The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 835-844

A Study on the Selection of Variogram Using Spatial Correlation[1)]

Key-Il Shin[2)], Ki-Jung Back, Jin-Mo Park[3)]

Abstract

A difficulty in spatial data analysis is to choose a suitable theoretical variogram. Generally mean squares error(MSE) is used as a criterion of selection. However researchers encounter the case that the values of MSE are almost the same whereas the estimates of parameters are different. In this case, the selection criterion based on MSE should take into account the parameter estimates. In this paper we study on the method of selecting a variogram using spatial correlation.

Keyword : Moran's I, Geary's C, Theoretical variogram.

1. 서론

최근 컴퓨터의 발달은 공간 통계분석의 이론적 어려움을 극복하게 하였다. 그러나 공간 통계분석의 많은 부분을 차지하는 Geostatistic 자료 분석에서 오차의 상관관계 모형화는 아직도 명확하게 설명되지 않고 있다. 일반적으로 오차의 상관관계 모형화는 변이도(Variogram)를 이용한다. 변이도를 이용한 오차의 모형화는 크게 두 단계로 나누어진다. 첫 단계는 변이도의 추정이다. 특히 자료에 이상점이 존재할 경우 변이도의 추정값은 많은 영향을 받게 된다. 이를 극복하기 위한 연구는 Cressie와 Hawkins(1980), Hawkins와 Cressie(1984), Genton(1998) 그리고 Mugglestone 등(2000)이 있다. 두 번째 단계가 변이도의 추정값에 이론적인 변이도를 적합시켜 이론적 변이도에 포함된 모수를 추정하는 것이다. 그러나 알려진 여러 이론적 변이도 중에서 자료를 잘 설명할 수 있는 변이도를 선택하는 것은 쉬운 일이 아니다. 일반적으로 많이 사용되는 이론적 변이도는 구형 변이도(Spherical variogram), 지수 변이도(Exponential variogram) 그리고 가우시안 변이도(Gaussian variogram)이다. 이 중에서 평균제곱오차(Mean Square Error: MSE)를 최소로 하는 변이도를 선택하는 것이 일반적이다. 그러나 실제 자료 분석에서 추정된 모수 추정값은 선택된 모형에 따라 차이를 보인다. 자료 분석의 목적에 따라서는 선택된 모형뿐만 아니라 모수 추정값 자체가 중요할 경우가 있다. 예를 들어 온실속의 벌레 밀도를 공간 통계학을 이용하여 분석할 때 어느

1) This research was supported by the research fund of Hankuk University of Foreign Studies, 2003.

2) Professor, Department of Statistics, Hankuk University of Foreign Studies, Yongin, Gyonggi, 449-791. E-mail: keyshin@stat.hufs.ac.kr

3) Graduate student, Department of Statistics, Hankuk University of Foreign Studies, Yongin, Gyonggi, 449-791.

거리까지 오차의 상관관계가 있는 지는 생물학적으로 매우 중요하다. 오차의 상관관계가 존재하는 최대거리는 변이도에 포함된 모수 중에서 범위(Range)에 해당되며, 따라서 범위 추정값 값 자체가 중요한 의미를 가질 수 있다. 이에 관한 내용은 Cho 등(2002)과 Park 등(2002)을 참조하기 바란다.

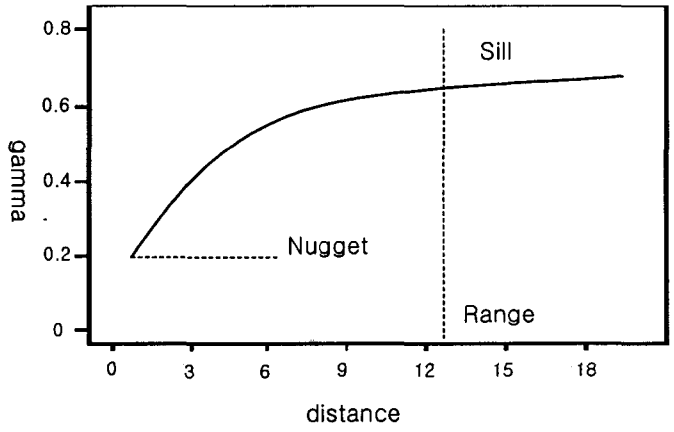
본 논문은 2절에서 변이도에 포함된 모수인 범위(Range)와 공간 독립 검정통계량으로 사용되는 Moran의 I 와의 관계를 살펴보았고, 3절에서는 충청남도 부여지방의 온실에서 얻어진 온실가루이 자료 분석을 통하여 Moran의 I 값과 Geary의 C 값이 변이도를 선택할 때 어떻게 사용되는지 살펴보았다. 끝으로 4절에 결론이 있다.

