget, sill and range) of semivariogram, which shows the relationship in the given two sites. A visual fit of the semivariogram parameters to a few standard models is widely used. But, it does not give the suitable results and not provide the automated process of Kriging. The gradient based nonlinear least squares is another choices to estimate three parameters, but it has some problems such as initial value problem.

In this paper, we suggest the genetic algorithm as a compatible alternative method to solve the above mentioned problem. Finally, we estimate three parameters of semivariogram of rain-fall by adapting the genetic algorithm, compute Kriging estimate and conclude its effectiveness and compatibility.

Keywords : Kriging, Semivariogram, Genetic algorithm

1. 서론

많은 과학적 현상에서 서로 다른 위치에 있는 자료들이 공간적인 상호작용에 의해 영향을 받는 복잡한 모형들이 연구되고 있다. 실제로, 지리학, 생태학, 환경과학 및 기계 공학 등의 분야에서 공간 데이터(spatial data)를 이용하여 새로운 지점에서의 데이터를 예측해야 하는 경우가 있다.

공간 데이터에 대한 통계적 접근의 출발은 실험계획에서 비롯되었다고 할 수 있다. 그러나, 실험계획에서의 데이터처럼 공간 데이터를 임의화, 블록화, 반복화할 수 없는 문제가 발생하게 되면서 확률과정 모형(random process model)으로 접근하게 된다.

확률과정 모형을 바탕으로 한 공간 데이터의 통계적 모형은 다음과 같다.

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

여기에서, D는 연속인 관심공간이며, s는 D에서의 관측지점이다.

이러한 확률과정 방법으로서 크리깅이 있다. 크리깅은 남아프리카 공화국의 광산 기술자였던 D.G. Krige(1951)의 이름을 본따 이름지어졌으며, 그는 1950년대에 샘플링된 광물질 등급(ore

1) Post doc, Center of Innovative Design Optimization Technology, Hanyang University, Seoul, 133-791, Korea.
Email : fbwptjs@ihanyang.ac.kr
2) Research Professor, The Research Institute of Natural Sciences, Hanyang University, Seoul, 133-791, Korea.
3) Professor, Department of Mathematics, Hanyang University, Seoul, 133-791, Korea.

grade)에 기초된 분포로부터 최적의 분포를 결정하기 위한 경험적 방법을 개발했다. 또한, 크리깅은 토양의 성질(soil characteristics; Webster, 1985), 강우량(rainfall; Bacchi 등, 1995; Lee, 2003), 유전빈도(gene frequency; Piazza 등, 1983), 이미지 열코딩(image sequence coding; Decenciere, 1998), 오존도(ozone level; Namkung 등, 2003)와 같은 공간적으로 분포된 변수의 관측되지는 않았지만, 가능한 값을 예측하는 데에 사용되고 있다.

크리깅 방법에는 여러 가지 방법이 있다. 평균을 안다는 가정하에 사용할 수 있는 단순크리깅(simple Kriging, Matheron, 1971)과 평균을 모르지만 일정하다고 가정하여 사용하는 범용크리깅(ordinary Kriging; Matheron, 1971; Jornel and Huijbregts, 1978), 그리고 공간에 대한 경향성을 고려한 일반크리깅(universal Kriging; Matheron, 1969; Matheron and Huijbregts, 1971) 등이 있으며, 중앙값을 이용한 방법인 Median Polish(Tukey, 1977)와 범용크리깅 방법을 혼합한 Median Polish 크리깅(Cressie, 1991) 방법 또한 자주 적용되고 있는 방법이다.

Sacks 등(1989)에 의해 발표된 최초의 논문제목인 전산실험(Design and Analysis of Computer Experiments; DACE) 모형에서 적용되고 있으며, 다분야통합최적설계(multidisciplinary optimal design) 등 공학분야에 널리 사용되고 있다 (Ryu 등, 2002).

최근 들어, 홍수 등과 같은 기후변동에 의한 피해를 줄이기 위해 기후자료에 대한 적절한 분포를 찾으려는 연구가 진행되고 있다. 실제로, 계획된 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 세워 보고자, 강우관측 지점에서 채취한 13년간의 강우량 데이터에 대하여 크리깅 방법을 적용하여 크리깅 추정값을 제시하였다. 크리깅 추정량은 관측값의 선형결합으로 결정되며, 양 혹은 음의 가중치를 산출하는 데에 목적이 있다. 이를 위해, 주어진 두 지점의 연관성을 나타내는 반변이도라는 상관함수가 필요하다. 반변이도 함수는 3가지의 모수(parameter)로 구성되어 있으며, 반변이도 모형을 추정하는 것은 3가지의 모수를 추정하는 데에 그 목적이 있다.

최근의 반변이도 모수 추정 프로그램(S-plus, SAS 등)은 시각적으로 결정하거나, 그래디언트 기반의 비선형 최소제곱법(nonlinear least squares)을 적용하고 있다(Stephen 등, 1996). 그러나, 시각적으로 결정하는 것은 객관적이지 못하며, 또한 사용자에 의한 결정단계가 필요하기 때문에 자동화 프로세스를 수행하는 데에 문제가 있다. 그래디언트 기반의 비선형 최소제곱법은 초기값 문제와 국부최소점(local minima)에 빠져 수렴하지 못하면 원하는 결과와 다른 결과를 주는 경우가 발생하는 단점이 있다. 따라서, 이러한 단점을 보완하고자 반변이도 모수 추정에서 확률적 최적화 방법을 적용한 연구가 있었으며(최종근 등, 2002), 본 연구에서는 유전자 알고리즘(genetic algorithm)을 적용하여 반변이도의 모수를 추정하였고, 이를 통해 실측된 강우량에 대한 크리깅 추정량을 산출할 수 있었다. 유전자 알고리즘은 확률적으로 전역화된 최적값을 제공하는 최적화 방법 중 하나로서, 주어진 시간 내에 초기값에 영향을 받지 않는 최적화 방법이라는 측면에서 여러 문제에 적용되어 왔다(김여근 등, 1997). 본 연구에서는 수학적 예제와 실측된 강우량 자료에 그래디언트 방법과 유전자 알고리즘을 적용하여 반변이도 모수를 추정하였고, 이를 통해 유전자 알고리즘은 반변이도 모수 추정에서 경쟁력 있는 대안으로서 그 효과성과 유용성을 제안하였다.

