응용통계연구 제 17권 3호, 2004년, pp. 403-418

온실가루이의 공간시계열 분석*

박진모 [1] 신기일 [2]

요 약

시간에 따라 얻어진 공간 자료를 공간시계열 자료라 하며 이러한 자료를 분석하기 위해 사용되는 모형이 공간시계열 모형이다. 최근 곤충학과 생태학에서 공간시계열 모형을 이용한 연구가 활발히 진행되고 있다. 본 논문에서는 온실에 있는 곤충의 마리수를 ARMA 모형과 자기회귀오차모형을 이용한 공간 시계열 모형으로 분석하였다. 자료에 포함된 이상점은 분산도(Variogram) 추정에 많은 영향을 주기 때문에 Mugglestone (2000)의 이상점 수정법을 이용하여 수정하였다. 공간시계열 모형들과 시계열 요인을 배제한 공간모형을 MSE와 MAPE를 이용하여 비교하였다.

주요용어: 분산도, ARMA 모형, 분산도 구름, 이상점 수정법

1. 서론

최근 많은 분야에서 공간에서 얻어진 공간자료가 얻어지고 있으며 이중에서 시간에 따라 반복적으로 자료가 얻어지는 경우가 있다. 이렇게 시간에 따라 얻어진 공간 자료를 공간시계열 자료라 한다. 각 지점에서 오존의 양을 시간에 따라 측정한 자료나 온실의 각 지점에서 시간에 따라 조사된 벌레 마리 수 등이 공간시계열 자료의 예라 할 수 있다. 공간시계열 모형은 공간통계 분석 방법과 시계열 분석 방법을 통합하여 분석할 경우에 사용된다. 최근 생태학과 곤충학에서 공간 통계학을 이용한 자료분석은 많은 학자들의 관심을 끌고 있으며 온실에서 발생하는 해충을 줄이기 위한 많은 연구도 진행 되고 있다. 해충을 줄이기 위하여 천적을 이용하기도 하지만 살충제를 사용한 방제가 일반적으로 사용되고 있다. 살충제 사용의 어려운 점은 살충제의 살포 시기와 최적량을 알아내야 한다는 것이다. 이러한 결정에 도움을 줄 수 있는 통계 분석기법이 공간 시계열 분석기법이다. 공간 시계열 분석 방법을 이용하게 되면 미래에 발생할 벌레 마리수를 예측할 수 있게 되고 이를 바탕으로 살충제 살포시기와 그 양을 조절할 수 있기 때문이다.

본 논문에서는 충남 부여 지방의 한 온실에서 7주간 얻어진 온실가루이 자료를 분석하였다. 일반적으로 벌레 마리수는 과산포 포아송 분포를 따르는 것으로 알려져 있으며 본 논문에서 사용된 자료 또한 과산포 포아송분포를 따르고 있다. 이러한 자료를 분석하는 여러 방법이 발표되었으며 주로 사용되고 있는 대표적인 방법이 TPL(Taylor's Power Law:

* 본 연구는 한국과학재단 특정기초연구 지원으로 수행되었음

1) (449-791) 경기도 용인시 모현면 왕산리 산 79, 한국외국어대학교 통계학과, 대학원
E-mail : jmpark@stat.hufs.ac.kr
2) (449-791) 경기도 용인시 모현면 왕산리 산 79, 한국외국어대학교 통계학과, 교수
E-mail : keyshin@hufs.ac.kr

Talor(1961)) 방법이다. 그러나 최근 공간통계학 분석 기법을 사용하여 분석한 많은 논문들이 발표되고 있다.

지리통계학(Geostatistics) 분석 기법을 사용하기 위하여는 먼저 과산포 포아송분포를 따르는 자료를 정규화시키기 위한 변환이 필요하게 된다. log 변환이 일반적으로 사용되고 있으나 log_{10} 변환도 사용되고 있다. 이에 관한 내용은 Ettema 등(2000)과 Tobin 등(2003)을 참조하기 바란다. 본 논문에서는 log 변환과 0.25 변환을 사용하였다. 이 두 변환을 사용한 이유는 Park 등(2004)을 참조하기 바란다.

공간통계분석에서 이상점은 분석 결과에 많은 영향을 미치게 된다. 특히 본 자료에서와 같이 이상점이 존재할 경우 이를 처리하지 않고 분석하게 되면 분산도의 범위(Range) 추정값이 매우 크게나와 분석 자체의 의미가 상실 될 수 있게 된다. 이상점의 영향력을 축소하기 위한 여러 방법이 제안되었다. 대표적인 방법이 분산도 추정량을 로버스트 방법으로 구하는 것이다. Cressie와 Hawkins(1980)가 로버스트 추정량을 제안한 후 Genton(1998)의 로버스트 추정량 등 많은 연구가 이루어졌다. 최근 Mugglestone(2000)은 자료에서 먼저 이상점을 수정한 후 분석을 실시하는 방법을 제안하였으며 Lee와 Shin(2004)은 Mugglestone이 제안한 방법이 지리통계분석에서도 사용할 수 있음을 보였다. 그러나 본 논문에서 사용한 자료는 벌레 마리수로 매우 큰 값을 갖는 이상점 자체가 의미를 갖고 있다. 곤충학에서는 이를 'Hot Spot'이라 부르며 이에 관한 많은 연구가 진행하고 있다. 따라서 'Hot Spot'에 관한 정보를 분석에 추가하여 분석을 하게 되면 더욱 의미있는 분석이 될 수 있을 것이다. 이를 위한 방법으로 과산포 포아송 분포를 가정한 베이지안 분석방법 또는 HGLM등이 고려될 수 있으리라 생각된다. 그러나 'Hot Spot' 자체가 시간에 따라 이동하고 있으며 생성과 소멸이 랜덤으로 형성되어 이를 통계적으로 분석하는 것은 간단한 일이 아니다. 이에 본 논문에서는 Mugglestone(2000)이 제안한 방법으로 이상점을 수정한 후 Cressie와 Hawkins (1980)가 제안한 로버스트 방법으로 분산도를 추정하는 방법을 선택하여 분석하였다.