2. 변이도와 Moran의 I 와의 관계

변이도는 거리에 따른 공간 상관관계를 규명하기 위해 사용되며, 이를 이용하여 오차의 상관관계 또는 오차의 구조를 찾아내게 된다. 전통적인 변이도(Classical variogram) 추정량은 다음과 같다.

$$2\hat{\gamma}(h)=\frac{1}{N_h}\sum_{i=1}^{N_h}(Z(U_i)-Z(U_i+h))^2$$

여기서 h는 거리, N_h는 거리가 h인 자료 쌍의 수, U_i는 공간상의 위치, 그리고 $Z(U_i)$는 공간 위치 U_i 에서 얻어진 자료값을 나타낸다. 위의 변이도 추정량에 적합시킬 2차 정상성을 만족하는 이론적 변이도의 일반적 형태는 다음과 같다.



1절에서 언급한데로 Geostatistic 자료분석에 많이 사용되는 이론적인 변이도는 다음과 같다.

1) 구형 변이도 $(d=1,2,3)$

$$\gamma(h,\theta)=\theta_1+\theta_2\left(\frac{3}{2}\frac{h}{\theta_3}-\frac{1}{2}\left(\frac{h}{\theta_3}\right)^3\right),\quad 0<h<\theta_3$$

$$\gamma(h,\theta)=\theta_1+\theta_2, h>\theta_3$$

2) 지수 변이도 ($d\geq 1$)

$$\gamma(h,\theta)=\theta_1+\theta_2\left(1-\exp\left(-\frac{h}{\theta_3}\right)\right),\quad 0<h$$

3) 가우시안 변이도 ($d\geq 1$)

$$\gamma(h,\theta)=\theta_1+\theta_2\left(1-\exp\left(-\frac{h^2}{\theta_3^2}\right)\right),\quad 0<h$$

위의 세 변이도는 모두 이차 정상성을 만족하며 범위(Range)와 문턱(Sill)이 존재한다. 여기서 d는 자료의 차원을 의미하고, θ_1 은 뭉치(Nugget)를 $\theta_1+\theta_2$ 는 문턱(Sill)을 나타낸다. 구형 변이도(Spherical variogram)의 경우 범위는 θ_3 이나 가우시안 변이도와 지수 변이도의 경우는 θ_3 를 이용하여 근사적으로 범위(Range)를 구한다. 본 논문에서는 지수 변이도의 범위는 $3\theta_3$ 를 가우시안 변이도의 범위는 $\sqrt{3}\theta_3$ 를 사용하였다. 이에 관한 내용은 Goovaerts(1997)을 참조하기 바란다.

변이도에서 범위(Range)는 문턱(Sill)이 얻어질 때의 거리를 나타내므로 범위보다 큰 거리에서는 두 변수의 상관관계가 존재하지 않게 된다. 따라서 범위는 공간독립 검정 통계량과 깊은 관계가 있음을 알 수 있다.

Lattice자료 분석에서 공간 상관관계가 있는지를 검정하기 위해 일반적으로 사용하는 검정 통계량은 Moran의 I 값과 Geary의 C 값이다. 먼저 자료 $Z_1, Z_2, \cdots, Z_n$가 연속이라 하자. 여기서 Z_i는 i번째 지역에서 얻어진 자료이다.