2장에서는 반변이도에 대한 설명과 함께, 반변이도 모수 추정에서 그래디언트 기반의 최소제곱법의 문제점을 제시하였다. 3장에서는 유전자 알고리즘의 소개와 적용방법에 대해 정리하였으며, 2장에서 발생된 문제의 대안적 해결 방법으로서 유전자 알고리즘을 적용하였다. 4장에서는 범용크

리깅에 대한 소개를 하였고, 5장에서는 실측된 강우량의 크리깅 추정치를 산출하기 위해 그래디언트 기반의 최소제곱법과 유전자 알고리즘을 이용한 반변이도 모수 추정 결과를 이용하였으며, 두 방법의 결과에 대한 비교를 하였다. 6장에서는 본 연구의 결론으로서 유전자 알고리즘은 반변이도 모수를 추정하는 데 있어 경쟁적 대안임을 설명하였다.

2. 반변이도와 그래디언트 방법을 적용한 모수 추정

2.1 반변이도(semivariogram)의 소개

반변이도는 일정한 거리에 있는 자료들의 연관성을 나타내는 척도이다. 따라서, 거리가 가까우면 그 값들이 비슷하므로 작은 값을 갖고, 멀어질수록 그 값이 크게 나타나는 것이 일반적인 경향이다. 즉, 관심 지역 $\boldsymbol{D}$ 안에 있는 두 개의 지점 s_1, s_2에서 확률과정 $Z(s)$에 의해 가정된 값들 사이에서의 연관성은, 다음과 같이 정의된 변이도 함수로 표현되어 질 수 있다.

$$2\gamma(s_1, s_2) = Var[Z(s_1) - Z(s_2)] = E[\{Z(s_1) - Z(s_2)\}^2].$$

실험적 반변이도는 Matheron(1962)에 의하여 제안되었으며, 추정량은 다음과 같다.

$$\hat{\gamma}(\boldsymbol{h}) = \frac{1}{2|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} (Z(s_i) - Z(s_j))^2,$$

$$N(\boldsymbol{h}) \equiv \{(s_i - s_j);\ s_i - s_j = \boldsymbol{h},\ i < j,\ i, j = 1, \cdots, n\}.$$

여기에서 $\hat{\gamma}(\boldsymbol{h})$을 실험적 반변이도라 하며, $|N(\boldsymbol{h})|$는 $N(\boldsymbol{h})$의 서로 다른 원소의 개수이다. 즉, $\hat{\gamma}(\boldsymbol{h})$는 평균이 상수라는 가정에서 적률법(method of moments)을 이용한 불편추정량이다.

실험적 반변이도를 통하여 관측된 공간적 변동을 표현하기 위해 적절한 이론적 모형에 대한 공식화가 필요하다. 관측점들의 반변이도 행렬을 양정치 행렬(positive definite matrix)이 되도록 사

[표 1] 반변이도의 모형

이름	이론적 변이도의 모형
spherical	$v_0 + \frac{v_1}{2}\left\{\frac{3\|\boldsymbol{h}\|}{v_2} - \left(\frac{\|\boldsymbol{h}\|}{v_2}\right)^3\right\}$
exponential	$v_0 + v_1\left\{1 - \exp\left(-\frac{\|\boldsymbol{h}\|}{v_2}\right)\right\}$
Gaussian	$v_0 + v_1\left\{1 - \exp\left(-\frac{\|\boldsymbol{h}\|}{v_2}\right)^2\right\}$
wave	$v_0 + v_1\left\{1 - v_2\sin\left(\frac{\|\boldsymbol{h}\|}{v_2}\right)/\|\boldsymbol{h}\|\right\}$
power	$p_0 + p_1\|\boldsymbol{h}\|^{p_2}$

용되는 반변이도 모형은 [표 1]과 같다. 여기에서 반변이도의 모수 v_0, v_1, v_2는 각각 nugget, sill,

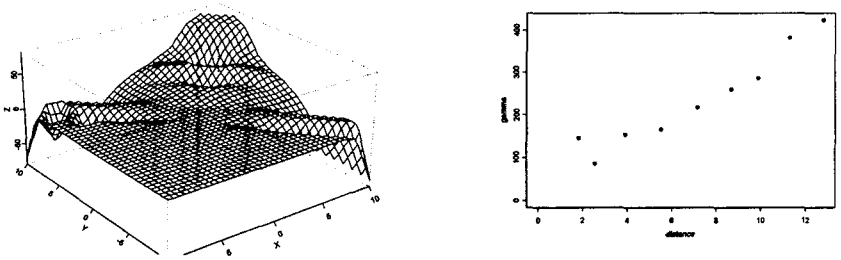
range라고 하며, p_0, p_1, p_2는 각각 절편, 기울기, 그리고 거리에 대한 지수를 나타낸다.