최근 Ettema 등(2000)은 5종의 벌레 마리수 분석을 위한 다음의 공간 시계열 분산도 모형을 사용하였다.

$$\gamma(h,t) = \theta_1 + \theta_2(1 - exp(-\frac{h}{\theta_3} - \frac{t}{\theta_4})) \tag{1.1}$$

여기서 θ_1은 뭉치(Nugget)를 $\theta_1 + \theta_2$는 문턱(Sill)을 그리고 θ_3는 범위(Range)를 나타낸다. 또한 시계열적 요인을 위한 모수로 θ_4를 사용하였다. Ettema 논문에서 사용된 자료는 시점 수가 4이며 등간격으로 조사가 이루어지지 않았기 때문에 시계열분석에서 얻어진 이론을 바로 적용하기는 어려웠을 것이다. 그러나 수식 (1.1)은 분산도 모형중에서 지수 모형이고 시계열 요인 분석을 위한 부분은 시계열 분석에서 사용하는 AR(1) 모형의 자기상관함수(ACF)와 동일하다는 것을 쉽게 알 수 있다. 즉 Ettema는 시계열 요인을 분석하기 위하여 변형된 AR(1)모형을 사용했다고 할 수 있다.

Wikle과 Cressie(1999)와 Huang과 Cressie(1996)는 Kalman-Filter 모형을 이용하여 공간 시계열 자료를 분석하였다. 그러나 Kalman-Filter 모형을 이용한 분석은 복잡한 계산과정이 필요하고 자료의 수가 많은 경우 분석에 어려움이 따른다. 이에 본 논문에서는 Kalman-Filter 모형의 특별한 경우인 ARMA 모형 또는 자기회귀오차모형과 공간 모형(분산도)를 이용한 공간시계열 모형을 이용하여 분석하였다.

논문의 구성은 다음과 같다. 먼저 2절에서 공간시계열 모형에 사용되는 Kalman-Filter모형에 관하여 살펴보았다. 또한 Cressie와 Hawkins가 제안한 로버스트 추정법과 Mugglestone (2000)이 제안한 방법도 살펴보았다. 3절에서는 자료 분석을 통하여 2절에서 살펴본 각 이론의 사용방법을 살펴보았으며 4절에서는 공간시계열모형을 이용한 경우와 그렇지 않은 경우의 분석 결과를 비교하였다. 5절에 결론이 있다.

2. 공간시계열에 관한 이론

2.1. 공간시계열 분석을 위한 시계열 모형

2.1.1. Kalman-Filter 모형

Kalman-Filter 모형은 다음의 두 방정식으로 구성되어 있다.

$$\xi_{t+1} = F\xi_t + v_{t+1} \tag{2.1}$$

$$y_t = A'x_t + H'\xi_t + \omega_t \tag{2.2}$$

(2.1)식은 상태 방정식(State Equation), (2.2)식은 관측 방정식(Observation Equation)이라 부른다. 여기서
F : 시계열 모형을 상태공간으로 표현할 때 나오는 행렬로 크기가 $r \times r$ 인 행렬,
A' : 독립변수 x_t와 종속변수 y_t 의 관계를 나타내는 계수행렬로 크기가 $n \times k$인 행렬,
H' : ξ_t 와 종속변수와의 관계를 나타내는 계수 행렬로 크기가 $n \times k$인 행렬이다. 또한

$$E(v_t v_\tau') = \begin{cases} Q, & \text{for } t = \tau \\ 0, & \text{otherwise} \end{cases}, \quad E(\omega_t \omega_\tau') = \begin{cases} R, & \text{for } t = \tau \\ 0, & \text{otherwise} \end{cases} \tag{2.3}$$

이며, v_t와 ω_t는 독립이라 가정한다. Kalman-Filter 모형에 관한 자세한 내용은 Hamilton (1994)를 참조하기 바란다. 위 식들을 이용하면 공간시계열 모형 중 시계열 부분을 설명할 수 있다. 또한 행렬 R은 각 조사점에서 얻어진 공간자료의 공간상관관계를 설명하고 있다. 공간시계열 모형에서 관측점의 개수가 n으로 표현되며 AR(p)모형을 고려할 경우 $r = n \times p$가 된다. 따라서 관측점의 개수 n이 큰 경우 분석이 복잡하게 되기 때문에 본 논문에서는 Kalman-Filter 모형의 특수한 경우인 ARMA(p, q)모형과 자기회귀오차 모형만을 고려하였다. Kalman-Filter 모형을 이용한 공간시계열 분석은 Huang과 Cressie(1996) 그리고 Wikle과 Cressie(1999)를 살펴보기 바란다.

2.1.2. 자기회귀오차 모형(Autoregressive Error Model)

자기회귀오차 모형은 해석이 용이하고 모형이 간단하여 다양한 분석에 사용될 수 있으며 일반적으로 사용하는 자기회귀오차 모형은 다음과 같다.

$$y_t = x_t'\beta + \varepsilon_t \quad (= \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + \varepsilon_t) \tag{2.4}$$

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \cdots + \phi_p \varepsilon_{t-p} + a_t \tag{2.5}$$

여기서 $a_t \sim i.i.d(0, \sigma^2)$를 따른다고 가정한다. 모형 (2.4)를 이용하면 쉽게 공간시계열 모형을 고려할 수 있다. 즉

$$y_{i,t} = x'_{i,t}\beta + \varepsilon_{i,t} + \omega_{i,t} \tag{2.6}$$

$$\varepsilon_{i,t} = \phi_1\varepsilon_{i,t-1} + \cdots + \phi_p\varepsilon_{i,t-p} + a_{i,t} \tag{2.7}$$

주어진 시점 t에서, $Var(\omega_{i,t}) = R$은 공간상관관계를 나타내며 $Cov(\omega_{i,t}, \omega_{i,\tau}) = 0$, $i \neq \tau$이다. 또한, $Cov(a_{i,t}, \omega_{i,t}) = 0$이고 $a_{i,t} \sim i.i.d(0, \sigma^2)$ 이라 가정한다. 추세가 있으며 오차 $\varepsilon_{i,t}$ 가 AR모형을 따를 경우 쉽게 고려할 수 있는 모형이며 Kalman-Filter 모형의 특수한 경우임을 쉽게 알 수 있다.