그리고

δ_{ij}=1, 만약 j번째 자료가 i번째 자료와 이웃이면

δ_{ij}=0, 만약 j번째 자료가 i번째 자료와 이웃이 아니면

이라 하자. 또한 $T=1/2\sum_{ij}\delta_{ij}$라 하자. 그러면 Moran의 I와 Geary의 C 통계량은 다음과 같이 정의된다.

Moran의 I

$$I=\frac{n}{2T}\frac{\sum_i\sum_j\delta_{ij}(Z_i-\bar{Z})(Z_j-\bar{Z})}{\sum_i(Z_i-\bar{Z})^2} \tag{2.1}$$

Geary의 C

$$C=\frac{n-1}{4T}\frac{\sum_i\sum_j\delta_{ij}(Z_i-Z_j)^2}{\sum_i(Z_i-\bar{Z})^2} \tag{2.2}$$

위의 두 통계량에 관한 자세한 내용은 Cressie (1993)을 살펴보기 바란다.
Moran의 I 값과 Geary의 C 값은 정의된 이웃정보(Neighbor Information), δ_{ij},에 의존하며 이러한 이웃정보는 거리에 따라 정의 될 수 있다. 따라서 거리를 이용한 Moran의 I 값과 Geary의 C 값을 이용하면 자료가 어느 거리까지 상관관계가 있는지를 구할 수 있게 된다. 최근 Diblasi와 Bowman(2001)은 변이도가 공간 상관관계의 유무를 검정하는데 사용 될 수 있다고 주장하였다. 또한 이웃정보 δ_{ij}를 거리에 의해 정의하고 $\rho(h)$를 Correlogram이라 하면 Moran의 I는

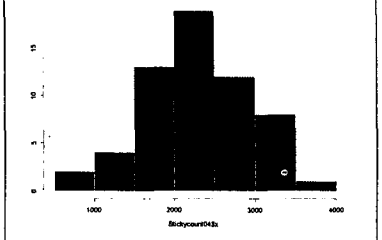
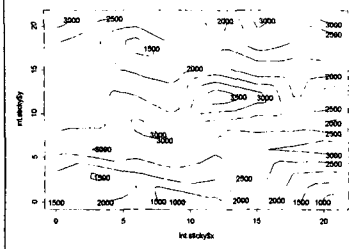
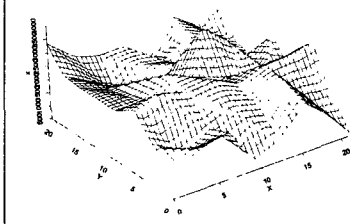
$$I=\frac{n}{2T}N_h\frac{\sum_{i=1}^{N_h}(Z_i-\bar{Z})(Z_j-\bar{Z})/N_h}{\sum_i(Z_i-\bar{Z})^2}=\frac{n}{2T}N_h\,\rho(h)$$

가 되어 Correlogram에 비례하게 된다. 여기서 N_h는 거리 h에 속한 자료 쌍의 개수이다. 또한 Geary의 C 통계량의 분자는 전통적 변이도를 구하는 공식과 일치한다. 따라서 변이도와 Correlogram은 Moran의 I 와 Geary의 C 통계량과 깊은 관계가 있음을 알 수 있으며 공간 상관관계의 한계를 말해주는 변이도의 범위는 Moran의 I 와 Geary의 C를 거리에 따라 구할 때 상관관계가 있는 거리의 최대값과 일치하여야 한다. 따라서 잔차제곱합을 이론적인 변이도의 선택 기준으로 사용하는 것 뿐 아니라 거리를 이용한 Moran의 I 와 Geary의 C의 결과를 이론적 변이도 선택에 반영하면 더 좋은 분석 결과를 얻을 수 있을 것으로 판단된다.