2.2 반변이도 모수 추정을 위한 그래디언트 기반의 비선형 최소제곱법 적용

1장에서도 설명한 바와 같이 반변이도의 모수를 시각적으로 결정하는 것은 주관적이며, 자동화된 프로세스를 구현할 수 없다. 또한, 그래디언트 기반의 비선형 최소제곱법을 적용하는 문제에서 실제 사례로서 수치 최적화를 위한 수학적 함수

$$z = x_1 \times \sin(\exp(x_1 + x_2)) \times x_2 \times \sin(\exp(x_1 + x_2)) \quad (1)$$

를 고려해 보자. [그림 1]의 (a)는 변수 x_1과 x_2에 대하여 구간 $[-10, 10]$에서 등간격으로 12등분하여 식 (1)을 나타낸 그림이며, 이에 대한 실험적 반변이도는 (b)와 같다. [표 2]는 기존의 반변이도 모수 추정 방법인 그래디언트 기반의 비선형 최소제곱법을 적용하기 위해 통계 패키지인 S-plus 2000의 nls 함수를 적용한 결과이다. 반변이도 모형으로서 spherical 모형을 선택하였고 초기 nugget, sill, range값은 시각적인 경험에 의한 값으로, 각각 10, 500, 10을 입력하였다. [표 2]에서 nls함수 수행 결과 에러 메시지를 제공하고 있음을 알 수 있다. 또한, 위의 초기값 이외에 $5 \leq \text{nugget} \leq 15$, $300 \leq \text{sill} \leq 700$, $8 \leq \text{range} \leq 12$에 속하는 몇 개의 값을 이용하여 추정한 결과 모두 발산하는 것을 알 수 있고, 궁극적으로 수렴하지 않는다고 판단된다. 이를 통해, 반변이도의 모수 추정에 있어, 그래디언트 방법을 적용하는 것은 초기값 문제 등으로 원하는 결과를 얻지 못하는 경우가 발생할 수도 있음을 알 수 있다.



(a) 수학적 함수 (b) 실험적 반변이도

[그림 1] 수학적 함수와 실험적 반변이도

[표 2] 반변이도 모수 추정에서의 비선형 최소제곱법 오류

```
> nls(meangamma~a+b*((3/2)*(meandist/c)-(1/2)*(meandist/c)^3),data=vario,
  start=list(a=10, b=500, c=10))
  Error in nls(meangamma ~ a + b * ((3/2) * (meandist/c) - (1/2) *
(meandist/c)..: step factor reduced below minimum
```

3. 유전자 알고리즘을 적용한 반변이도의 모수 추정

이 장에서는 반변이도의 3가지 모수에 대하여 좋은 추정량을 선택하기 위한 방법으로서 유전자 알고리즘에 대하여 소개하고, 본 연구의 목적이라 할 수 있는, 실험적 반변이도에 대하여 오차제곱합(sum of squared error)을 최소로 하는 반변이도의 모수를 추정하고자 한다.

3.1 유전자 알고리즘의 소개

유전자 알고리즘은 자연 생태계에서의 진화(evolution)과정, 즉 자연선별과 유전법칙을 모방한 탐색적 알고리즘으로, 해공간이 방대하여 해를 찾는 것이 어려운 최적화 문제에서 주어진 시간 안에 최적에 가까운 해를 찾기 위하여 사용되는 방법이다. 특히, 복잡한 해공간의 탐색성능이 우수하여 변수와 제약이 많은 대형 수리문제를 해결하는 데에 적합하며, 모형에 대한 유연성이 높아 제약 첨가나 목적함수의 변경이 용이하다는 장점을 가진다. 수행과정은 [표 3]과 같다.

[표 3]을 좀더 자세히 설명하면, 우선 해를 하나의 염색체(chromosome)라 하고 여러 개의 초기해로 해집단(population)을 구성한다 [initialization]. 해집단 내의 개체에 대한 평가를 수행하고 [evaluation], 이들 중 일부를 다음 세대를 구성하기 위한 개체로 선택한다. 다음 세대를 구성하기 위해 개체를 선택할 때에는 평가치가 높은 개체들이 높은 확률로 선택되어지도록 하며 [selection], 새로운 개체를 만들어 내기 위한 유전연산으로는 교차[crossover] 및 돌연변이 [mutation] 연산자를 사용한다.

이러한 연산자들을 반복 시행하면서 평가치가 높은 개체들을 선택해 나가면 세대가 진행될수록 평가치가 높은 해집단을 얻을 수 있게 된다 [evaluation]. 해집단 내에 일정한 점수 이상의 평가치를 갖는 개체가 생성되거나, 또는 일정한 세대가 진행되어도 더 나은 개체가 생성되지 않으면 알고리즘을 종료하게 된다 [end] (김여근 등, 1997).

[표 3] 유전자 알고리즘의 수행과정

```
t=0
P(t) initialization
P(t) evaluation
while(do not satisfy the termination condition) do
        t=t+1
        P(t) selection, crossover, mutation
        P(t) evaluation
end
```

3.2 반변이도 모수 추정을 위한 유전자 알고리즘의 적용

반변이도 모수 추정을 위한 유전자 알고리즘의 적용에 대한 [그림 2]를 자세히 설명하면 다음과 같다. 데이터에 대한 이론적 반변이도를 결정하기 위해서는 실험적 반변이도를 산출한다. 우선 lag의 수 $nlag$를 결정하고 [최대거리로부터 lag의 길이 결정], 관측점들 사이에서의 거리 $|\boldsymbol{h}|$를

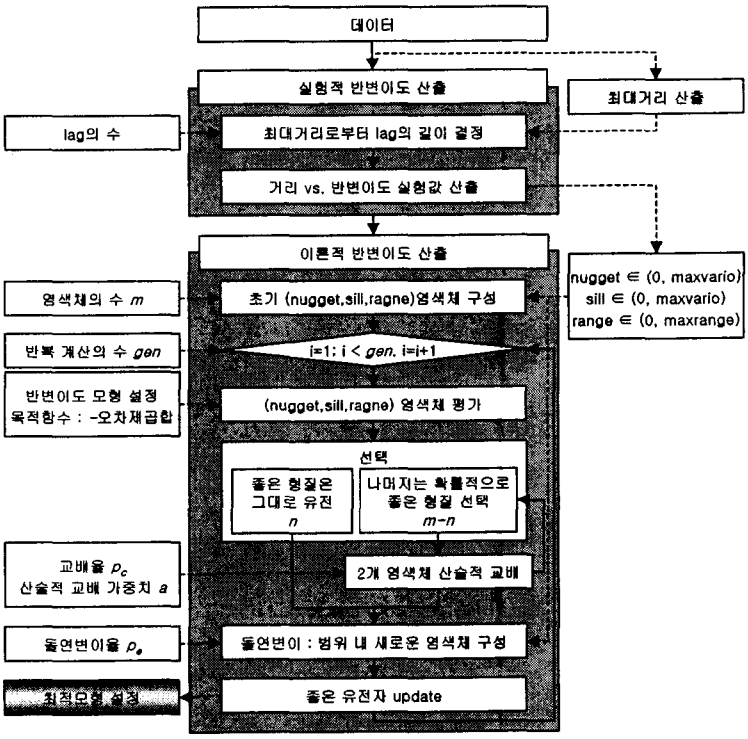
산출한다. 이때, 최대 거리 $|\boldsymbol{h}|_{\max}$에 대하여 실험적 반변이도의 산출범위는 $|\boldsymbol{h}|_{\max}/2$으로 한다. 다음으로, 산출범위를 $nlag$로 나누어 등분한 지점에서의 실험적 반변이도는 k번째 lag를 중심으로 구간

