Journal of the Korean Data & Information Science Society
2017, **28**(1), 11–20

http://dx.doi.org/10.7465/jkdi.2017.28.1.11
한국데이터정보과학회지

여러 가지 가중행렬을 가진 공간 시계열 모형들의 예측[†]

이성덕[1] · 주수인[2] · 이소현[3]

[12]충북대학교 정보통계학과 · [3]국립보건연구원 유전체역학과

접수 2016년 12월 19일, 수정 2017년 1월 3일, 게재확정 2017년 1월 4일

요 약

시간의 변화뿐만 아니라 공간 위치의 변화를 함께 고려한 자료를 공간 시계열 자료라고 한다. 공간 시계열 자기회귀 이동평균 모형과 공간 시계열 중선형 모형에 대해 소개하고 각각의 Kalman Filter 방법에 의한 모수 추정의 과정을 거쳐 최종 선택된 모형의 예측력을 비교하였다. 또한 공간 시계열 자료의 모형에 포함되는 가중행렬에 대하여 기존의 방법인 동일한 가중치와 더불어 거리에 비례한 가중치와 인구수에 비례한 가중치를 제안하였다. 실증분석을 위해 한국질병관리본부에서 수집한 유행성 이하 선염 자료를 활용하여 가중치를 달리한 공간 시계열 모형을 적합시키고 예측하였다. 예측 오차 제곱합을 활용하여 어느 모형이 가장 효과적인 모형인지 판정하였다.

주요용어: 가중행렬, 예측 오차 제곱합, 유행성 이하 선염, Kalman Filter, STARMA 모형, STBL 모형.

1. 서론

특정 위치에서 관측되어진 자료가 과거의 시간과 주변 공간의 영향을 동시에 받는 성질을 가진 자료를 공간 시계열 자료 (spatial time series data)라고 한다. 공간 시계열 선형 모형으로 공간 시계열 자기회귀 이동평균 (space-time series autoregressive moving average; STARMA) 모형과 공간 시계열 중선형 (space-time bilinear; STBL) 모형을 고려하고자 한다. 공간 시계열 모형은 기존 시계열 모형에 공간의 위치를 반영하는 가중행렬이 포함된다. 가중행렬이 지리적으로 인접한 지역일수록 공간 의존도 (spatial dependence)가 높은 것을 반영하는데 본 연구는 인접한 지역에 동일한 가중치를 주는 기존의 방법과 더불어 인구수에 비례하여 인구수가 많은 곳에 더 큰 가중치를 주는 방법과 인접한 지역들 사이의 거리에 비례하여 거리가 가까운 곳에 더 큰 가중치를 주는 방법을 제안하였다.

기존의 연구로 Cliff와 Ord (1975)는 자기회귀 이동평균 (autoregressive moving average; ARMA) 모형을 확장하여 공간 시계열 자료를 적합할 수 있는 STARMA 모형을 제시하였다. 이 모형은 기존의 ARMA 모형에 공간항을 더 추가시켜 모형 자체를 시간과 공간으로 확장시킨 것이다. Pfeifer와 Deutsh (1980a)은 STARIMA 모형의 특성과 모형 구축의 절차를 제시하였다. 또한, Pfeifer와 Deutsh (1980b)는 공간 시계열 자기상관함수 (space-time autocorrelation function; STACF)와 공간 시계열 부분 자기상관함수 (space-time partial autocorrelation function; STPACF)에 대해 제시하고, 공간

[†] 이 논문은 2016년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 이공학 개인기초연구지원사업 연구비 지원에 의하여 연구되었음 (NRF-2016R1D1A3B03932557).
[1] 교신저자: (28644) 충북 청주시 서원구 충대로 1, 충북대학교 정보통계학과, 교수.
E-mail: sdlee@chungbuk..ac.kr
[2] (28644) 충북 청주시 서원구 충대로 1, 충북대학교 정보통계학과, 석사과정.
[3] (28159) 충북 청주시 흥덕구 오송읍 오송생명2로 187, 국립보건연구원 유전체역학과, 연구원.