3. 자료 분석

3.1 자료 소개

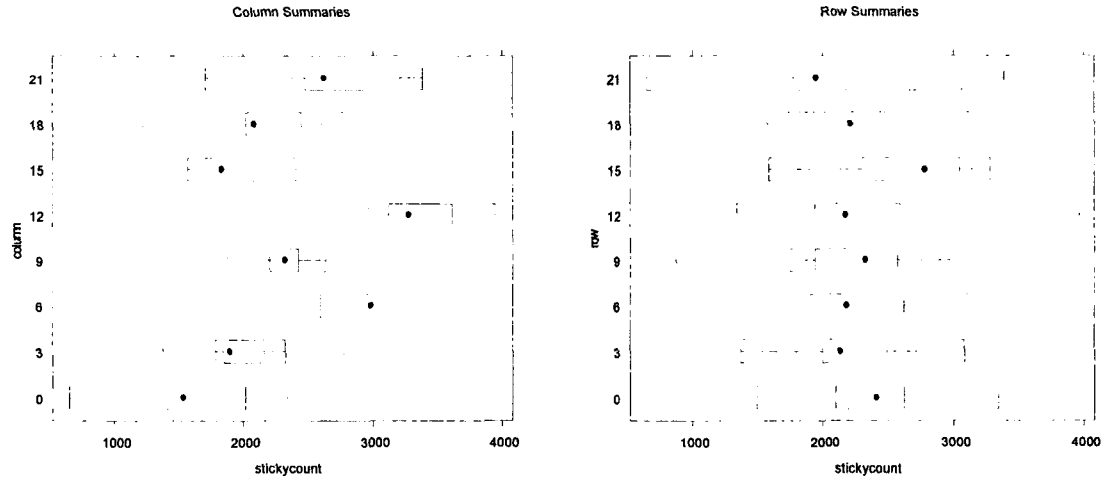
2001년 충청남도 부여지방의 한 온실에서 끈끈이 트랩을 이용하여 잡은 벌레 (온실가루이)의 마리 수를 조사하였다. 끈끈이 트랩은 3m의 간격을 두고 동서 방향으로 8개, 남북 방향으로 8개를 설치하였다. 64개의 자료 중 3개 자료를 제외한 61개 자료가 분석에 사용되었다. 다음은 온실가루이 자료의 히스토그램, Contour plot 그리고 3-D 그림이다. 자료의 히스토그램을 살펴보면 변환 없이 원자료를 사용하여 분석하여도 별 무리가 없어 보인다.

원 자 료	Contour Plot	3-D Perspective Plot
		

< 그림 1 > 자료의 히스토그램과 Contour plot 그리고 3-D plot

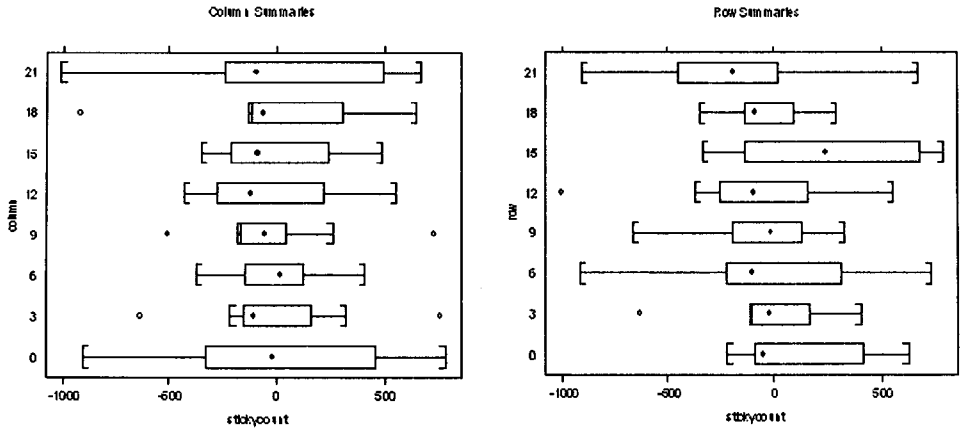
3.2 추세식별

각 행과 열에 따라 추세가 존재하는지 또는 평균의 차이가 있는지 살펴보기 위하여 행별, 열별 Box-Wisker plot을 살펴보았다.