$$k\times\frac{|\boldsymbol{h}|_{\max}}{2\times nlag}-\frac{1}{2}\frac{|\boldsymbol{h}|_{\max}}{2\times nlag}<\boldsymbol{h}<k\times\frac{|\boldsymbol{h}|_{\max}}{2\times nlag}+\frac{1}{2}\frac{|\boldsymbol{h}|_{\max}}{2\times nlag},$$

내의 점들에 대하여 평균 반변이도와 평균 거리를 산출한다 [거리 vs. 반변이도 실험값 산출].

다음으로 산출된 실험적 반변이도와 유전자 알고리즘을 이용하여 이론적 반변이도의 3가지 모수를 결정하게 된다. 유전자 알고리즘에서 사용할 내부적인 숫자의 표현방법은 정밀도 문제에서 실수가 이진수보다 우수하기 때문에 실수를 취하여 분석하기로 한다. 또한, 컴퓨팅 시간을 고려하면서 염색체의 수 m과 세대수 gen을 결정한다. 즉, 염색체의 수 m은 부모세대 염색체인 반변이도의 3가지 모수(nugget, sill, range)를 후보로 갖는 후보 염색체의 개수이며 세대수 gen은 진화과정, 즉, 반복계산의 수를 의미한다 [초기 (nugget, sill, range) 염색체 구성]. 목적함수는 오차제곱합으로서 평가한다. 유전자 알고리즘은 함수의 최대화를 목적으로 하고 있으므로, 오차제곱합의 최소화를 위하여 마이너스(-)가 필요하다 [(nugget, sill, range)염색체 평가]. 선택 단계에서는 임의의 개수는 좋은 형질의 유전자, 즉 오차제곱합이 낮은 유전자를 그대로 선택하고 나머지에 대하여는 확률바퀴 방법을 적용하였다 [선택]. 교배연산을 위해서는 산술적 교배 방법을 적용하였다. 이는 두 부모 g_i^t와 g_j^t가 교배될 때, 그 자손세대는 1차 결합(linear combination)

$$g_k^{t+1}=ag_i^t+(1-a)g_j^t \text{ if } r(0,1)\le p_c,\ g_i^t>g_j^t$$



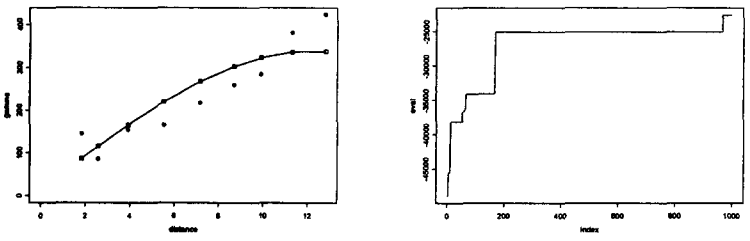
[그림 2] 반변이도의 3가지 모수 추정

으로 정의된다. 여기서 g_k^t는 t세대의 임의의 k번째 염색체이고, a는 가중치이다. $r(0, 1)$은 (0,1)에서 균일분포를 따르는 난수이며, p_c는 교배율이다 [2개 염색체 산출적 교배]. 또한, 돌연변이 연산자는 균등 돌연변이를 적용하였으며, 이때, t세대의 임의의 염색체 $g_i^t = \langle g_1, \cdots, g_m \rangle$ 내의 각 원소 g_i, $i = 1, \cdots, m$는 같은 돌연변이 확률 p_e를 갖는다. 또한, 돌연변이 연산에서는 전역적 최적화된 점들을 찾기 위해 돌연변이율에 의해 선택된 점의 근방에 난수를 발생시키는 가우스돌연변이 방식을 적용하였다 [돌연변이]. 이와 같은 과정을 계속 업데이트 [좋은 유전자 update] 하여 최종 리턴시킨다 [최적모형 설정].

3.3 유전자 알고리즘을 이용한 반변이도 모수의 추정

본 절에서는 식 (1)의 수학적 함수에 대한 반변이도 모수 추정을 위해 유전자 알고리즘을 적용하고자 한다. 이를 위해 [그림 2]에서 제안된 순서도를 적용한다.

이론적 모형은 spherical 모형으로 하였다. 염색체의 수는 $m = 20$으로 하고, 세대수는 1000으로 하였다. 선택 과정에서는 좋은 유전자는 그대로 유지하고, 나머지는 확률적으로 좋은 유전자를 선택하였다. 산술적 교배 연산자의 가중치는 $a = 0.7$, 교배율은 $p_c = 0.3$, 그리고 돌연변이율은 $p_e = 0.1$로 하였다. [그림 3]은 유전자 알고리즘을 이용하여 반변이도의 모수를 추정한 결과로서 (a)는 spherical 모형을 적용한 최소제곱 오차추정선을 나타내며, (b)는 세대별 진화과정을 보여주고 있다. [표 4]는 유전자 알고리즘을 통해 얻은 nugget, sill, range의 최적값과 오차제곱합을 보여주고 있다. [표 4]의 값과 [표 2]에서 사용한 초기값을 비교해 볼 때, 비록 초기값으로 [표 4]에서 구해진 값과 비슷한 값을 사용하더라도 그래디언트 기반의 방법은 발산하는 것을 알 수 있고, 적절한 해를 제공하지 못하는 경우가 있다는 것을 알 수 있다. 이를 통해, 유전자 알고리즘은 반변이도 모수를 추정하는 데 있어 효과적으로 적용이 가능하였으며, 그래디언트 방법에 대한 경쟁력 있는 대안임을 알 수 있었다.