2.1.3. ARMA 모형(Autoregressive Moving Average Model)

정상성을 만족하는 시계열 자료는 ARMA(p, q) 모형을 이용하여 분석한다. 일반적인 ARMA(p, q)모형은 다음과 같다.

$$\Phi(B)(Y_t - \mu) = \Theta(B)a_t \tag{2.8}$$

여기서 $\Phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$ 와 $\Theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$는 공통근을 갖고 있지 않으며 정상성(Stationarity) 과 가역성(Invertibility)을 만족한다고 가정한다. 또한 $a_{i,t} \sim i.i.d(0, \sigma^2)$를 따른다고 가정한다. 이러한 ARMA$(p, q)$ 모형은 Kalman-Filter모형의 특수한 경우이고 이에 관한 자세한 내용은 Hamilton(1994) 를 참조하기 바란다. Huang과 Cressie(1996) 는 Kalman-Filter 모형 중에서 ARMA 모형을 이용하여 공간시계열 모형을 분석하였다. 본 논문에서는 ARMA(p, q) 모형과 공간상관관계를 고려한 다음의 모형을 고려하였다.

$$Y_{i,t} - \mu = \phi_1(Y_{i,t-1} - \mu) + \cdots + \phi_p(Y_{i,t-p} - \mu) + a_{i,t} - \theta_1 a_{i,t-1} - \cdots - \theta_q a_{i,t-q} + \omega_{i,t} \tag{2.9}$$

여기서도 주어진 시점 t에서, $Var(\omega_{i,t}) = R$은 공간상관관계를 나타내며, $Cov(\omega_{i,t}, \omega_{i,\tau}) = 0$, $i \neq \tau$이다. 또한 $Cov(a_{i,t}, \omega_{i,t}) = 0$이고, $a_{i,t} \sim i.i.d(0, \sigma^2)$이라 가정한다.

2.2. 분산도(Variogram)

분산도는 공간통계분석에서 가장 중요하게 사용되는 통계량이다. 이 분산도를 이용하여 오차의 상관구조를 파악하게 된다.

2.2.1. 경험적 분산도(Empirical Variogram)

정상성을 만족하는 공간자료의 분산도를 $2\gamma(h) = Var\{Y(u_i) - Y(u_i + h)\}$라 하면 $2\gamma(h) = E\{Y(u_i) - Y(u_i + h)\}^2$가 성립한다. 분산도의 전통적인 경험적 분산도 추정량은 다음과 같다.

$$2\hat{\gamma}(h) = \frac{1}{N_h}\sum_{i=1}^{N_h}(Y(u_i) - Y(u_i + h))^2 \tag{2.10}$$

여기서 N_h는 두 지점간의 거리가 h인 쌍의 개수를 말하며 $Y(u_i)$는 u_i좌표에서의 관측값을 의미한다. 이러한 전통적 분산도 추정량은 이상점에 크게 영향을 받는 것으로 알려져 있으며 Cressie와 Hawkins(1980)는 다음의 Robust추정량을 제안하였다.

$$2\hat{\gamma}(h) = \frac{1}{B(h)}\{\frac{1}{|N_h|}\sum_{i=1}^{N_h}|Y(u_i) - Y(u_i+h)|^{1/2}\}^4 \tag{2.11}$$

여기서 $B(h) = 0.457 + 0.494/|N_h|$이다. 본 논문에서는 식 (2.11)을 이용하여 경험적 분산도 추정값을 계산하였다.

2.2.2. 경험적 분산도의 모형화

공간자료분석에서 사용되는 많은 이론적 분산도 중에서 본 논문에서는 다음의 세 모형을 고려하였다.

(1) 구형(Spherical) $(d = 1, 2, 3)$

$$\gamma(h,\theta) = \theta_1 + \theta_2(\frac{3}{2}\frac{h}{\theta_3} - \frac{1}{2}(\frac{h}{\theta_3})^3), \quad 0 < h < \theta_3 \tag{2.12}$$

$$\gamma(h,\theta) = \theta_1 + \theta_2, \quad h > \theta_3, \quad 0 < h \tag{2.13}$$

(2) 지수(Exponential) $(d \geq 1)$

$$\gamma(h,\theta) = \theta_1 + \theta_2(1 - exp(-\frac{h}{\theta_3})), \quad 0 < h \tag{2.14}$$

(3) 가우시안(Gaussian) $(d \geq 1)$

$$\gamma(h,\theta) = \theta_1 + \theta_2(1 - exp(-\frac{h^2}{{\theta_3}^2})), \quad 0 < h \tag{2.15}$$

위의 세 분산도는 정상성을 모두 만족하며 범위(Range)와 문턱(Sill)이 존재한다. 여기서 θ_1은 뭉치(Nugget)를 $\theta_1 + \theta_2$는 문턱(Sill)을 나타낸다. 구형분산도의 경우 범위는 θ_3 이나 가우시안과 지수 분산도의 경우는 θ_3 를 이용하여 실제 범위를 구한다. 즉 지수 분산도의 실제 범위는 $3\theta_3$가 사용되고 가우시안 분산도의 실제 범위는 $\sqrt{3}\theta_3$가 사용된다. 본 논문에서 사용된 모수 추정방법은 가중제곱합(WLS)을 최소화 시키는 방법으로 다음을 최소화 시키는 방법을 이용하여 모수를 추정하였다.

$$\sum_{i=1}^{K}|N_h(j)|\{\frac{\hat{\gamma}(h(j))}{\gamma(h(j);\theta)} - 1\}^2 \tag{2.16}$$

여기서 N_h는 두 지점간의 거리가 h인 쌍들의 개수를 의미한다. 추정된 모수를 이용하여 공간통계학의 예측 방법인 Kriging이 이루어졌다.

2.3. 이상점 수정

여러 이상점 수정법이 제안되었지만 본 논문에서 사용된 이상점 수정법은 최근에 발표된 Mugglestone 등(2000)의 방법이 사용되었다. 먼저 Nierel 등(1998)은 데이터에서 이상점을 찾는 방법을 제안하였다. 먼저 다음의 식을 고려하자.