<그림 2> 열과 행에 대한 Box and Whisker plot

<그림 2>를 살펴본 결과 추세는 존재하지 않는 것으로 판단되며, 행의 경우 평균이 일정하다고 판단할 수 있으나 열의 경우는 많은 차이가 있음을 알 수 있다. 열에 대하여 평균이 일정하지 않기 때문에 Median Polish기법을 이용하여 평균을 일정하게 만들어 주었다. 다음이 Median Polish 방법을 열에 적용한 후의 Box-Whisker plot이다.

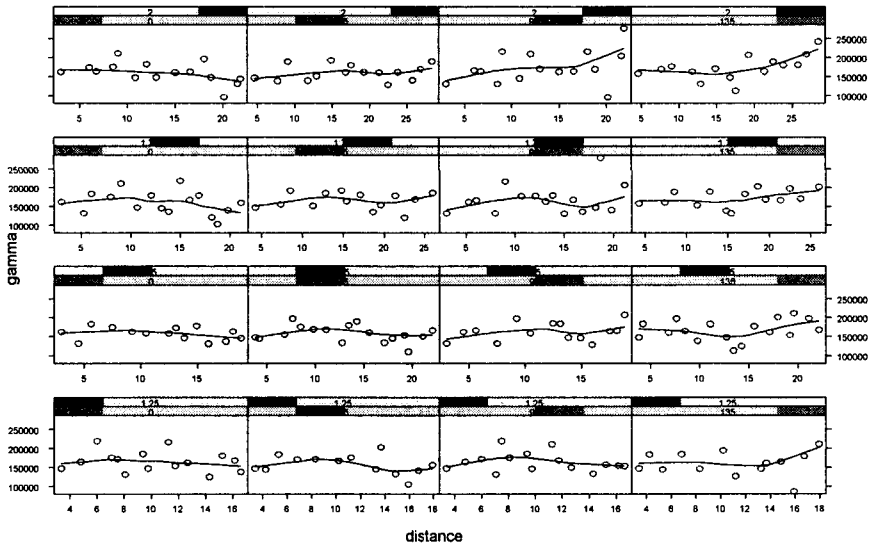


<그림 3> Median Polish후의 열과 행에 대한 Box and Whisker plot

<그림 3>을 살펴보면, Median Polish를 한 결과 열에 대하여 평균이 거의 일정함을 알 수 있다.

3.3 Anisotropy Check

Median Polish을 실시한 원 자료를 이용하여 자료가 Isotropy인지를 확인하기 위하여 방향과 거리의 비를 이용하여 변이도를 그려보았다.



<그림 4> Anisotropy Check를 위한 변이도 그림

위에서 거리의 비는 1.25, 1.5, 1.75 그리고 2가 사용되었으며 방향은 0°, 45°, 90°, 135°가 사

용되었다. 위의 그림을 살펴본 결과 Isotropy를 가정하고 분석을 하여도 별 무리가 없다고 판단할 수 있다.

3.4 Moran의 I 와 Geary의 C

(2.1)과 (2.2)식에 의해 구해진 Moran의 I와 Geary의 C는 다음과 같다.

Distance		Moran의 I	p-value	Geary의 C	p-value
4	4.0	0.2479	0.007361	0.6633	0.001213
	4.2	0.2479	0.007361		
	4.4	0.09764	0.1045		
	4.6	0.09764	0.1045		
	4.8	0.09764	0.1045		
5		0.09764	0.1045	0.824	0.04276
6	6.0	0.09764	0.1045	0.824	0.04276
	6.2			0.8344	0.01672
	6.4			0.8344	0.01672
	6.6			0.8344	0.01672
	6.8			0.9168	0.1938
7		0.01686	0.432	0.9168	0.1938

<표 1> 거리별 Moran의 I와 Geary의 C

<표 1> 에서 알 수 있듯이 Moran의 I 경우, 거리가 4.4m 이상에서 P-값이 0.1보다 크므로 공간 상관관계가 있는 최대 거리는 약 4.0m에서 4.4m까지 인 것으로 판단할 수 있다. 반면에 Geary의 C를 이용할 경우 얻어진 최대 거리는 약 6.4m에서 6.8m까지로 나타났다.

