한국통계학회논문집
2010, 17권, 2호, 263-273

공간시계열모형의 결측치 추정방법 비교

이성덕[1,a], 김덕기[a]

[a]충북대학교 정보통계학과

요 약

시계열의 결측값은 미지의 모수 또는 확률변수로 취급할 수 있으며 이에 따른 최대가능도방법과 확률변수방법에 의해 결측치를 추정할 수 있으며 또한 주어진 자료 하에서 미지의 값에 대한 조건부기대치로 예측할 수 있다. 이 연구의 주된 목적은 불완전한 자료에 대해 기존에는 ARMA모형만을 고려하였는데 이를 확장하여 공간시계열모형인 STAR모형에 적용하여 두 가지 추정방법을 이용해 결측값의 추정 정밀도를 비교하는데 있다. 사례분석을 위해 한국질병관리본부에서 전산보고 하고 있는 전염병 자료 중에서 2001~2009년 동안의 월별 Mumps 자료를 이용하여 두 가지 추정방법의 추정 정밀도와 예측정확도를 비교하였다.

주요용어: 최대가능도추정법, 확률변수법, STAR모형, Mumps 자료, 가중치행렬, STACF, STPACF, Kalman-Filter, 예측오차제곱합.

1. 서론

최근에 컴퓨터와 인터넷네트워크의 발달로 인한 산업분야의 형태 또한 다양해지고 있으며, 다양한 산업분야에서 효과적인 정보 분석에 대한 요구가 높아지면서 방대한 자료에 대한 데이터웨어하우스(DW)구축에 관심이 높아지면서 각 기업들은 DB구축을 통해 방대한 자료를 저장해 온 것이 현실이지만 정보 분석을 위해 데이터 정제과정을 거쳐야만 효과적인 통계분석이 이루어질 수 있다는 면에서 데이터 정제과정에서 발생하는 결측치(Missing Values)를 어떻게 처리할 것인가가 통계학에서 꾸준히 다뤄 온 과제 중 하나이다. 최근의 결측치를 보정해 주는 방법으로 MCMC(Markov Chain Monte Carlo), EM(Expectation and Maximization) 알고리즘이 널리 이용되고 있다.

본 연구는 DB자료의 상당수가 시간의 흐름에 따라 얻어지는 시계열자료의 특성을 갖고 있어서 시계열 자료에서 존재하는 결측치를 대체하는 방법을 연구하였다. 시계열분석에서 발생하는 결측치를 대체하는 방법은 크게 두 가지로 나뉘어져 있는데, 첫째는 최대가능도추정법을 이용하여 결측값을 대체하는 방법이고 둘째는 베이지안 방법을 통한 확률변수방법이다.

일반적으로 결측값을 추정하기 위한 최대가능도추정법의 이용이나 확률변수의 결측값의 이용은 논쟁거리가 되어왔다. 왜냐하면 연구자별로 추정량을 얻기 위해 각기 다른 가능도함수를 사용하기 때문이다 (Bayarri 등, 1986).

시계열자료에서도 각 모형에 따라 가능도함수가 다르기 때문에 이러한 문제점을 갖고 있지만 대부분의 연구자들은 ARMA모형이 시계열모형 중 가장 기본이 되는 모형이기 때문에 이모형에 적용시킬 수 있다면 다른 어떠한 모형에도 적용하기가 쉬울 것으로 생각하였다. Dunsmuir와 Robinson (1981)은 결측값이 존재할 때 결측값 대체방법으로 가능도함수를 이용한 최대가능도방법을 제안했고, Pena와 Tiao (1991)은 결측값 대체방법으로 확률변수방법과 최대가능도방법을 ARMA모형에 대해 제안하였

이 논문은 2008학년도 충북대학교 학술연구지원 사업의 연구비 지원에 의하여 연구되었음.
[1] 교신저자: (361-763) 충북 청주시 흥덕구 충북대학교 정보통계학과, 교수. E-mail: sdlee@cbnu.ac.kr

고, Rubin (1994)는 결측값 대체방법으로 Bootstrap방법을 제안하였고, Kim (2005)은 결측치가 있는 시계열자료에 대해 확률변동모형을 고려하여 EM알고리즘을 이용한 모수추정방법을 제안하였고, Horton과 Kleinman (2007)은 회귀모형에서의 결측치 추정방법을 비교분석하였다.