시계열 자기회귀 (space-time autoregressive; STAR) 모형의 STAR(1_1) 모형과 공간 시계열 이동평균 (spce-time moving average; STMA) 모형의 STMA(1_1) 모형에 대해 추론하였다. Dai와 Billard (1998)는 선형 모형인 ARMA 모형을 비선형으로 확장 시킨 BL (bilinear) 모형에 다시 공간 시계열 자료를 적합시킬 수 있도록 확장시킨 공간 중선형 시계열 모형 (STBL)을 제안하고 식별 과정을 제시하였다. Dai와 Billard (2003)은 STBL 모형에 대한 모수 추정 방법 및 STBL 모형의 최대우도추정 (MLE)을 제시하였다. Lee 등 (2005)은 미국의 12개 주에 대한 유행성 이하 선염 (mumps) 자료에 대해 STAR 모형과 STARMA 모형에 대한 추정과 예측력을 비교하였다. 또한 Kim 등 (2005)은 Kriging 모형과 통계학적 공간자료 분석 모형인 지리적 가중 회귀 모형을 고려하고 미지의 위치에 대한 예측력을 비교하였다. Sung과 Sohn (2013)은 일반화 극단분포를 활용해 서울시 강우량의 사후예측분포를 생성한 후, 실제값과 비교하였다. Park (2015)은 큰 공간 데이터로 인해 발생하는 계산적인 문제를 해결하기 위한 세 가지 방법을 비교한 결과, 마코프 가우스 체 (Markov Gaussian field) 접근법이 가우스 체 (Gaussian field)에 적합하다는 것을 확인하였다.

본 논문의 2절에서는 여러 공간시계열 모형으로 STARMA 모형과 STBL 모형을 소개하고 가중행렬에 대해 새로운 방법을 제안하였다. 3절에서는 공간시계열 모형의 모수 추정 방법 중 하나인 Kalman Filter 방법과 예측에 대해서 살펴보았으며, 4절에서는 전염성이 강한 질병인 Mumps (유행성 이하 선염) 자료로 실증분석을 실시하여 각 모형에 대한 예측 결과를 비교하고, 마지막으로 5절에서는 결론을 맺었다.

2. 공간 시계열 모형

2.1. STARMA (Space-Time Autoregressive Moving Average) 모형

대표적인 공간 시계열 선형 모형인 STARMA 모형은 Cliff와 Ord (1975)에 의해서 제안되었다. STARMA 모형은 시간과 공간에 의존하는 모형으로 $\lambda = [\lambda_1, \lambda_2, \cdots, \lambda_p]$, $\eta = [\eta_1, \eta_2, \cdots, \eta_q]$인 경우에 STARMA (p_λ, q_η)로 표기하며 모형은 다음과 같다.

$$z(t) = \sum_{i=1}^{p} \sum_{m=0}^{\lambda_i} \phi_m^i W^{(m)} z(t-i) - \sum_{j=1}^{q} \sum_{n=0}^{\eta_j} \theta_n^j W^{(n)} e(t-j) + e(t).$$

여기서, $z(t)$는 $[z_1(t), \cdots, z_n(t)]^T$인 $n \times 1$확률벡터과정 (random vector process)
p는 최대 자기회귀차수, q는 최대 이동평균차수
λ_i는 i번째 자기회귀항의 차수, η_j는 j번째 이동평균항의 차수
ϕ_m^i은 시간차수 i와 공간차수 m의 자기회귀 모수
θ_n^j은 시간차수 j와 공간차수 n의 이동평균 모수
$W^{(m)}$은 공간차수 m에 대한 $n \times n$가중행렬
$e(t) = [e_1(t), e_2(t), \cdots, e_n(t)]^T$는 백색잡음 (white noise)벡터이다.

예를 들어, STARMA $(1_1,\ 1_1)$의 모형식은 다음과 같다.

$$z(t) = \phi_0 W^{(0)} z(t-1) + \phi_1 W^{(1)} z(t-1) + \theta_0 W^{(0)} e(t-1) + \theta_1 W^{(1)} e(t-1) + e(t). \tag{2.1}$$

2.2. STBL (Space-Time Bilinear) 모형

Dai와 Billard (1998)가 제안한 STBL 모형은 비선형항인 자기회귀항과 이동평균항이 서로 곱해져 있는 비선형항을 갖는 대표적인 공간 시계열 비선형 모형이다. STARMA 모형은 STBL 모형의 특수한 경우로써 STBL 모형의 비선형항이 모두 0으로 항이 나타나지 않는 경우이다.

$\lambda = [\lambda_1, \lambda_2, \cdots, \lambda_p]$, $\eta = [\eta_1, \eta_2, \cdots, \eta_q]$, $\xi = [\xi_1, \xi_2, \cdots, \xi_r]$, $\mu = [\mu_1, \mu_2, \cdots, \mu_s]$인 경우에 STBL $(p_\lambda,\ q_\eta,\ r_\xi,\ s_\mu)$로 표기하며 모형은 다음과 같다.

$$z(t) = \sum_{i=1}^{p}\sum_{m=0}^{\lambda_i}\phi_m^i[W^{(m)}z(t-i)] + \sum_{j=1}^{q}\sum_{n=0}^{\eta_j}\theta_n^j[W^{(n)}e(t-j)] + \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{m=0}^{\xi_i}\sum_{n=0}^{\mu_j}\beta_{mn}^{ij}[W^{(m)}z(t-i)]WELL[W^{(n)}e(t-j)].$$