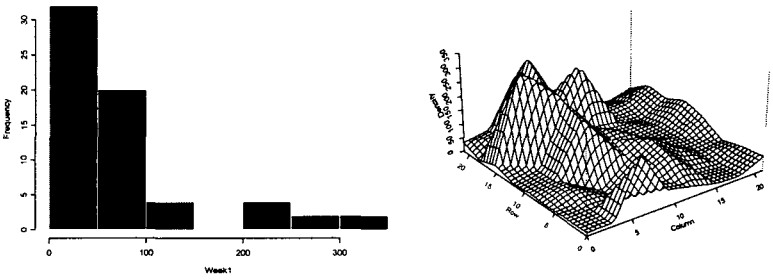
$$\psi(y_{u,v}; M, g^0, g^1) = \begin{cases} y_{u,v}, & \text{if } |y_{u,v} - g^0(y_{u,v})| \le M \\ g^1(y_{u,v}), & \text{otherwise} \end{cases} \tag{2.17}$$

여기서 $y_{u,v}$는 관측값이고 주어진 자료의 중앙값인 $g^0(y_{u,v})$의 값과 자료값의 차의 절대값이 주어진 M값보다 큰 경우, 즉 $|y_{u,v} - g^0(y_{u,v})| > M$ 경우 이상점으로 판단한다. 이상점이 아니라고 판단된 경우에는 원자료를 그대로 사용하고 이상점이라 판단된 자료는 $g^1(y_{u,v})$로 대체한다. 여기서 $g^1(y_{u,v})$는 이상점이라 판단된 자료 주위에 있는 네 개의 이웃들의 중앙값이다. 이 방법을 NM(Neighboring Median Cleaner)이라 한다. NM방법은 이상점의 이웃 자료가 이상점일 경우 이상점 수정이 잘 안되는 것이 문제점인 것으로 밝혀졌다. 이를 보완한 방법이 Mugglestone 등(2000)이 제안한 NTM(Neighboring Truncated Median)방법이다. 본 논문에서는 NTM방법이 사용되었으며 이에 관한 자세한 내용은 Mugglestone 등(2000)을 참고하기 바란다.

3. 자료분석

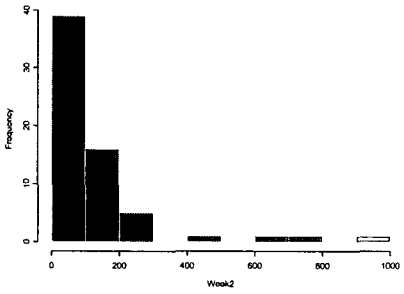
3.1. 데이터 설명

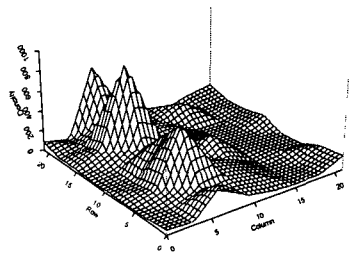
본 논문에서 사용된 자료는 2001년 충남 부여지방의 한 온실에서 7주간에 걸쳐 격자 모양의 64개 위치별 트랩에 유인된 온실가루이의 마리수이다. 조사된 격자는 8×8의 격자형태로 되어있으며 행별, 열별 자료의 거리는 $3m$ 이다.



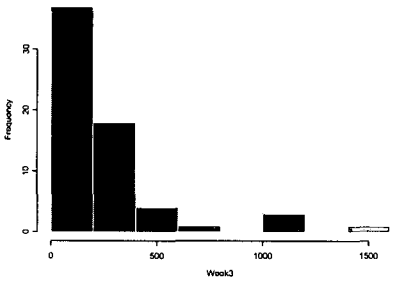
(1주)

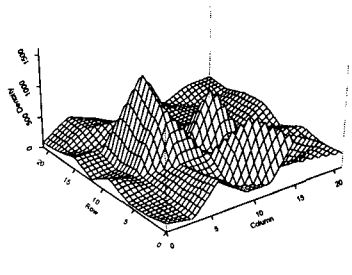
그림 3.1: 각주별 자료의 히스토그램과 3D-Perspective Plot



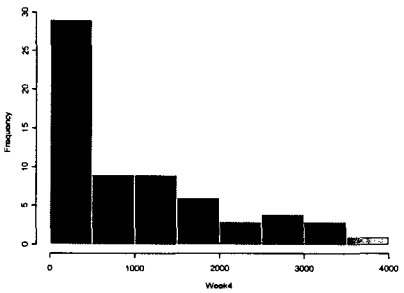


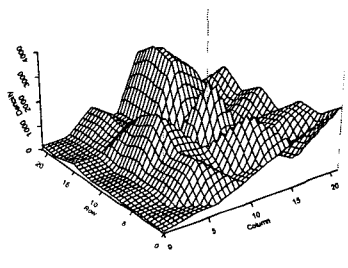
(2주)



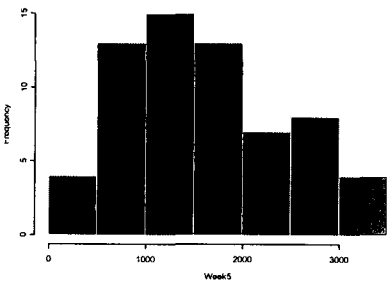


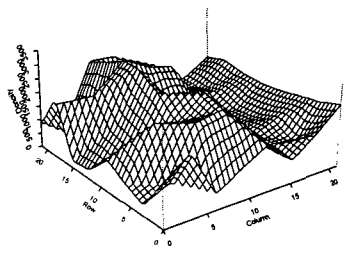
(3주)





(4주)





(5주)

그림 3.1: 각주별 자료의 히스토그램과 3D-Perspective Plot (계속)