기존의 논문에서 이성덕과 김덕기 (2009)는 실증분석을 위해 한국질병관리본부에서 제 2군 전염병으로 관리하고 있으며 시간과 공간에 따라 전염성이 매우 강한 특징을 보이는 Mumps(유행성이하선염) 자료를 ARMA모형에 적합 시켜 동일한 결측값을 최대가능도방법과 확률변수방법을 이용하여 추정하고 두 방법에 따른 예측결과를 비교분석하였다.

본 논문에서는 기존의 실증분석 사례로 사용된 자료인 한국의 Mumps자료를 최근 자료까지 업데이트하여 ARMA모형에 대해 앞에서 언급한 두 가지 방법으로 결측값에 대한 추정의 정밀도와 예측력을 비교하였고, 시간과 공간에 따라 전염성이 매우 강한 특징을 잘 반영시켜주는 공간시계열자기회귀(STAR)모형에 대한 결측값 대체방법에 대해 연구하였고, 실증분석에서 16개시도 자료를 8개의 도별 자료로 재분류하고 공간과 시간에 대한 특성을 반영하여 분석하였다.

본 논문의 구성은 2절에서 자기회귀이동평균(ARMA)모형에서의 최대가능도방법과 확률변수방법으로 결측값을 추정하는 방법을 비교 설명하였고, 3절에서는 공간시계열자기회귀(STAR)모형에서의 최대가능도방법과 확률변수방법으로 결측값을 추정하는 방법을 비교 설명하였다. 4절에서는 사례연구로써 Mumps자료에 대해 결측값을 추정하여 추정의 정도를 비교하고, 5절에서는 각 방법에 대한 도별 Mumps자료에 대한 월별 예측력을 비교하고 6절에서는 결론 및 향후 연구 과제를 제시하였다.

2. 자기회귀이동평균(ARMA)모형에서 결측값 대체

2.1. 최대가능도추정법에 의한 결측값 대체

시계열 모형에서 결측값 대체방법을 살펴보기 위해 시계열 z_t는 정상 시계열이라고 가정하자.

$$z_t = \phi z_{t-1} + e_t, \quad e_t \sim iid \text{ Normal}\left(0, \sigma^2\right).$$

계산상의 편리를 위해 ϕ와 σ^2은 알려져 있다고 가정하자. n개의 관찰값 $z_t (t = 1, \ldots, n)$ 중 z_m을 결측값이라 가정하면 $(1 \le m \le n)$에 대하여, $Z_n = (z_1, \ldots, z_n)'$는 $n \times 1$ 벡터로 정의되고 $Z_{(m)}$는 Z_n에서 z_m을 제외하고 얻어진 $(n-1) \times 1$벡터로 정의된다. z_m이 주어진 이용 가능한 자료 $Z_{(m)}$의 결합밀도함수는 다음과 같다.

$$f(Z_{(m)}|z_m) = \frac{f(Z_n)}{f(z_m)}, \tag{2.1}$$

여기서

$$f(Z_n) = \left(2\pi\sigma^2\right)^{\frac{1}{2}n} \left(1-\phi^2\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(1-\phi^2\right)z_1^2 + \sum_{t=2}^{n}(z_t - \phi z_{t-1})^2\right]\right\} \tag{2.2}$$

$$f(z_m) = \left(2\pi\sigma^2\right)^{\frac{1}{2}} \left(1-\phi^2\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(1-\phi^2\right)z_m^2\right]\right\}. \tag{2.3}$$

식 (2.1)에서 z_m는 $Z_{(m)}$에 대한 미지의 모수로 취급할 수 있다. 그러면 z_m의 가능도함수는 다음과 같다.

$$l(z_m|Z_{(m)}) = \left(2\pi\sigma^2\right)^{\frac{1}{2}(n-1)} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{t=1}^{m-1}(z_t - \phi z_{t+1})^2 + \sum_{t=m+1}^{n}(z_t - \phi z_{t-1})^2\right]\right\}, \tag{2.4}$$

여기서 합의 범위가 양의 값이 아니라면 지수항은 보이지 않는다. 그러므로 z_m의 최대가능도추정량은 다음과 같다.