여기서, r은 중선형항의 최대 자기회귀차수, s는 중선형항의 최대 이동평균차수
λ_i는 i번째 자기회귀항의 차수, η_j는 j번째 이동평균항의 차수
ξ_i는 i번째 중선형항의 최대 자기회귀차수, μ_j는 j번째 중선형항의 최대 이동평균차수
ϕ_m^i는 공간차수가 m, 시간차수가 i인 자기회귀 모수
θ_n^j는 공간차수가 n, 시간차수가 j인 이동평균 모수
β_{mn}^{ij}은 자기회귀와 이동평균의 i, j번째 시간차수에서 m, n번째 공간차수를 갖는 중선형 모수 이다.
#은 행렬의 원소 간의 곱을 나타낸다. 예를 들어, $A = (a_{ij})$, $B = (b_{ij})$에서 $a_{ij}b_{ij}$의 값을 계산하는 연산자이다. 나머지 기호는 앞의 STARMA 모형의 정의와 같다.

예를 들어, STBL$(1_1, 1_1, 1_1, 1_1)$의 모형식은 다음과 같다.

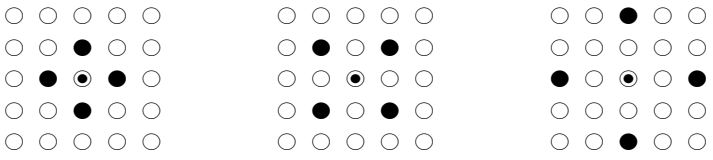
$$\begin{aligned} z(t) =& \phi_0 W^{(0)}z(t-1) + \phi_1 W^{(1)}z(t-1) + \theta_0 W^{(0)}e(t-1) + \theta_1 W^{(1)}e(t-1) \\ &+\beta_{00}z(t-1)WELLe(t-1) + \beta_{01}z(t-1)WELL[W^{(1)}e(t-1)] \\ &+\beta_{10}[W^{(1)}z(t-1)]WELLe(t-1) + \beta_{11}[W^{(1)}z(t-1)]WELL[W^{(1)}e(t-1)] + e(t), \\ & t = 1, 2, \cdots, T. \end{aligned} \tag{2.2}$$

2.3. 가중행렬 (Weight Matrix)

가중행렬 구성은 지리적으로 인접한 지역일수록 공간의존도 (spatial dependence)가 높다는 것에서 착안하여 두 지역 간의 상호작용 여부를 고려하였다. 가중행렬은 모든 지역의 인접성 여부를 통해 동일한 크기의 가중치를 가지고 있기 때문에 그대로 모형에 적용시킬 경우 해석상 중대한 오류가 발생하게 되므로 표준화를 하여 사용하며 다음과 같이 나타낼 수 있다. 한 지역과 연계되어 있는 지역들의 가중치의 합은 1로 지정해준다.

$$W^{(l)} = \begin{cases} 1/n^l, & n^l : l\text{번째 이웃하고 있는 이웃의 수} \\ 0, & \text{그외의 경우.} \end{cases}$$

가중행렬 $W^{(l)}$은 공간에서 이웃하고 있는 순서에 관계가 있다. 가중행렬에서 1차 이웃값 (first order neighbour)은 한 위치에서 가장 가깝고 동일한 거리를 갖는 값의 집합이고, 2차 이웃값 (second order neighbour)은 1차 이웃값보다 약간 먼 동일한 위치에 있는 값들로 나타낸다. 3차 이웃값도 동일한 방법으로 표현할 수 있고, Figure 2.1은 관심 있는 위치에서 첫 번째 이웃하고 있는 구조, 두 번째 이웃하고 있는 구조, 세 번째 이웃하고 있는 구조의 세 가지 경우를 표현한 것이다.