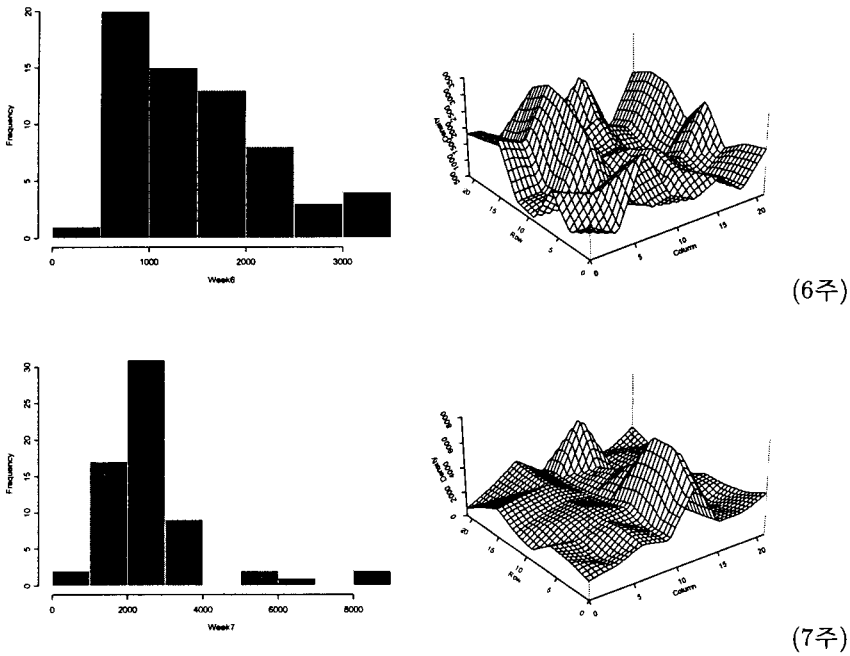
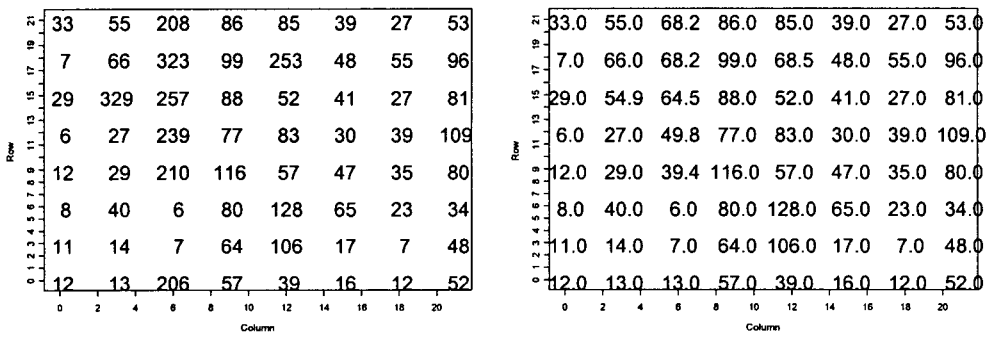


그림 3.1: 각주별 자료의 히스토그램과 3D-Perspective Plot (계속)

그림 3.1에 나와 있는 7주간 자료의 히스토그램을 살펴보면 각 주별 자료에서 이상점이 존재한다는 것을 알 수 있다. 따라서 공간시계열 분석에 앞서 이상점의 유무를 확인한 후 2.3절에서 기술한 Mugglestone(2000)의 NTM 이상점 수정법을 사용하여 이상점을 수정하였다.

M=133.9758 m.c.v=-83.97575 M.C.V=183.9758



1주(수정전) 1주(수정후)

그림 3.2: 이상점 수정후의 자료의 형태

7주간에 걸쳐 얻어진 자료를 모두 수정하였으며 본 논문에서는 이 중에서 1주만을 그림 3.2에 나타냈다. 1주 자료는 이상점이 몰려 있는 패치 이상점임을 알 수 있다. 이는 NM 방법보다 NTM 방법을 사용하여야 함을 보여주고 있다.

3.2. 시계열 모형

3.2.1. 자기회귀오차 모형 적합

다음은 7주간의 자료를 이용하여 64개의 각 지점별 공간자료를 시계열 자료의 형태로 고친 후 그린 그래프이다.

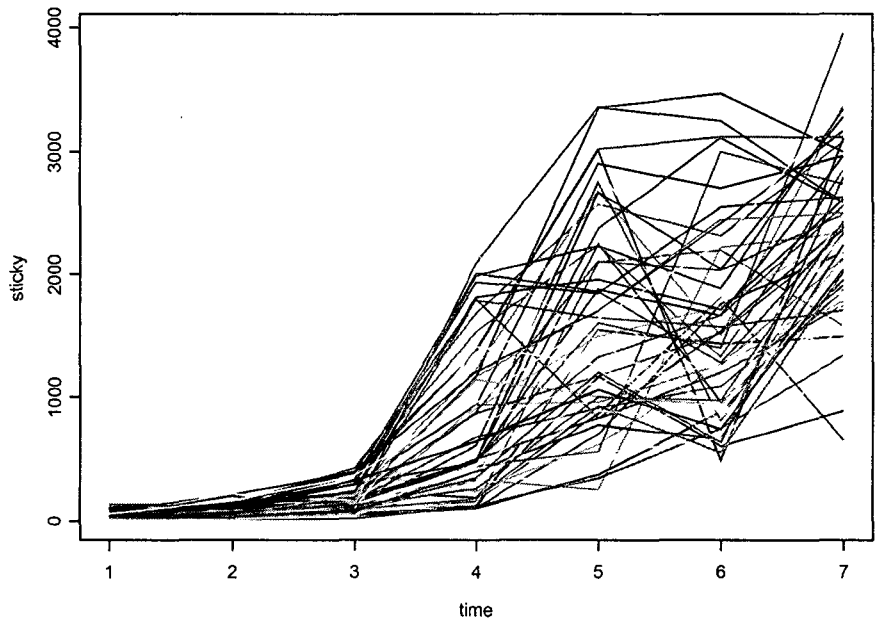
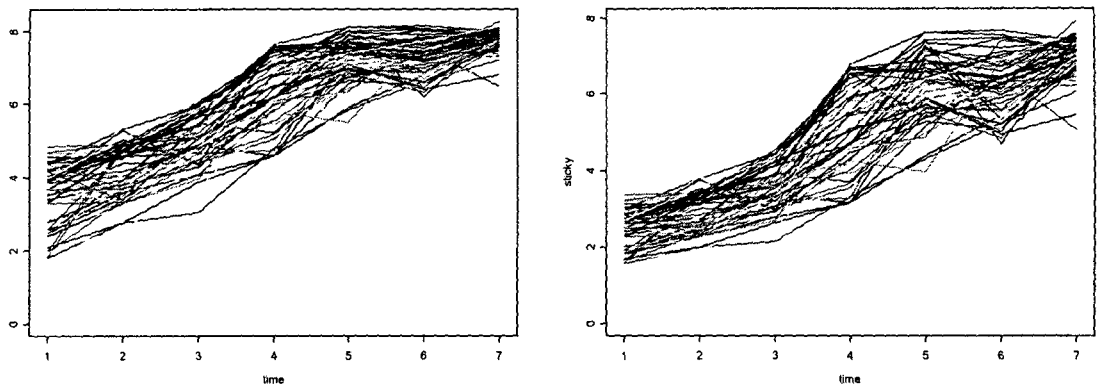