$$\hat{z}_m = \begin{cases} \phi^{-1}\mathbf{z_{m+\delta}}, & m = 1 \text{ or } n, \\ (2\phi)^{-1}(z_{m+1} + z_{m-1}), & 1 < m < n, \end{cases} \tag{2.5}$$

$m = 1$이면 $\delta = 1$이고 $m = n$이면 $\delta = -1$이다. $\hat{z}_m$의 MSE는 식 (2.4)를 통해 다음과 같이 구할 수 있다.

$$\text{MSE}(\hat{z}_m) = \begin{cases} \phi^{-2}\sigma^2, & m = 1 \text{ or } n, \\ \left(2\phi^2\right)^{-1}\sigma^2, & 1 < m < n. \end{cases} \tag{2.6}$$

이를 ARMA(p, q)에 적용하면 다음과 같이 최대가능도추정법에 의해 결측값을 대체할 수 있다.

$$\hat{z}_m = \sum_{i=1}^{p} \frac{z_{m+i} + z_{m-i}}{2\phi_i} + \sum_{j=1}^{q} \frac{z_{m+j} + z_{m-j} - \hat{z}_{m+j} - \hat{z}_{m-j}}{2\theta_j}. \tag{2.7}$$

2.2. 확률변수에 의한 결측값 대체

그 동안 알려진 자료에 대한 결측값에 관하여 식 (2.2)의 결합밀도함수 $f(Z_n)$을 최대로 하는 최소제곱추정량이나 최대가능도추정량을 이용하여 대체하는 연구들이 활발히 진행되어왔다 (Brubacher와 Wilson, 1976). 이 경우 추정량은 다음과 같이 쉽게 얻어진다.

$$\tilde{z}_m = \begin{cases} \phi\mathbf{z_{m+\delta}}, & m = 1 \text{ or } n, \\ \dfrac{\phi}{1+\phi^2}(z_{m+1} + z_{m-1}), & 1 < m < n. \end{cases} \tag{2.8}$$

그러나 이 추정량 $\tilde{z}_m$는 미지의 모수로 고려되는 z_m에 대한 함수 $f(Z_n)$가 결합밀도함수가 아니기 때문에 최대가능도추정량이 아니다. 그러므로 결측값 z_m와 측정값 $Z_{(m)}$에 대한 $f(Z_n)$는 가능도함수가 아니다.

식 (2.8)의 의미를 설명하기 위해 식 (2.2)의 확률구조를 따르는 확률변수 z_m를 고려하자. 그러면 $Z_{(m)}$가 주어진 z_m의 분포는 다음과 같다.

$$f(z_m|Z_{(m)}) = \frac{f(Z_n)}{f(Z_{(m)})}, \tag{2.9}$$

여기서 $f(Z_{(m)})$는 $f(Z_n)$으로 부터 z_m를 제거한 적분값으로 얻어진다. 잘 알려진 대로 식 (2.9)의 분포는 다음과 같은 정규분포이다 (Pena, 1987).

$$E(z_m|Z_{(m)}) = \tilde{z}_m \tag{2.10}$$

$$\text{VAR}(z_m|Z_{(m)}) = \begin{cases} \sigma^2, & m = 1 \text{ or } n, \\ \left(1+\phi^2\right)^{-1}\sigma^2, & 1 < m < n. \end{cases} \tag{2.11}$$

식 (2.10)과 (2.11)이 성립됨을 보이면 다음과 같다. z_m의 사후분포의 기댓값과 분산을 구하기 위해 식 (2.3)과 (2.4)를 이용해 z_m의 사후분포 $f(z_m|Z_{(m)})$을 유도하면 다음과 같다.

$$\begin{aligned} f(z_m|Z_{(m)}) &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}}\left(1-\phi^2\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(1-\phi^2\right)z_1^2 + \sum_{t=2}^{n}(z_t - \phi z_{t-1})^2\right]\right\} \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}}\left(1-\phi^2\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(1-\phi^2\right)z_1^2 + (z_2 - \phi z_1)^2 + \cdots + (z_m - \phi z_{m-1})^2\right.\right. \\ &\quad \left.\left. + (z_{m+1} - \phi z_m)^2 + \cdots\right]\right\} \end{aligned}$$