(a) 최소제곱 추정선 (b) 세대별 진화과정

[그림 3] 유전자 알고리즘을 이용한 반변이도의 모수 추정

[표 4] 이론적 반변이도 모수와 최소제곱 추정값

nugget	sill	range	sum of squared error
14.5339	324.3048	12.1792	22691.4231

4. 범용크리깅 (Ordinary Kriging)

범용크리깅은 모평균을 모르지만 고정되어 있다는 전제에서 사용이 가능하다. 또한, 가장 많이 적용되고 있는 크리깅 방법 중 하나이다. 이에 대한 가정과 새로운 지점 s_0에서의 크리깅 예측값의 가정은 다음과 같다.

$$Z(s) = m + \delta(s), \quad s \in D, m \in R$$

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

이때, 불편성의 조건

$$\sum_{i=1}^{n} \lambda_i = 1$$

에 의하여 범용크리깅의 평균제곱오차(mean squared error)

$$\sigma_{OK}^2 = E[Z(s_0) - p(Z; s_0)]^2$$

를 최소화하는 $\lambda_1, \cdots, \lambda_n$을 구하고자 한다. 따라서, 라그랑지 승수 μ 에 대하여

$$E[Z(s_0) - \sum_{i=1}^{n} \lambda_i Z(s_i)]^2 - 2\mu\left(\sum_{i=1}^{n} \lambda_i - 1\right) \tag{2}$$

을 최소화하는 $\lambda_1, \cdots, \lambda_n$과 μ를 구해야 한다. 식 (2)는

$$-\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(s_i - s_j) + 2\sum_{i=1}^{n} \lambda_i \gamma(s_0 - s_i) - 2\mu\left(\sum_{i=1}^{n} \lambda_i - 1\right)$$

이 되며, $\lambda_i, \cdots, \lambda_n$, 그리고 μ 에 의하여 편미분하여 0이 되도록 하면, 각각

$$-\sum_{j=1}^{n} \lambda_j \gamma(s_i - s_j) + \gamma(s_0 - s_i) - \mu = 0, \quad \sum_{i=1}^{n} \lambda_i = 1$$

이 된다. 범용크리깅의 최량선형비편향추정량(best linear unbiased estimate)을 추정하기 위해 크리깅 계수 $\lambda' = (\lambda_1, \cdots, \lambda_n)$를 산출하면,

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

이다. 여기에서 1_n는 길이가 n이고 모든 원소가 1인 벡터이고, Γ는 (i, j)번째 원소가 $\gamma(s_i - s_j)$인 $n \times n$ 행렬이며, $\gamma = (\gamma(s_0 - s_1), \cdots, \gamma(s_0 - s_n))'$이다. 이와 같이 산출된 크리깅 계수는 평균제곱예측오차를 최소화시키며, 이를 크리깅 분산이라 부르고 다음과 같다.

$$\sigma_{OK}^2 = \gamma' \Gamma^{-1} \gamma - \frac{(1 - 1'_n \Gamma^{-1}\gamma)^2}{1'_n \Gamma^{-1} 1_n}.$$

또한, 공간변수 $Z(s_0)$가 정규분포를 따른다고 가정하면, $Z(s_0)$의 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$(\hat{Z}(s_0) - z_{\alpha/2}\sigma_{OK}(s_0), \ \hat{Z}(s_0) + z_{\alpha/2}\sigma_{OK}(s_0)).$$

5. 실증분석

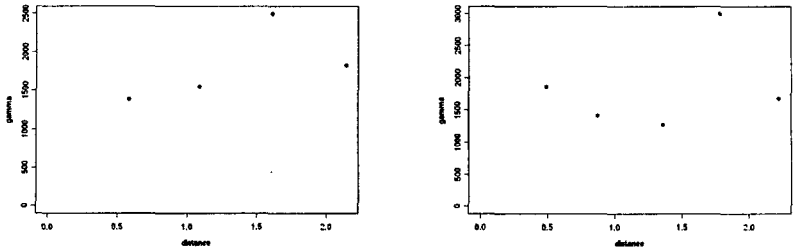
본 연구에 사용된 강우량 데이터는 통계청 홈페이지 (http://www.nso.go.kr/)의 통계 DB인 KOSIS에서 제공하고 있는 강우량에 관한 자료와 기상청에서 발간하고 있는 기상연보(2001)를 활용하였다. KOSIS에서 제공하고 있는 수량 데이터는 강우량 관측 지점명과 월별 강우량에 대하여 제공하고 있으며, 기상연보에서 관측지점별 경도와 위도를 제공하고 있다. 본 연구에서는 과거 13년(1990 ~ 2002) 동안 강우량이 가장 많았던 8월의 데이터에 대한 평균값을 적용하였다. 여기에서, 울진 데이터는 결측치로 인하여 2000년 이후에 측정한 결과를 이용하였고, 울릉도 데이터는 이상치로서 제외하였다. 이에 대한 자료는 [표 5]와 같다. 본 연구에서의 크리깅 방법 및 유전자 알고리즘은 S-Plus의 사용자 함수를 이용하여 구현하였다.