(a) first order (b) second order (c) third order

Figure 2.1 Neighborhood structure in spatial orders

하지만 지역이 인접하다 할지라도 영향이 미치는 정도는 다를 것이라 예상할 수 있다. 따라서 본 연구는 여러 가지 가중행렬을 고려하였다. 즉, 인구수가 많으면 더 많은 교류가 있고, 그에 따라 영향을 미치는 정도가 더 커질 것이라고 기대할 수 있다. 따라서 인구수가 많은 지역에 더 큰 가중치를 주는 방법을 추가적으로 고려하였다. 예를 들어, 강원도에 인접해 있는 경기, 충북, 경북의 가중치를 구하고자 할 경우 각 지역의 인구수를 구한 뒤, 그 인구수에 비례하여 계산한다. 즉, 경기에는 0.777, 충북은 0.050, 경북은 0.173으로 가중치를 줄 수 있다. 또한 경도, 위도 상의 좌표 거리를 고려해 거리가 가까운 지역에 더 큰 가중치를 주는 방법도 고려하였다. 공간 자료는 일반적으로 거리가 가까울수록 상관관계가 높고, 멀어질수록 상관관계가 낮아지는 특성을 가지고 있다. 전염병에 대한 공간상관성을 확인하고자 한다면 공간상의 거리가 먼 지역보다 가까운 지역으로의 전염이 더 잘되므로 거리가 가까운 지역에 더 큰 가중치를 주는 것이다. 예를 들어, 강원도에 인접해 있는 경기, 충북, 경북의 가중치를 구하고자 하는 경우 각 도의 도청 소재지를 기준으로 경도, 위도 상의 좌표 위치를 조사한 다음, 그 좌표 위치를 이용하여 두 지역의 거리를 계산한 후에 그 거리에 비례하여 가중행렬을 구한다. 즉, 경기에는 0.518, 충북에는 0.324, 경북에는 0.158의 가중치를 구할 수 있다.

3. Kalman Filter 방법을 이용한 모수 추정과 예측

3.1. 상태 공간 모형

Kalman과 Bucy (1961)에 의해 도입된 상태 공간 모형은 두 개의 방정식, 즉 시점 t에서의 체계의 상태를 표현하는 상태 방정식과 관측되지 않은 상태 벡터와 관측 오차의 함수인 관측 방정식으로 구성된다. 이렇게 상태 공간 모형으로 표현된 모형은 Kalman Filter를 통하여 모형을 반복적으로 계산함으로써 추정오차를 줄이는 모형을 선택할 수 있다 (Harvey, 1990).

STARMA$(1_1,\ 1_1)$의 식 (2.1)을 상태 공간 모형으로 표현하면 다음과 같다.

$$y(t) = Zb(t)b$$
$$(t) = T_t b(t-1) + Fe(t).$$

여기서, $y(t) = z(t)$이고 $Z = [I_n, O_n]$, $b(t) = \begin{bmatrix} z(t) \\ e(t) \end{bmatrix}$, $F = \begin{bmatrix} I_n \\ O_n \end{bmatrix}$, $T_t = \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix}$이며, $a = \phi_0 + \phi_1 W^{(1)}$, $b = \theta_0 + \theta_1 W^{(1)}$, I_n은 단위행렬, O_n은 영행렬이다.

STBL$(1_1, 1_1, 1_1, 1_1)$의 식 (2.2)를 상태 공간 모형으로 표현하면 다음과 같다.

$$y(t) = Zb(t),$$
$$b(t) = T_t[y(t)]b(t-1) + Fe(t).$$

여기서, $y(t) = z(t)$이고 $Z = [I_n, O_n]$, $b(t) = \begin{bmatrix} z(t) \\ e(t) \end{bmatrix}$, $F(t) = \begin{bmatrix} I_n \\ I_n \end{bmatrix}$, $T_t[y(t)] = \begin{bmatrix} a & b + c(t) \\ 0 & 0 \end{bmatrix}$, $a = \phi_0 + \phi_1 W^{(1)}$, $b = \theta_0 + \theta_1 W^{(1)}$이고, $c(t) = \beta_{00} diag[z(t-1)] + \beta_{01} diag[z(t-1)] W^{(1)} + \beta_{10} diag[W^{(1)} z(t-1)] + \beta_{11} diag[W^{(1)} z(t-1)] W^{(1)}$, I_n은 단위행렬, O_n은 영행렬이다.

3.2. Kalman Filter 방법에 의한 모수 추정

Kalman Filter 방법은 상태 공간 모형을 추정하는 도구로써 상태 공간 모형에서 $u(t)$와 $v(t)$의 정규성이 만족될 때, 시점 t에서 이용 가능한 정보에 기초하여 시점 t의 비관측 변수들인 상태 공간 벡터 $b(t)$의 조건부 평균과 분산을 반복과정을 통해 계산해 내는 기법이다. 시점 t까지의 이용 가능한 정보를 바탕으로 $b(t)$를 추정하는 것이 기본 필터 (basic filter)이다. Kalman Filter 방정식은 새로운 관측치가

사용 가능하게 됨과 동시에 추정치가 Update 되는 것을 가능하게 해주는 방정식의 집합이라고 할 수 있다. $\beta(t|t-1)$에 의한 관찰치 $y(1), y(2), \cdots, y(t-1)$가 주어졌을 때 상태 벡터인 $b(t)$를 t시점에서의 실현치에 의한 모수 추정치로 정의하자. 정의에 의하여 행렬 $\beta(t|t-1)$에 대한 평균 잔차 제곱합 행렬 (mean squared error matrix)은 다음과 같이 표현할 수 있다.