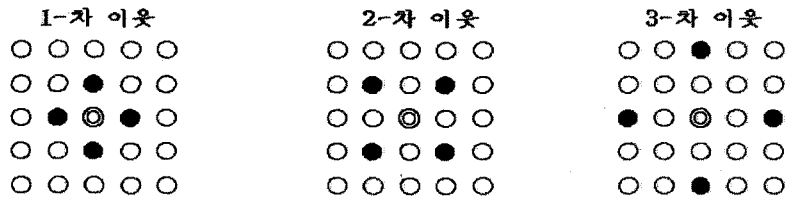


그림 1: *공간차수의 격자자료에서 이웃하고 있는 구조*

$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \left(1-\phi^2\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(1+\phi^2\right)\left[z_m^2 - 2\frac{\phi}{(1+\phi^2)}\left(z_{m-1}+z_{m+1}\right)z_m + C\right]\right\},$$

여기서 C는 z_m에 의존하지 않는 나머지 항을 나타낸다.

그러므로, $E(z_m|Z_{(m)}) = \{\phi/(1+\phi^2)\}(z_{m-1}+z_{m+1})$, $\mathrm{VAR}(z_m|Z_{(m)}) = \sigma^2/(1+\phi^2)$이 된다. 따라서 이를 ARMA$(p,q)$에 적용하면 다음과 같이 확률변수방법에 의해 결측값을 대체할 수 있다.

$$\hat{z}_m = \sum_{i=1}^{p} \frac{\phi_i(z_{m+i}+z_{m-i})}{1+\phi_i^2} + \sum_{j=1}^{q} \frac{\theta_j(z_{m+j}+z_{m-j}-\hat{z}_{m+j}-\hat{z}_{m-j})}{1+\theta_j^2}, \tag{2.12}$$

여기서, ϕ_i은 i-번째 자기회귀계수, θ_j은 j-번째 이동평균계수를 나타낸다.

3. 공간시계열모형에서 결측값 대체

3.1. 공간시계열자기회귀모형(STAR)

공간시계열 자기회귀모형(Space-Time Autoregressive Model)은 STAR(P_{λ_i})로 표현할 수 있으며, 모형은 다음과 같다.

$$Z_t = \sum_{i=1}^{p}\sum_{m=0}^{\lambda_i} \phi_m^i \mathbf{W}^{(m)} Z(t-i) + e(t), \tag{3.1}$$

p : 최대 자기회귀 차수, λ_i : i번째 자기회귀항의 차수
ϕ_m^i : 공간차수가 m, 시간차수가 i인 자기회귀모수
$W^{(m)}$: 공간차수가 m인 $n \times n$ 가중치행렬(Weighting Matrix)
$z(t) = [z_1(t), z_2(t), \ldots, z_n(t)]^T$: $n \times 1$ 확률벡터과정(random vector process)
$e(t) = [e_1(t), e_2(t), \ldots, e_n(t)]^T$: $n \times 1$ 확률잡음벡터(random noise vector)

$$\mathrm{E}\left[e(t)e(t+j)^T\right] = \begin{cases} \sigma^2\mathbf{I_n}, & j=0, \\ 0, & \text{otherwise}, \end{cases} \quad \mathrm{E}\left[\mathbf{z}(t)\, e(t+j)^T\right] = 0, \text{ if } j>0.$$

3.2. 가중치행렬(Weight matrix)

그림 1은 공간차수에서 격자자료(lattice-data)의 1~3차 이웃(neighbor)하고 있는 구조의 예를 나타낸 그림이다. 1차 이웃의 경우 한 위치(◎)에서 가장 가까우며 동일한 유클리디안 거리를 갖는 집합(●)이며, 2차 이웃의 경우 1차 이웃보다 먼 동일한 유클리디안 거리를 갖는 집합으로 나타낸다. 3~4차 이웃의 경우도 마찬가지 방법으로 나타낼 수 있다.