5.1 강우량 자료의 실험적 반변이도 산출

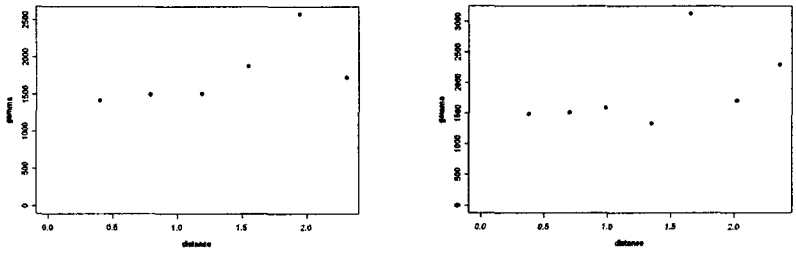
실험적 반변이도는 관측 지점으로부터 얻은 13년간의 평균값을 이용하여 산출하였는데, 우선 24개의 관측된 지점에서 모든 가능한 조합을 산출하였다. 다음으로, 주어진 lag의 수에 따라 관측된

[표 5] 과거 13년(1990~2002년)간 8월달 평균 강우량

관측지점	북위(Lat)	동경(Long)	평균강우량
속초	38.15	128.34	296.22
춘천	37.54	127.44	332.10
강릉	37.45	128.54	332.39
서울	37.34	126.58	431.92
인천	37.28	126.38	333.93
수원	37.16	126.59	347.90
서산	36.46	126.30	340.63
울진	36.59	129.25	293.80
청주	36.38	127.27	317.66
대전	36.22	127.22	337.79
추풍령	36.13	128.00	264.35
포항	36.02	129.23	255.25
군산	35.59	126.42	282.32
대구	35.53	128.37	242.51
전주	35.49	127.09	261.48
울산	35.33	129.19	292.18
광주	35.10	126.54	320.62
부산	35.06	129.02	334.04
통영	34.51	128.26	289.09
목포	34.49	126.23	206.80
여수	34.44	127.45	280.13
제주	33.31	126.32	281.98
서귀포	33.15	126.34	335.72
진주	35.12	128.07	331.12



(a) lag = 5 (b) lag = 6



(c) lag = 7 (d) lag = 8

[그림 4] lag에 따른 강우량 측정치의 실험적 반변이도

[표 6] lag=7에서의 실험적 반변이도

meandist	meangamma	npoint
0.3963	1414.683	9
0.7913	1495.487	31
1.1863	1503.542	27
1.5465	1874.956	38
1.9463	2570.954	38
2.3089	1722.566	36

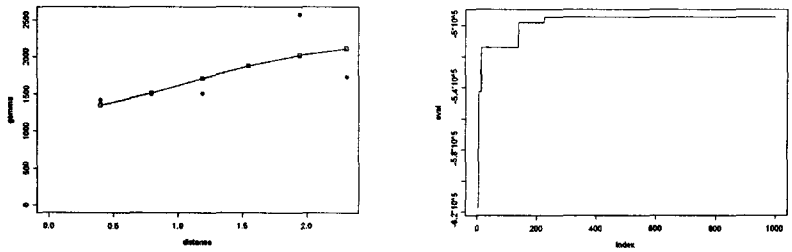
거리의 평균과 산출된 반변이도의 평균값을 구하였다. 실제로 lag의 수를 5, 6, 7, 8개로 정했을 때의 거리 vs. 반변이도의 그림은 [그림 4]와 같다. 본 연구에서는 오차가 적을 것으로 판단된 lag=7인 결과를 이용하여 반변이도의 모수를 결정하기로 하였으며, 각 lag 별 평균거리와 평균 반변이도 값은 [표 6]에서와 같다. 여기에서, meandist는 평균 거리이며 meangamma는 반변이도의 평균값 그리고 npoint는 lag에서의 점들의 개수이다.

5.2 유전자 알고리즘을 이용한 강우량 자료의 반변이도 모수 추정

이론적 모형은 Gaussian 모형으로 하였다. 컴퓨팅 시간을 고려하여 염색체의 수는 $m = 20$으로 하고, 세대수를 1000으로 하였다. 선택 과정에서는 염색체 중 임의의 개수는 우수한 형질의 염색체를 그대로 유전시키고, 나머지는 확률적으로 선택하도록 하였다. 산술적 교배 연산자의 가중치

[표 7] 이론적 반변이도 모수와 최소제곱 추정값

nugget	sill	range	sum of squared error
1255.016	880.7061	1.35454	490508



(a) 최소제곱오차 추정선 (b) 세대별 진화과정

[그림 5] 유전자 알고리즘을 적용한 반변이도 모수 추정

는 $a = 0.7$, 교배율은 $p_c = 0.3$, 그리고 돌연변이율은 $p_e = 0.1$로 하였다. 분석결과의 평가를 위하여 산출된 평균거리에서의 평균 반변이도와 추정량의 차의 제곱값을 최소화되도록 하는 모수를 산출하였다. 반변이도의 3가지 모수와 오차의 제곱합은 [표 7]에서와 같다. [그림 5]에서는 유전자 알고리즘을 이용한 최소제곱 추정선(a)과 세대별 진화과정(b)을 보여주고 있다.

5.3 그래디언트 기반의 비선형 최소제곱법을 이용한 강우량 자료의 반변이도 모수 추정

[표 8]은 그래디언트 기반의 비선형 최소제곱법을 이용하여 반변이도의 모수를 추정한 결과이다. [그림 4]의 (c)에서 시각적 경험에 의해 결정한 초기값으로서 nugget과 sill 값으로 1000을 입력하고 range의 값에 1을 입력한 결과, nugget은 1262.914, sill은 871.2803, 그리고 range는 1.356886을, 오차제곱합은 490433.3의 값을 얻었다. [표 7]과 [표 8]을 비교해 보면, 두 방법에서 얻어진 추정값에 큰 차이를 보이지 않고 있음을 알 수 있다. 따라서, 유전자 알고리즘을 적용한 방법은 그래디언트 기반의 방법에 대한 경쟁력 있는 대안이 될 뿐 아니라, 발산문제도 해결할 수 있는 이점을 제공한다고 할 수 있다.