그림 3.3: 각 주별 64개 지점에 대한 그래프

그림 3.3을 살펴보면 시간이 흘러갈수록 벌레수가 증가하는 것을 볼 수 있다. 또한 분산도 커지는 것을 확인할 수 있다. 이에 대하여 다음의 Box-Cox 변환을 고려하였다.

$$Y = \frac{X^{\lambda} - 1}{\lambda} \tag{3.1}$$

여기서 λ는 변환 모수(Transformation Parameter)이며 $\lambda = 0$인 경우가 log 변환을 의미한다.



a) $\lambda = 0$인 경우(Log변환) b) $\lambda = 0.25$인 경우

그림 3.4: 변환후의 그래프

위의 그림 3.4를 보면 log변환과 0.25변환을 사용할 경우 분산이 안정되어 있음을 확인할 수 있다. 따라서 분석에 사용한 변환은 log변환과 0.25변환이다. 이제 Y를 변환된 자료라 하자. 모형 (2.6), (2.7)과 그림 3.4를 이용하여 다음의 모형을 설정하였다.

$$Y_{i,t} = \beta_0 + \beta_1 t + \varepsilon_{i,t} + \omega_{i,t} \tag{3.2}$$

$$\varepsilon_{i,t} = \phi\varepsilon_{i,t-1} + a_{i,t} \tag{3.3}$$

모수 β_0, β_1 그리고 ϕ의 추정은 SAS/Proc Autoreg 를 이용하였다. 여기서 $Y_{i,t} = X_{i,t}^{\lambda}$는 변환된 i번째 위치의 t주 관측치를 나타내며 $Var(\omega_{i,t}) = R$로 공간상관관계를 나타내며 $Cov(\omega_{i,t}, \omega_{i,\tau}) = 0$, $i \neq \tau$이다. 다음이 SAS를 이용하여 자기회귀오차 모형을 적합한 결과이다.

$$X_t^{0.25} = 1.6704 + 0.7615t + \varepsilon_t \tag{3.4}$$

$$\varepsilon_t = 0.377664\varepsilon_{t-1} + a_t \tag{3.5}$$

$$logX_t = 3.0463 + 0.7051t + \varepsilon_t \tag{3.6}$$

$$\varepsilon_t = 0.441435\varepsilon_{t-1} + a_t \tag{3.7}$$

이제 (3.4), (3.5)식 또는 (3.6), (3.7)식을 이용하여 모형으로부터 잔차

$$\hat{\omega_{i,t}} = y_{i,t} - \hat{\beta_0} - \hat{\beta_1} t - \hat{\varepsilon_{i,t}}, \quad \lambda = 0, 0.25 \tag{3.8}$$

를 구한 후 이를 표준화하여 공간모형 분석에 사용하였다.

3.2.2. ARMA 모형을 적합시킨 경우

ARMA 모형을 이용한 분석도 3.2.1절에서와 같이 $\lambda = 0, 0.25$를 이용한 변환을 실시하였다. 그러나 3.2.1절에서도 알 수 있듯이 추세가 존재한다. 정상 ARMA 모형을 적합시키기 위해서는 추세를 제거해야 하므로 각 주별로 자료를 표준화시켰다. 또한 분석을 간단하게 하기 위하여 다음의 AR(1) 모형을 선택하였다.

$$Y_{i,t} = \phi Y_{i,t-1} + \omega_{i,t}, \quad i = 1, 2, 3, \cdots, 64, \quad t = 1, 2, 3, \cdots, 7 \tag{3.9}$$

여기서도 $Y_{i,t} = X_{i,t}^{\lambda}$는 변환된 i번째 위치의 t주 관측치를 나타내며 $Var(\omega_{i,t}) = R$로 공간상관관계를 나타내며 $Cov(\omega_{i,t}, \omega_{i,\tau}) = 0$, $i \neq \tau$이다. 모수 ϕ의 추정은 시계열 분석에서 일반적으로 사용하는 다음 식을 이용하였다.

$$\hat{\phi} = \frac{\sum_{i=1}^{64} \sum_{t=2}^{6} Y_{i,t} Y_{i,t-1}}{\sum_{i=1}^{64} \sum_{t=2}^{6} Y_{i,t}^2} \tag{3.10}$$

(3.10)식을 이용하여 얻은 모수추정값은 다음과 같다.

1) *log* 변환된 자료의 AR(1)모형

$$logX_{i,t} = 0.461 logX_{i,t-1} + \omega_{i,t} \tag{3.11}$$

2) 0.25변환된 자료의 AR(1)모형

$$X_{i,t}^{0.25} = 0.459 X_{i,t-1}^{0.25} + \omega_{i,t} \tag{3.12}$$

이제 (3.11)과 (3.12)식을 이용하여 잔차 $r_{i,t} = y_{i,t} - \hat{\phi} y_{i,t-1}$를 구한 후 이를 공간 분석에 사용하였다.