위 $5 \times 5 (n = 25)$ 격자자료의 경우 각 위치는 왼쪽 위에서부터 오른쪽 아랫방향으로 가중행렬(Weighting Matrix)을 그림 2와 같이 결정할 수 있다.

$$W^{(1)} = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \cdots \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \cdots \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \cdots \\ 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

그림 2: *1차 이웃한 가중치행렬*

3.3. STAR모형에서의 결측치 추정방법

(1) 최우추정방법에 의한 결측값 대체

식 (3.1)에서 $p = 1$(1차 자기회귀), $m = 1(W^{(1)})$: 1차 이웃한 가중치 행렬)인 경우의 공간시계열 모형을 가정하자. 즉, STAR(1_1)모형을 가정하자.

$$\mathbf{z}(t) = \phi_0 I\mathbf{z}(t-1) + \phi_1 \mathbf{W}^{(1)}\mathbf{z}(t-1) + \mathbf{e}(t), \tag{3.2}$$

여기서, $I = W^{(0)}$, $W^{(1)} = w_{ij}$로 그림 2와 같다고 가정한다.

식 (3.2)는 특정 공간(i)에 대하여 다음과 같이 표현될 수 있다.

$$\mathbf{z}_i(t) = \phi_0 \mathbf{z}_i(t-1) + \phi_1 \sum_{j=1}^{n} w_{ij} z_j(t-1) + \mathbf{e}_i(t). \tag{3.3}$$

식 (3.2)는 아래의 다변량 ARMA모형의 특별한 경우로써 표현될 수 있다.

$$\mathbf{z}(t) = \Phi\mathbf{z}(t-1) + \mathbf{e}(t), \tag{3.4}$$

여기서, $\Phi = \phi_0 I + \phi_1 W^{(1)}$이고, $e(t)$는 평균 0, 분산 σ^2을 따르는 정규분포로 가정한다. 계산상의 편리를 위해 ϕ_0, ϕ_1, σ^2는 알려져 있다고 가정하자. $Z(t) = [z_1(t), z_2(t), \ldots, z_n(t)]^T$ $t = 1, \ldots, T$인 $(n \times 1)$벡터이고, 특정 s-공간의 특정 m-시점 결측값을 $z_s(m)$이라 가정하면 $(1 \le s \le n, 1 \le m \le T)$에 대하여, $Z_s((m))$는 $Z_s(t) = [z_s(1), \ldots, z_s(T)]^T$, $s = 1, \ldots, n$에서 $z_s(m)$을 제외하고 얻어진 $(T-1) \times 1$벡터로 정의된다. 그리고 Z는 $n \times T$ 행렬로 정의되고, Z^*는 Z에서 $z_s(m)$을 제외하고 얻어진 행렬로 정의된다. $z_s(m)$이 주어진 이용 가능한 자료 Z^*의 결합밀도함수는 다음과 같다.

$$f(Z^*|z_s(m)) = \frac{f(Z)}{f(z_s(m))}, \tag{3.5}$$

여기서, 결합밀도함수 $f(Z)$는 다음과 같다.

$$f(Z) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{nT}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum\sum\left\{z_i(t) - \phi_0 z_i(t-1) - \phi_1 \sum_{j=1}^{n} w_{ij} z_j(t-1)\right\}^2\right\}. \tag{3.6}$$

식 (3.5)에서 $z_s(m)$은 Z^*에 대한 미지의 모수로 취급할 수 있다. 그러므로 $z_s(m)$의 가능도함수는 다음과 같다.

$$l(z_s(m)|Z^*) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{nT-1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\sum_{t=1}^{m-1}[A_i(t)]^2 + \sum_{t=m+1}^{T}[B_i(t)]^2\right\}\right\}, \tag{3.7}$$

여기서, $A_i(t) = z_i(t) - \phi_0 z_i(t+1) - \phi_1 \sum_{j=1}^{n} w_{ij} z_j(t+1)$, $B_i(t) = z_i(t) - \phi_0 z_i(t-1) - \phi_1 \sum_{j=1}^{n} w_{ij} z_j(t-1)$이다.

식 (3.7)의 지수 내부 항에서 $i = s$ 공간에 대한 Q로 정의하면 다음과 같이 나타낼 수 있다.

$$Q = \sum_{t=1}^{m-1} \left[z_s(t) - \phi_0 z_s(t+1) - \phi_1 \sum_{j=1}^{n} w_{sj} z_j(t+1) \right]^2 + \sum_{t=m+1}^{T} \left[z_s(t) - \phi_0 z_s(t-1) - \phi_1 \sum_{j=1}^{n} w_{sj} z_j(t-1) \right]^2$$

그러므로 $z_s(m)$의 최대가능도추정량은 다음과 같다.