[표 8] 그래디언트 기반의 비선형 최소제곱법을 이용한 반변이도 모수 추정

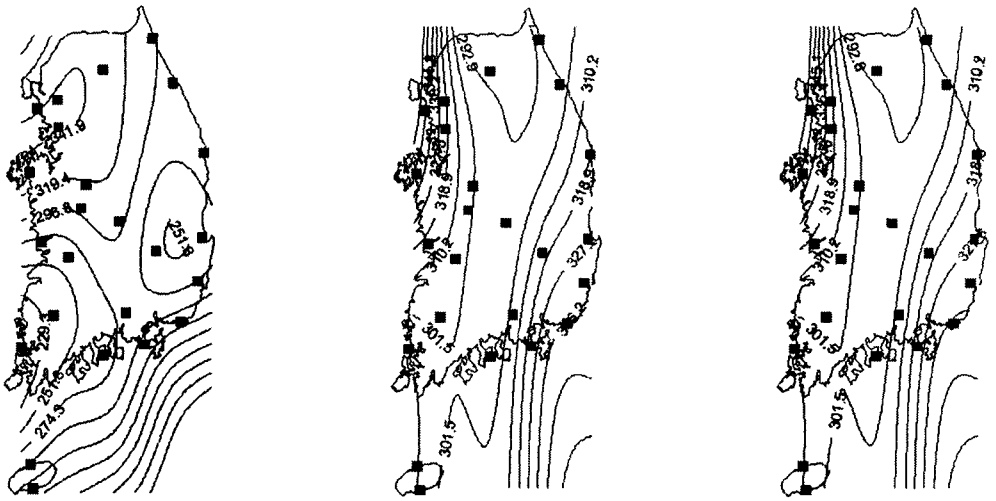
```
> nls(meangamma~nugget+sill*(1-exp(-(meandist/range)^2)),data=ttt,start=list
(nugget=1000,sill=1000,range=1))
Residual sum of squares : 490433.3
parameters:
  nugget      sill     range
 1262.914 871.2803 1.356886
formula: meangamma ~ nugget+sill*(1-exp(-(meandist/range)^2))
6 observations
```

5.4 범용크리깅 추정량

본 절에서는 크리깅 방법의 효과성을 살펴보기 위해, [표 5]에서의 강우량 데이터에서 관측지점 하나를 제외하고 나머지 관측지점으로 제외된 하나의 지점을 추정하는 cross-validataion 방법을 이용하였다. [표 5]에서의 데이터를 [표 7]과 [표 8]에서의 유전자 알고리즘과 그래디언트 기반의 비선형 최소제곱법을 적용한 반변이도 모수 추정값을 이용한 결과, [표 9]에서와 같이 각 지역별 크리깅 추정값을 얻었으며, 크리깅 분산을 계산함으로서 얻어진 크리깅 표준편차, 95% 하한값 그리고 95% 상한값이 제시되어 있다. [표 9]를 보면, 5.3절에서와 마찬가지로 두 방법에서의 결과값의 차이는 크지 않음을 알 수 있다. 또한, [그림 6]과 같이 실제 강우량과 유전자 알고리즘, 그리고 그래디언트 기반의 비선형 최소제곱법을 적용하여 크리깅 추정량을 시각화 한 결과, 두 가지 추정 방법에서는 유사한 등고선 형태를 보이고 있으며, 실제 강우량에서와 같이, 서북부와 동남부에 많은 강우량을, 그리고 동북부와 서남부에서는 적은 강우량을 나타내는 것으로 보인다. [표 5]와 [표 9]의 실제 강우량과 크리깅 추정 강우량을 비교해 보면, 서울에서는 상한을, 목포에서는 하한을 벗어난 것으로 나타났는데, 이는 실제 강우량에서 타 지역에 비해 상당히 큰 편차를 보이고 있기 때문으로 사료된다.

[표 9] 유전자 알고리즘과 그래디언트 기반의 비선형 최소제곱법으로 추정된 반변이도 모수 추정값을 적용한 관측지점 별 크리깅 추정값

	유전자 알고리즘에 의한 크리깅 추정값				그래디언트 기반의 비선형 최소제곱법에 의한 크리깅 추정값			
관측지점	추정값	표준편차	95%하한	95%상한	추정값	표준편차	95%하한	95%상한
속초	320.96	43.19	236.32	405.60	320.91	43.24	236.16	405.66
춘천	336.75	42.24	253.96	419.53	336.57	42.29	253.69	419.45
강릉	299.60	42.39	216.51	382.68	299.66	42.45	216.45	382.86
서울	330.22	39.95	251.93	408.52	330.06	40.04	251.59	408.54
인천	353.82	40.60	274.24	433.40	353.49	40.69	273.74	433.23
수원	348.57	39.20	274.74	425.40	348.27	39.30	271.24	425.29
서산	325.70	41.03	245.27	406.12	325.64	41.10	245.09	406.19
울진	293.20	43.73	207.48	378.91	293.28	43.77	207.50	379.06
청주	317.89	39.54	240.39	395.38	317.83	39.62	240.18	395.49
대전	302.50	39.46	225.74	379.26	302.56	39.26	225.62	379.50
추풍령	298.48	40.34	219.41	377.55	298.59	40.42	219.38	377.80
포항	292.11	41.27	211.23	373.00	292.25	41.34	211.22	373.27
군산	301.49	40.49	222.12	380.85	301.54	40.57	222.03	381.05
대구	292.52	40.14	213.84	371.19	292.64	40.21	213.82	371.46
전주	300.40	39.51	222.96	377.84	300.40	39.60	222.79	378.00
울산	289.90	40.74	210.04	369.76	289.98	40.82	209.96	369.99
광주	266.58	40.63	186.94	346.22	266.90	40.71	187.11	346.69
부산	289.79	40.92	209.58	370.00	289.80	41.00	209.45	370.16
통영	307.18	40.61	227.58	386.78	307.02	40.69	227.27	386.78
목포	303.99	43.28	219.17	388.81	304.00	43.31	219.10	388.89
여수	301.89	41.48	220.59	383.20	301.79	41.55	220.36	383.22
제주	307.02	42.68	223.37	390.67	306.97	42.75	223.17	390.77
서귀포	290.24	43.34	205.29	375.79	290.31	43.41	205.22	375.39
진주	284.99	39.79	207.01	362.98	285.06	39.87	206.91	363.20