$$\sigma^2 C(t|t-1) = E[\beta(t|t-1) - b(t)][\beta(t|t-1) - b(t)]^T.$$

t번째의 실현치가 나타나는 시간에 우리는 $\beta(t|t-1)$과 $C(t|t-1)$을 얻는데 이들은 $(t-1)$번째의 실현치에 의해 계산될 수 있다. t 번째의 실현치가 나타나는 $C(t|t)$를 따라 추정치 $\beta(t|t)$를 계산할 수 있으며 이 과정은 다음과 같은 예측 방정식으로 표현할 수 있다.

$$\beta(t|t) = \beta(t|t-1) + C(t|t-1)Z^T H^{-1}(t)v(t),$$
$$C(t|t) = C(t|t-1) - C(t|t-1)Z^T H^{-1}(t)ZC(t|t-1).$$

여기서, $v(t) = y(t) - Z\beta(t|t-1)$이고 $H(t) = ZC(t|t-1)Z^T$이다. $v(t)$는 1단계의 예측 추정치의 오차이며 $\sigma^2 H(t)$는 $v(t)$에 대한 분산공분산 행렬이다. 이 경우 우리는 다음의 예측 방정식에 의해 주어진 $C(t+1|t)$를 따라 1단계 예측 추정치인 $\beta(t+1|t)$를 계산함으로써 t번째의 실현치를 완성시킬 수 있다.

$$\beta(t+1|t) = T[y(t+1)]\beta(t|t), \tag{3.1}$$
$$C(t+1|t) = T[y(t+1)]C(t|t)T[y(t+1)]^T + FQ(t+1)F^T. \tag{3.2}$$

상태벡터 $\{b(t)\}$가 정상적이라는 가정 하에서 Kalman Filter의 초기치로 사용할 추정치가 주어져야 하는데 이 값은 $\beta(1|0) = E[b(0)]$, $C(1|0) = FQ(0)F^T = \sigma^{-2}var[b(0)]$으로 각각 부여할 수 있다.

3.3. 예측

모수 추정 단계를 거쳐 주어진 자료의 공간 시계열에 대한 모형화를 완료하면 예측을 하게 되는데 예측 역시 Kalman Filter에 의한 방법으로 이루어진다. 즉, 예측은 Kalman Filter의 예측 방정식을 풀어서 얻을 수 있는 Kalman Filter 방정식의 종합적인 과정이다. 예측방정식은 관찰방정식 $y(t) = Z_t[(y(t)]b(t) + u(t),\ t = 1, 2, \cdots$와 동치를 이루고 있으며 다음과 같이 나타낼 수 있다.

$$\bar{y}(t+1|t) = Z\beta(t+1|t). \tag{3.3}$$

식 (3.3)에서는 예측치인 $\bar{y}(t+1|t)$를 제공해준다. 예측 오차인 $v(t)$의 값 또한 계산할 수 있다. 때때로 상태 벡터의 다중 예측치 (multi-step-ahead estimate)가 요구되는 경우가 있는데 이는 $k > 1$인 경우에 $\beta(t+k|t)$의 추정치가 필요함을 의미한다. 이 과정은 식 (3.1)의 예측 방정식을 k번 반복적으로 적용시킴으로써 수행되어질 수 있다.

4. 실증분석

4.1. 자료설명

실증분석에서 사용한 자료는 Mumps (유행성 이하 선염) 자료이다. Mumps는 공간적, 시간적으로 퍼져나가는 전염성이 강한 질병으로 예방백신을 접종받지 않으면 청소년과 노년층 사이에서 주기적으로 크게 발생하는 특성이 있다.

실증분석을 위해 한국질병관리본부에서 한국의 16개 시도 (서울, 부산, 대구, 인천, 대전, 광주, 울산, 경기, 강원, 충남, 충북, 전북, 전남, 경북, 경남, 제주)에 대해서 2001년 1월부터 2008년 8월까지의 Mumps 자료를 사용하였다. 한편 공간적인 관점에서 자료 분석의 효율성을 위해 16개 지역 중 공간적으로 같은 공간이라고 볼 수 있는 8개 지역 (경기, 강원, 충남, 충북, 전북, 전남, 경북, 경남)으로 재분류하였다. Table 4.1은 각 지역에서 첫 번째 이웃하고 있는 지역의 집합을 나타내고 있다.