3.2.3. 변환만을 이용한 경우

3.2.1절과 3.2.2절에서 처럼 자기회귀오차 모형이나 ARMA 모형을 적합시키지 않고 각 주별 자료가 독립이라는 가정하에 변환만을 실시한 자료를 이용하여 변이도를 추정하여 모형을 적합시켰다. 위의 시계열 모형과 비교하기 위해 $\lambda = 0$인 *log* 변환과 $\lambda = 0.25$ 변환을 실시하였다. 변환된 자료를 표준화하여 얻어진 자료를 분석에 사용하였다.

3.3. 분산도 적합

분산도 적합을 위해 구형(Spherical), 지수(Exponential) 그리고 가우시안(Gaussian) 분산도가 사용되었다. 각 주별로 분산도가 구해지나 전체의 자료를 대표할 수 있는 하나의 분산도를 적합하기 위해서 각 주별로 얻어진 분산도를 분산도구름(Variogram Cloud)으로 만들어 가중최소제곱추정법으로 모수를 추정하였다. 다음이 적합된 분산도이다.

표 3.1: 각 모형을 이용한 적합

모형		Range	Sill	Nugget	Objective
ARMA (0.25변환)	Spher	8.4465844	0.5452555	0.3261507	2.9838
	Exp	2.7154861	0.8953389	0.00000	2.9288
	Gauss	2.8632931	0.8593089	0.00000	3.0236
ARMA (Log변환)	Spher	5.6520066	0.8203786	0.00000	2.9916
	Exp	2.4691969	0.8502421	0.00000	2.8971
	Gauss	2.7382334	0.8234596	0.00000	2.9745
자기회귀 (0.25변환)	Spher	5.827326	1.050246	0.00000	4.3521
	Exp	2.549399	1.090914	0.00000	4.2441
	Gauss	2.818517	1.054719	0.00000	4.3186
자기회귀 (Log변환)	Spher	5.4381865	0.9925628	0.00000	3.7539
	Exp	2.271147	1.021041	0.00000	3.6556
	Gauss	2.6322739	0.9955387	0.00000	3.7383
	Spher	6.934999	1.131731	0.00000	5.9834
(0.25변환)	Exp	2.979262	1.185076	0.00000	5.989
	Gauss	3.188892	1.132604	0.00000	5.934
	Spher	6.692417	1.081806	0.00000	5.9507
(Log변환)	Exp	2.825199	1.126648	0.00000	5.9282
	Gauss	3.102826	1.082997	0.00000	5.8884

4. 각 방법의 비교

3절에서 얻어진 각 분산도를 이용하여 각 지점에 대한 예측을 실시하였다. 예측력을 비교하기 위해 Leave one out 방법이 사용되었으며 각 모형별로 MSE, MAPE 통계량을 구하였다. 여기서 $\hat{y_{-i}}$는 i번째 위치의 자료를 제거한 후 나머지 자료를 이용하여 i번째 위치의 자료를 예측한 예측값이고 y_i는 i번째 자료의 실제 자료값이다. 표 4.1은 6주 동안 얻어진 자료를 이용하여 모형을 설정한 후 7주 자료를 이용하여 예측력을 비교한 결과이다. 표에서 정의된 MSE와 MAPE는 다음과 같다. 여기서 n은 자료의 수이다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_{-i}})^2 \tag{4.1}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_{-i}}}{y_i}\right| \tag{4.2}$$

표 4.1: 각 모형별 MSE, MAPE

모형		MSE	MAPE
ARMA (0.25변환)	Spher	225052.5	0.172048
	Exp	218319.1	0.170284
	Gauss	236033.4	0.175164
ARMA (Log변환)	Spher	229238.1	0.170526
	Exp	221348.5	0.169858
	Gauss	234354.7	0.172279
자기회귀 (0.25변환)	Spher	201566.6	0.166787
	Exp	194754.2	0.165201
	Gauss	214457.8	0.170749
자기회귀 (Log변환)	Spher	243215.6	0.185027
	Exp	241124.0	0.182416
	Gauss	247042.7	0.185963
(0.25변환)	Spher	251823.1	0.185273
	Exp	226456.6	0.175691
	Gauss	206544.0	0.170831
(Log변환)	Spher	240262.6	0.179977
	Exp	226972.6	0.173849
	Gauss	209256.6	0.171468

표 4.1을 살펴보면 전체적으로 자기회귀오차 모형중에서 0.25변환을 사용한 분석기법이 MSE와 MAPE 모두에서 우수한 것으로 나타났다. 그러나 시계열 분석을 적용하지 않고 0.25변환과 log변환만 사용한 가우시안 모형은 MSE를 기준으로 하였을 때 가우시안 모형 중에서 가장 우수한 것으로 나타났다. 또한 자기회귀 모형과 log변환을 이용한 분석이 전체적으로 나쁜 결과를 주고 있다. 이는 7주 자료 중에서 하나의 자료만을 제거한 후 예측하여 비교하는 Leave one out 방법을 사용하였기 때문에 시계열 요인에의한 설명력이 공간 상관에 의한 설명력에 비해 상대적으로 약해지기 때문인 것으로 판단된다.

이제 다른 기준을 이용하여 모형의 우수성을 비교하자. 일반적으로 곤충학자들은 적은 수의 표본을 조사하였을 때 MSE 또는 MAPE가 작은 모형이 우수한 모형이라 판단을 내린다. 이를 위하여 먼저 7주 자료에서 10개 자료를 임의로 추출한 후 나머지 54개 자료를 예측하였으며 얻어진 값을 이용하여 MSE, MAPE를 구하였다. 이 과정을 100번 반복한 후 얻어진 값을 평균하여 얻은 결과를 표 4.2에 정리하였다.