(a) 실제 강우량 (b) 그래디언트 기반의 최소제곱법 (c) 유전자 알고리즘

[그림 6] 강우량 등고선 그림

6. 결론

본 연구는 크리깅을 수행하기 위해 유전자 알고리즘을 이용한 반변이도 모수의 추정을 목적으로 하였다. 유전자 알고리즘은 확률적(random)으로 최적화 된 값을 제공할 뿐만 아니라, 보다 정확한 값을 얻기 위해서는 보다 많은 시간이 필요하다는 단점을 갖고 있으며, 그래디언트 기반의 비선형 최소제곱법은 좋은 초기값을 갖고 있을 때 결정적(deterministic)인 최적값을 제공한다는 점에서 장점을 갖고 있다. 그러나, 유전자 알고리즘의 적용에 따른 장점을 요약하면 다음과 같다.

첫째, 기존의 반변이도 모수의 추정은 시각적인 확인 과정을 필요로 하며, 이는 중간단계에서의 검증절차가 요구될 뿐만 아니라, 그 결과에 대한 객관적 신뢰성이 부족하다. 이에 비해, 유전자 알고리즘은 세대가 증가함에 따라 최적화된 결과를 제공함과 동시에 중간단계에서 검증하는 단계가 필요하지 않으므로 자동화 프로세스를 구축할 수 있다.

둘째, 일부 반변이도 모수 추정을 위한 패키지의 경우에는 그래디언트 기반의 비선형 최소제곱법을 적용하고 있으나, 이는 초기치 문제 등으로 인하여 발산하는 경우, 사용자가 원하는 결과를 제공하지 못하는 경우가 발생하기도 한다. 이에 비해, 유전자 알고리즘은 사용자가 계산 시간을 미리 지정할 수 있기 때문에 유한한 시간에서의 전역적으로 최적화된 값을 찾을 수 있다.

따라서, 본 연구에서는 그래디언트 기반의 비선형 최소제곱법의 경쟁력 있는 대안으로서 유전자 알고리즘의 적용방법과 그 가능성을 제시하였다.

참고 문헌

[1] 통계청 홈페이지, http://www.nso.go.kr/

[2] 「기상연보」 (2001). 기상청.

[3] 김여근, 윤복식, 이상복 (1997). 「메타휴리스틱」. 영지문화사

[4] 오은선 (2002). 「강수량 분포에 적용되는 Kappa분포의 모수추정」, 전남대학교 석사학위논문.

[5] 이동률 (2002). 기후변동과 확률강우량의 변화, 「건설기술정보」.

[6] 이재형, 유양규, 정재성 (2002). 강우관측망 최적설계 기법 개선에 관한 연구, 「대한토목학회 논문집」, Vol.22(5;B), pp.671-677.

[7] 최종근, 정대인 (2002). Simulated Annealing 기법을 이용한 실험적 베리오그램의 모델링, 「한국지하수토양환경학회 추계학회 논문집」.

[8] Bacchi, B. and Kottegoda, N.T. (1995). Identification and calibration of spatial correlation patterns of rain fall, *Journal of Hydrology*, Vol.165, pp.311-348.

[9] Cressie, N. (1991). *Statistics for spatial data*, John Wiley & Sons, New York.

[10] Deceneiere, E., Fouquet, C. and Meyer, F. (1998). Applications of Kriging to image sequence cooling, Signal processing; *Image communication*, Vol.13, pp.227-249.

[11] Journel, A.G. and Huijbregts, C.J. (1978). *Mining geostatistics*, Academic Press, London.

[12] Krige, D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand, *Journal of the Chemical, Metallurgical and mining society of south africa*, Vol.52, pp.119-139.

[13] Lee, E.K., (2003). A space model to annual rainfall in South Korea, *The Korean Communications in Statistics*, Vol.10(2), pp.445-456.

[14] Matheron, G. (1969). Le Krigeage universel, Cahiers du Centre de Morphologie Mathematique, *Fontainebleau*, Vol.1, France.

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

[16] Matheron, G. and Huijbregts, C.J. (1971). Universal Kriging, *In proceedings of ninth international symposium on techniques for decision-making in the mineral industry*, The Canadian Institute of Mining and Metallurgy, Vol.12, pp.159-169.

[17] Namkung, P., Jang. J.H. and Hong, T.K. (2003). Spatial data analysis using the Kriging method, *The Korean Communications in Statistics*, Vol.10(2), pp.423-432.

[18] Piazza. et al. (1983). The making and testing of geographic gene-frequency maps. *Biometrics*, Vol.37, pp.635-659.

[19] Ryu, J.S., Kim, M.S., Cha, K.J, Lee, T.H. and Choi, D.H. (2002). Kriging interpolation methods in geostatistics and DACE Model, *KSME international journal*, Vol.16(5), pp.619-632.

[20] Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, P.H. (1989). Design and analysis of computer experiments, *Statistical Science*, Vol.4(4), pp.409-435.

[21] Stephen, P.K., Silvia, C.V., Tamre, P.C. and Alic, A.C. (1996). *S+ spatialstats user's Manual*. MathSoft, Inc.

[22] Tukey, J.W. (1977). *Exploratory data analysis*, Addison-Wesley, Reading, MA.

[23] Webster, R. (1985). Quantitative spatial analysis of soil in the field, *Advances is soil science*, Vol.3, B.A., Stewart(ed.), New York, Springer Verlag, pp.1-70.

[2004년 3월 접수, 2004년 6월 채택]