3.5 모형의 적합 및 예측

이론적인 변이도 모형은 여러 가지의 모형이 있으나 2절에서 소개한 구형 변이도(Spherical variogram), 지수 변이도(Exponential variogram) 그리고 가우시안 변이도(Gaussian variogram)등 세 가지 모형을 이용하여 적합시켰다. 또한 적합시킨 모형을 이용하여 하나의 자료를 제거하고 나머지 자료를 이용하여 제거된 자료를 예측하는 Leave-Oun-Out 방법을 이용하여 모형을 비교하였다. <표 2>에서 사용된 S_{opt}는 각 모형별로 S-PLUS가 제공하는 최적의 모수 추정값을 이용한 결과이고 Moran의 I와 Gear의 C는 범위(Range)를 3.4절에서 구한 값, 4.2m와 6.6m로 각각 고정시킨 후, 다른 모수는 최적의 모수 추정값을 사용하여 얻은 결과이며 R_{opt}는 범위를 <표 1>에서 구한 4.0m에서 4.4m 구간으로 제한한 후 구한 최적의 모수 추정값을 사용하여 얻은 결과이다. 또

한 Root SSres는 S-PLUS에서 출력해 주는 값으로 잔차제곱합에 제곱근을 취한 값을 나타낸다. 그리고 MSE와 MAPE 그리고 MSPE의 정의는 다음과 같다. 여기서 Z_i는 i번째 지점의 실제 자료값이고 $\hat{Z}_{(-i)}$는 i번째 자료를 제외한 자료를 이용하여 얻은 i번째 자료값의 예측값을 의미한다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \hat{Z}_{(-i)})^2$$

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i - \hat{Z}_{(-i)}}{Z_i}\right)^2$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Z_i - \hat{Z}_{(-i)}}{Z_i}\right|$$

다음의 <표 2>에 분석 결과를 정리하였다.

모형		range	Root SSres	MSE	MAPE	MSPE
구형	Sopt	5.52	73529.0	165540.7	0.168643	0.078228
	Moran의 I	4.20	74737.6	164509.5	0.168546	0.078974
	Geary의 C	6.60	83983.6	175582.8	0.171782	0.071654
	Ropt	4.40	73756.6	162532.4	0.167568	0.078086
지수	Sopt	8.80	108462.7	172889.6	0.172199	0.081296
	Moran의 I	4.20	120884.7	168978.3	0.170509	0.080778
	Geary의 C	6.60	196795.3	170069.0	0.170732	0.081036
	Ropt	4.00	114317.9	169126.3	0.170590	0.080788
가우시안	Sopt	3.95	73605.1	164297.1	0.168150	0.078235
	Moran의 I	4.20	89659.7	161980.7	0.167292	0.075964
	Geary의 C	6.60	128879.3	198627.5	0.186171	0.068613
	Ropt	4.00	86044.4	162584.3	0.167277	0.077010

<표 2> 각 모형별 범위 추정값과 MSE, MAPE, MSPE

<표 2>를 살펴보면 먼저 S-PLUS는 Root SSres를 최소로 해주는 최적의 모수 추정값을 출력해 주고 있음을 알 수 있다. 이제 Root SSres를 기준으로 모형을 결정한다면, 구형 변이도에서는 범위를 5.52m로 하는 모형을 선택할 것이며 가우시안 변이도에서는 범위를 3.95m로 하는 모형을 선택 할 것이다. 상대적으로 지수 변이도의 경우 Root SSres가 커서 모형으로 결정하기에는 어려

움이 있어 보인다. 두 모형 중에서 하나의 모형을 선택한다면 최소의 Root SSres를 주는 구형 모형을 선택할 것이다. 그러나 서론에서도 언급한 바와 같이 Root SSres의 차이는 매우 미미한 반면 범위(Range)의 차이는 상대적으로 크다.