Table 4.1 1st neighborhood set in regions

Location	Region	1st neighborhood region
1	Gyeonggi	Gangwon, Chungnam, Chungbuk
2	Gangwon	Gyeonggi, Chungbuk, Gyeongbuk
3	Chungnam	Gyeonggi, Chungbuk, Jeonbuk
4	Chungbuk	Gyeonggi, Gangwon, Chungnam, Jeonbuk, Gyeongbuk
5	Jeonbuk	Chungnam, Chungbuk, Jeonnam, Gyeongbuk, Gyeongnam
6	Jeonnam	Jeonbuk, Gyeongnam
7	Gyeongbuk	Gangwon, Chungbuk, Jeonbuk, Gyeongnam
8	Gyeongnam	Jeonbuk, Jeonnam, Gyeongbuk

4.2. 자료의 사전 처리와 가중행렬 구조 설명

지역별 Mumps 자료는 가산자료로 정상성 가정을 만족시키기 위해 분산 안정화 변환, 표준화 과정 및 계절차분을 실시하였다. 다음의 Figure 4.1은 8개 지역 중 전남 지역 하나의 Mumps 자료의 시계열도이다. 앞에서 STBL 모형의 특징을 언급했던 것처럼 불규칙적인 시간 주기로 갑자기 매우 큰 진폭이 발생하는 것을 볼 수 있으므로 STBL 모형에 더 적합하리라는 예상을 할 수 있다.

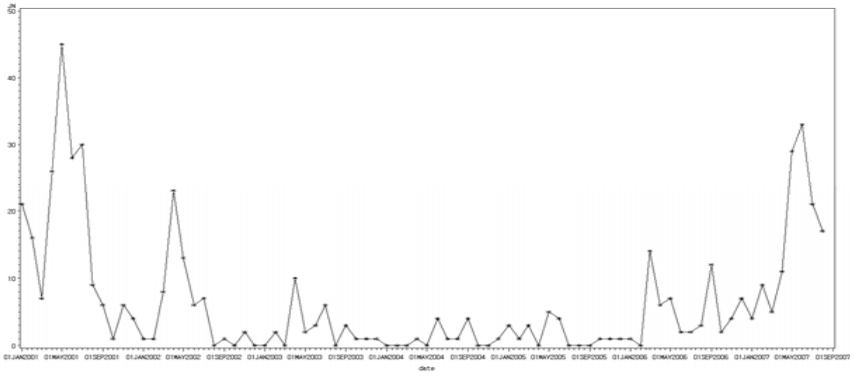


Figure 4.1 Time series plot of mumps data in Jeonnam

각 지역의 가중행렬을 고려한 첫 번째 방법은 1차로 이웃하고 있는 지역에 각각 동일한 가중치를 주는 방법, 두 번째는 인구수에 비례하여 1차 이웃 지역 중 인구수가 많은 곳에 더 큰 가중치를 주는 방법, 마지막 세 번째는 거리에 비례하여 1차 이웃 지역 중 거리가 가까운 곳에 더 큰 가중치를 주는 방법을 사용하였다. 각 방법에 대한 가중행렬은 다음과 같이 나타낼 수 있다.

• Weight matrix of equal proportion allocation

$$W_1^{(1)} = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 & 0 & 1/3 & 0 \\ 1/3 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 0 & 1/5 & 1/5 & 0 & 0 \\ 0 & 0 & 1/5 & 1/5 & 0 & 1/5 & 1/5 & 1/5 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 1/4 & 0 & 1/4 & 1/4 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{bmatrix}$$

• Weight matrix of reciprocal proportion with population

$$W_2^{(1)} = \begin{bmatrix} 0.000 & 0.234 & 0.533 & 0.233 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.777 & 0.000 & 0.000 & 0.050 & 0.000 & 0.000 & 0.173 & 0.000 \\ 0.875 & 0.000 & 0.000 & 0.056 & 0.069 & 0.000 & 0.000 & 0.000 \\ 0.661 & 0.043 & 0.097 & 0.000 & 0.052 & 0.000 & 0.147 & 0.000 \\ 0.000 & 0.000 & 0.161 & 0.070 & 0.000 & 0.156 & 0.245 & 0.368 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.190 & 0.000 & 0.000 & 0.810 \\ 0.000 & 0.119 & 0.000 & 0.118 & 0.145 & 0.000 & 0.000 & 0.618 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.177 & 0.321 & 0.503 & 0.000 \end{bmatrix}$$

• Weight matrix of reciprocal proportion with distance

$$W_3^{(1)} = \begin{bmatrix} 0.000 & 0.497 & 0.242 & 0.260 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.518 & 0.000 & 0.000 & 0.324 & 0.000 & 0.000 & 0.158 & 0.000 \\ 0.131 & 0.000 & 0.000 & 0.716 & 0.153 & 0.000 & 0.000 & 0.000 \\ 0.104 & 0.124 & 0.526 & 0.000 & 0.137 & 0.000 & 0.110 & 0.000 \\ 0.000 & 0.000 & 0.189 & 0.231 & 0.000 & 0.180 & 0.245 & 0.156 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.632 & 0.000 & 0.000 & 0.368 \\ 0.000 & 0.140 & 0.000 & 0.255 & 0.336 & 0.000 & 0.000 & 0.269 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.352 & 0.205 & 0.442 & 0.000 \end{bmatrix}$$

4.3. 공간 시계열 모형의 모수 추정

공간 시계열 자기상관함수 (spatial time autocorrelation function)와 공간 시계열 부분 자기상관함수 (spatial time partial autocorrelation function)를 이용하여 STARMA(1_1, 1_1) 모형과 STBL $(1_1, 1_1, 1_1, 1_1)$ 모형으로 식별하였다.