표 4.2: 각 모형별 MSE, MAPE

모형		MSE	MAPE
ARMA (0.25변환)	Spher	427574.5	0.265696
	Exp	427269.4	0.265832
	Gauss	426004.1	0.265739
ARMA (Log변환)	Spher	434912.8	0.266536
	Exp	435139.3	0.266442
	Gauss	434482.3	0.266523
자기회귀 (0.25변환)	Spher	359851.3	0.241876
	Exp	361063.6	0.242311
	Gauss	360303.9	0.242226
자기회귀 (Log변환)	Spher	407142.4	0.255116
	Exp	408369.0	0.255304
	Gauss	407319.0	0.255226
	Spher	431325.5	0.269709
(0.25변환)	Exp	433531.4	0.270577
	Gauss	430155.7	0.269301
	Spher	433255.0	0.268353
(Log변환)	Exp	436028.2	0.269284
	Gauss	432168.1	0.267986

표 4.2을 살펴보면 변환만을 이용한 분석 결과가 시계열 모형을 이용한 분석에 비해 상대적으로 나쁜 결과를 주고 있다. 이 결과는 자료가 시계열적으로 얻어졌고 예측에 사용된 자료의 수가 10개로 작기 때문에 얻이진 예견된 결과로 받아들일 수 있을 것이다. 이제 시계열 모형이 포함된 결과를 살펴보자. 두 모형의 비교에서는 자기회귀오차 모형이 우수한 것을 알 수 있다. 이는 ARMA 모형의 경우, 정상성을 만족시키기 위해 자료를 표준화 하여 추세를 제거하고 이때 64개 자료 중 10개만 평균 추정에 사용되기 때문에 평균에 기초한 Ordinary kriging을 사용하는 ARMA 모형은 상대적으로 추세를 그대로 유지하고 있는 자기회귀오차 모형에 비해 나쁜 결과를 주는 것으로 판단할 수 있다. 결론적으로 본 자료에서와 같이 자료수가 작고, 추세가 있는 경우에는 추세를 살린 모형인 자기회귀오차 모형을 사용하면 좋은 결과를 얻을 수 있을 것이다.

5. 결론

공간시계열 모형에 관한 연구가 활발히 진행되고 있으며 그 응용범위도 증가하고 있다. 자료가 공간적, 시간적으로 얻어진 경우 공간 시계열 모형을 이용하여 분석하면 더 좋은 결

과를 얻을 수 있게 되며 이는 당연한 결과이다. Kalman-Filter 모형을 이용한 공간시계열 모형과 Ettema가 제안한 공간 시계열 모형 뿐 아니라 서론에서도 언급한, 과산포 포아송 분포를 가정한 베이지안 분석 방법을 이용하거나 HGLM 등의 방법을 이용한 분석 방법도 이용될 수 있을 것이다. 그러나 이러한 모형들은 공간요인에 관한 모수와 시계열 요인에 관한 모수를 동시에 추정해야 하는 등 분석이 쉽지 않다. 물론 이러한 방법들은 본 논문에서 사용한 방법에 비해 더욱 좋은 결과를 줄 수 있으리라 생각된다. 본 논문에서는 시계열 요인을 모형화 하기 위한 방법으로 자기회귀오차 모형과 ARMA 모형을 이용한 공간시계열 모형을 살펴보았고 모수 추정을 단계별로 하는 등 분석을 간단히 하였음에도 불구하고, 표 4.2에서도 알 수 있듯이 공간모형에 비해 MSE와 MAPE를 기준으로 했을 때 우수한 결과를 주고 있음을 확인할 수 있었다.

참고문헌

Cressie, N. (1993). *Statistics for Spatial Data.* John Wiley, Sons, Inc.

Cressie, N. and Hawkins, D. M. (1980). Robust estimation of the variogram. I. *Mathematical Geology*, **12**, 2, 115-125.

Ettema, C. H., Rathbun, S. L. and Coleman, D. C. (2000). On spatiotemporal patchness and the coexistence of five species of Chronogaster(Nematoda: Chronogasteridae) in a riparian wetland. *Oecologia*, **125**, 444-452.

Genton, M. C. (1998). Highly robust variogram estimation. *Mathematical Geology*, **30**, 2, 213-221.

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton University Press.

Huang, H.-C. and Cressie, N. (1996). Spatio-Temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics and Data Analysis*, **22**, 159-175.

Lee, J.-H. and Shin, K-I. (2004). On the efficiency of outlier cleaners in spatial data analysis. *Journal of Korean Applied Statistics*, **17**, 2, 327-336.

Nierel, R., Moira. and Mugglestone, M. A. (1998). Outlier-Robust spectral estimation for spatial lattice process. *Communications and Statistics: Theory and Methods*, **27**, 3095-3111.

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.

Park, J-J. Shin, K-I. and Cho, K. (2004). Selection of appropriate data transformation for analyzing geostatistics: Case study with data of greenhouse white flies on greenhouse cherry tomatoes. *Journal of Asia-Pacific Entomology*, To appear.

Taylor, L. R. (1961) Aggregation, variance and the mean. *Nature*, **189**, 732-735.

Tobin, P. C., and Bjornstad O. N. (2003) Spatial dynamics and cross-correlation in a transient predator-prey system. *Journal of Animal Ecology*, 72, 460-467.

Wikle, C. K. and Cressie, N. (1999). A dimension reduction approach to space-time Kalman filtering. *Biometrika*, **86**, 815-829.

[2004년 4월 접수, 2004년 7월 채택]

Space Time Data Analysis for Greenhouse Whitefly *

Jin-Mo Park [1] Key-Il Shin [2]

ABSTRACT

Recently space-time model in spatial data analysis is widly used. In this paper we applied this model to analysis of greenhouse whitefly. For handling time component, we used ARMA model and autoregressive error model and for outliers, we adapted Mugglestone's method. We compared space-time models and geostatistic model with MSE and MAPE.

Keywords: Variogram, ARMA model, Outlier cleaner.

* This work was funded by a grant from the KOSEF(R01-2003-000-1024-0)

1) Graduate student, Department of statistics, Hankuk University of Foreign Studies, San 79, Wangsan, Mohyun, Yongin, Kyonggi, 449-791 Korea
E-mail : jmpark@stat.hufs.ac.kr

2) Professor, Department of statistics, Hankuk University of Foreign Studies, San 79, Wangsan, Mohyun Yongin, Kyonggi, 449-791 Korea
E-mail : keyshin@hufs.ac.kr