본 자료 분석에서와 같이 범위는 벌레의 행동반경을 나타내 주는 매우 중요한 지표이므로 범위 값 자체가 갖는 의미는 매우 크다. Moran의 I 값에 비중을 더 둔다면 가우시안 변이도에서 범위(Range)가 3.95m인 모형을 선택하는 것이 바람직 하고, Geary의 C 값을 참조한다면 구형 변이도 중에서 범위(Range)가 5.52m인 모형을 선택하는 것이 타당하리라 판단된다. 두 모형의 Root SSres, MAPE 그리고 MSPE를 비교하면 거의 일치하는 것으로 판단되나 MSE 면에서는 가우시안 변이도가 우수한 것으로 판단되어 두 모형중에서 하나를 선택한다면 가우시안 변이도를 선택하는 것이 타당해 보인다. 그러나 가우시안 변이도중에서 범위가 3.95m인 모형과 범위가 4.2m인 모형을 비교하면 Root SSres는 차이가 나지만, 비슷한 범위를 가지면서 MSE, MSPE 그리고 MAPE 면에서 범위가 4.2m인 모형이 우수한 것으로 판단된다.

4. 결론

공간통계분석에서 사용되는 여러 이론적 변이도 중에서 자료를 잘 설명할 수 있는 변이도를 선택하는 것은 매우 중요하다. 지금까지는 구형 변이도, 지수 변이도 그리고 가우시안 변이도 중에서 잔차제곱합을 최소로 하는 변이도를 선택하는 것이 일반적인 방법이다. 그러나 농작물에 해를 미치는 곤충을 연구하는 곤충학자들에게 있어 해충의 행동반경(범위)을 알아내는 것은 매우 중요하며 따라서 공간상관관계의 최대 거리를 나타내는 범위는 그 추정값 자체가 중요할 수 있다. 이러한 경우에는 모수 추정값, 특히 범위(Range)의 추정값을 고려한 모형을 선택하는 것이 바람직하다. 본 논문에서는 Moran의 I 값과 Geary의 C값을 고려한 변이도 선택에 관하여 살펴보았다. 자료 분석 결과 공간독립 검정에 사용되는 두 통계량을 변이도 선택에 함께 고려함으로써 MSE, MAPE 그리고 MSPE를 기준으로 했을 때 더 우수한 모형을 선택할 수 있었다.

5. 참고 문헌

[1] Cho, K. Shin, K-I. and Park, J-J. (2002), Mapping Insect Population Distributions in Greenhouses by Applying Kriging, Proceedings of the VIII Intecol International Congress of Ecology.

[2] Cressie, N. (1993). Statistics for Spatial Data, John Wiley & Sons, Inc.

[3] Cressie, N. and Hawkins, D. M. (1980). Robust Estimation of the Variogram, I. Mathematical Geology, Vol. 12, No. 2, 115-125.

[4] Diblasi, A. and Bowman, A. W. (2001). On the Use of the Variogram in Checking for Independence in Spatial Data, Biometrics, 57, 211-218.

[5] Genton, M. C. (1998). Highly Robust Variogram Estimation, Mathematical Geology, Vol. 30, No. 2, 213-221.

[6] Goovaerts, P (1997), Geostatistics for Natural Resources Evaluation, Oxford.

[7] Hawkins, D. M. and Cressie, N. (1984). Robust Kriging - A Proposal, Journal of the Intrnational Association of Mathematical Geologist, 16, 3-18.

[8] Mugglestone, M. A., Baenet, V., Nirel, R. and Murray, D. A. (2000). Modeling and Analysing Outlier in Spatial Lattice Data, Mathematical and Computer modeling, 32, 1-10.

[9] Park, J-J. Shin, K-I. and Cho, K. (2002), Geostatistical Description of Spatial Distribution of Trialeurodes Vaporariorum in Cherry Tomato, Meeting of the Entomological Society of America, D0624.

[2003년 8월 접수, 2003년 11월 채택]