Kalman Filter 방법으로 추정한 STARMA$(1_1, 1_1)$ 모형의 모수 값은 Table 4.2와 같고, STBL $(1_1, 1_1, 1_1, 1_1)$ 모형의 모수 값은 다음의 Table 4.3과 같다.

Table 4.2 Parameter estimation of STARMA$(1_1, 1_1)$model with weight matrices

Coefficient	Weight		
	Equal	Population	Distance
ϕ_0	0.6618	0.6614	0.6789
ϕ_1	0.2310	0.1739	0.1830
θ_0	-0.2506	-0.2448	-0.2691
θ_1	-0.2514	-0.1715	-0.1765
MSE	0.7185	0.7213	0.7208
AIC	-171.8197	-169.7055	-170.0877

Table 4.3 Parameter estimation of STBL$(1_1, 1_1, 1_1, 1_1)$model with weight matrices

Coefficient	Weight		
	Equal	Population	Distance
ϕ_0	0.6781	0.6776	0.6687
ϕ_1	0.2097	0.1603	0.2018
θ_0	-0.2983	-0.2899	-0.2904
θ_1	-0.2458	-0.1699	-0.2095
β_{00}	0.0124	0.0163	0.0165
β_{01}	-0.2489	-0.1340	-0.1689
β_{10}	-0.0484	-0.1196	-0.1285
β_{11}	0.0565	0.0438	0.0342
MSE	0.6943	0.6957	0.6916
AIC	-182.4807	-181.4026	-184.6218

모수 추정 결과, STBL 모형에서의 MSE 값과 AIC 값이 STARMA 모형에서보다 대체적으로 더 작은 것을 알 수 있고, STBL 모형에서는 거리에 비례한 가중치를 이용했을 때 MSE 값과 AIC 값이 가장 작게 나왔다.

4.4. 공간 시계열 모형에서의 예측

8개 지역의 Mumps 자료에 대해서 2007년 9월부터 2008년 8월까지의 12개월의 자료를 예측하고, 예측된 값과 실제 관측값을 예측 오차 제곱합 (sum of square forecast error; SSF)을 이용하여 비교하였다. SSF를 계산하는 식은 다음과 같다.

$$SSF = \sum_{i=1}^{12}\sum_{j=1}^{12}(\text{관측값}_{ij} - \text{예측값}_{ij})^2,\ i = \text{도},\ j = \text{월}\ .$$

Table 4.4 Comparison with SSF for weight matrices

Year/Month (province)	Obs.	Equal		Population		Distance	
		STARMA $(1_1, 1_1)$	STBL $(1_1, 1_1, 1_1, 1_1)$	STARM A $(1_1, 1_1)$	STBL $(1_1, 1_1, 1_1, 1_1)$	STARMA $(1_1, 1_1)$	STBL $(1_1, 1_1, 1_1, 1_1)$
0711 (Gyeonggi)	117	17	17	16	17	17	16
0712 (Gyeonggi)	128	17	17	16	17	17	17
0711 (Chungnam)	20	22	22	22	22	21	23
0712 (Chungnam)	37	10	10	10	9	9	10
0711 (Chungbuk)	21	9	9	9	9	9	9
0712 (Chungbuk)	20	12	12	12	12	11	12
0711 (Gyeongbuk)	186	84	85	84	83	82	85
0712 (Gyeongbuk)	278	172	173	172	170	170	174
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
SSF		533322	532259	535091	534396	536443	532241

Table 4.5 Comparison of SSF for models

Weight	STARMA$(1_1, 1_1)$	STBL$(1_1, 1_1, 1_1, 1_1)$
Equal	533322	532259
Distance	536443	532241
Population	535091	534396

Table 4.4와 Table 4.5와 같이 가중치를 달리한 각각의 모형에 대한 2007년 9월부터 2008년 8월까지의 실제 관측값과 예측된 값과 함께 각 모형의 예측 오차 제곱합의 값을 구하여 비교한 결과, STARMA$(1_1, 1_1)$ 모형보다는 STBL$(1_1, 1_1, 1_1, 1_1)$ 모형의 예측력이 더 높은 것으로 나타나 STBL $(1_1, 1_1, 1_1, 1_1)$모형이 더 적합하다고 할 수 있고, 그 중 거리에 비례한 가중치를 사용하여 예측한 결과의 예측력이 더 좋은 것으로 나타났다.

5. 결론

질병관리본부에서 제공한 Mumps 자료를 활용하여 실증분석을 실시하였다. STARMA 모형, STBL 모형으로 모형을 가정하고 모수 추정은 Kalman Filter 방법을 사용하였다. 또한 가중치를 인구수와 거리에 대해 달리 하였다. 그 결과, 세 가지 방법의 가중치에 대해 각각 STARMA$(1_1, 1_1)$, STBL$(1_1, 1_1, 1_1, 1_1)$ 모형이 선택되었다. STARMA$(1_1, 1_1)$ 모형보다 STBL$(1_1, 1_1, 1_1, 1_1)$ 모형에서 SSF 값이 더 작게 나와 본 논문에서 사용한 Mumps 자료는 STARMA$(1_1, 1_1)$ 모형보다 STBL

$(1_1, 1_1, 1_1, 1_1)$ 모형이 더 적합한 것을 알 수 있었다. 또한 거리 비례 가중치를 사용한 STBL$(1_1, 1_1, 1_1, 1_1)$ 모형의 예측력이 가장 좋은 것을 확인할 수 있었다.

References

Kim, S. W., Jeong, A. R. and Lee, S. D. (2005). Comparison between Kriging and GWR for the spatial data. *The Korean Journal of Applied Statistics*, **18**, 271-280.

Lee, S. D., Kim, I. K., Kim, D. K. and Jeong, A. R. (2009). Bayes inference for the spatial time series model. *The Korean Journal of Applied Statistics*, **16**, 31-40.

Dai, Y. and Billard, L. (1998). A space-time bilinear model and its identification. *Journal of Time Series Analysis*, **19**, 657-679.

Dai, Y. and Billard, L. (2003). Maximum likelihood estimation in space time bilinear model. *Journal of Time Series Analysis*, **24**, 25-44.

Cliff, A. D. and Ord, J. K. (1975). Space time modeling with an application to regional forecasting. *Transactions of the institute of British Geographers*, **64**, 119-128.

Harvey, A. C. (1990). F *orecasting, structural time series models and the kalman filter*, Cambridge University Press, Cambridge.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering*, **83**, 95-108.

Park, J. H. (2015). Review on statistical methods for large spatial Gaussian data. *Journal of the Korean Data & Information Science Society*, **26**, 495-504.

Pfeifer, P. E. and Deutsch, S. J. (1980a). A three-stage iterative procedure for space-time modeling. *Technometrics*, **22**, 25-47.

Pfeifer, P. E. and Deutsch, S. J. (1980b). Identification and interpretation of first order space-time ARMA models. *Technometrics*, **22**, 397-408.

Sung, Y. K. and Sohn, J. K. (2013). Prediction of extreme rainfall with a generalized extreme value distribution. *Journal of the Korean Data & Information Science Society*, **24**, 857-865.

Journal of the Korean Data & Information Science Society
2017, **28**(1), 11–20

http://dx.doi.org/10.7465/jkdi.2017.28.1.11
한국데이터정보과학회지

Prediction for spatial time series models with several weight matrices[†]

Sung Duck Lee[1] · Su In Ju[2] · So Hyun Lee[3]

[12]Department of Information and Statistics, Chungbuk National University
[3]Department of genome Epidemiology, Korea National Institute of Health

Received 19 December 2016, revised 3 January 2017, accepted 4 January 2017

Abstract

In this paper, we introduced linear spatial time series (space-time autoregressive and moving average model) and nonlinear spatial time series (space-time bilinear model). Also we estimated the parameters by Kalman Filter method and made comparative studies of power of forecast in the final model. We proposed several weight matrices such as equal proportion allocation, reciprocal proportion between distances, and proportion of population sizes. For applications, we collected Mumps data at Korea Center for Disease Control and Prevention from January 2001 until August 2008. We compared three approaches of weight matrices using the Mumps data. Finally, we also decided the most effective model based on sum of square forecast error.

Keywords: Kalman filter, Mumps, SSF, STARMA model, STBL model, weight matrix.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(NRF-2016R1D1A3B03932557).
[1] Corresponding author: Professor, Department of Information and Statistics, Chungbuk National University, Chungcheongbuk-do 28644, Korea. E-mail: sdlee@chungbuk.ack.kr
[2] Graduate student, Department of Information and Statistics, Chungbuk National University, Chungcheongbuk-do 28644, Korea.
[3] Researcher, Department of genome epidemiology, Korea National Institute of Health, Chungcheongbuk-do 28159, Korea.