$$\hat{z}_s(m) = \begin{cases} \dfrac{z_s(m+\delta)}{\phi_0} - \dfrac{\phi_1}{\phi_0} \sum\limits_{\substack{j=1 \\ j \neq s}}^{n} w_{sj} z_j(m), & m = 1 \text{ or } T, \\ \dfrac{z_s(m-1) + z_s(m+1)}{2\phi_0} - \dfrac{\phi_1}{\phi_0} \sum\limits_{\substack{j=1 \\ j \neq s}}^{n} w_{sj} z_j(m), & 1 < m < T, \end{cases}$$

여기서, $m = 1$일 때 $\delta = +1$이고, $m = T$일 때 $\delta = -1$이고, w_{ij}는 s-공간과 이웃한 j-공간의 가중치를 나타낸다.

(2) 확률변수방법에 의한 결측값 대체

앞 절에서 $z_s(m)$은 Z^*에 대한 미지의 모수로 취급하였는데, 확률변수방법에서는 식 (3.6)을 따르는 확률변수 $z_s(m)$을 고려하여, Z^*가 주어진 $z_s(m)$의 분포는 다음과 같다.

$$f(z_s(m)|Z^*) = \frac{f(Z)}{f(Z^*)}, \tag{3.8}$$

여기서, Z^*는 $f(Z)$로부터 $Z_s(m)$을 제거한 적분 값으로 얻어지며, $f(Z)$는 $f(z_s(m)|Z^*)$와 비례하기 때문에 $z_s(m)$의 조건부 기댓값 $E(z_s(m)|Z^*)$은 2.2절에서 설명하였듯이 같은 방법으로 식 (3.6)의 결합밀도함수 $f(Z)$을 최대로 하는 최소제곱추정량이나 최대가능도추정량을 이용하여 쉽게 얻어질 수 있다. 이는 베이지안 관점에서 생각하면 사후분포 $f(z_s(m)|Z^*)$의 평균으로 다음과 같다.

$$\hat{z}_s(m) = \begin{cases} \phi_0 z_s(m+\delta) + \phi_1 \sum\limits_{j=1}^{n} w_{sj} z_j(m+\delta), & m = 1 \text{ or } T, \\ \dfrac{\phi_0 [z_s(m-1) + z_s(m+1)] + \phi_1 \sum_{j=1}^{n} w_{sj} z_j(m-1) - \phi_0 \phi_1 \sum_{j \neq s}^{n} w_{sj} z_j(m)}{(1+\phi_0^2)}, & 1 < m < T, \end{cases}$$

여기서, $m = 1$일 때 $\delta = +1$이고, $m = T$일 때 $\delta = -1$이고, w_{sj}는 s-공간과 이웃한 j-공간의 가중치를 나타낸다.

4. 사례연구

4.1. 자료설명

유행성이하선염(이하 Mumps)은 홍역, B형 간염 등과 더불어 우리나라에서 제 2군 전염병으로 분류하고 있으며, 직접적인 비말(droplet) 또는 다른 오염된 물질이 코나 입으로 들어가서 감염된다. Mumps는 1998년 4,461명의 환자가 발생한 이 후 점차 감소하는 추이를 보이지만 여전히 전국에 걸쳐 높은 발생률을 보이는 전염병으로 알려져 있다 Mumps는 뇌수막염, 고환염, 부고환염, 난소염 췌장염 등의 합병증을 동원하고 일단 감염되면 특이요법이 없어 예방을 최우선으로 한다. 이에 예방을 위해 우

표 1: 모형식별 및 모수추정결과

지역	모형	모수추정	AIC
경기	ARMA(1,0)	$\phi = 0.6898$	189.90
강원	ARMA(1,0)	$\phi = 0.4197$	220.31
충남	ARMA(1,0)	$\phi = 0.4249$	221.52
충북	ARMA(1,0)	$\phi = 0.5343$	208.24
전북	ARMA(1,1)	$\phi = 0.8312$, $\theta = 0.4738$	212.14
전남	ARMA(1,1)	$\phi = 0.9298$, $\theta = 0.6711$	210.74
경북	ARMA(1,0)	$\phi = 0.8625$	142.51
경남	ARMA(1,1)	$\phi = -0.6658$, $\theta = -0.9547$	227.92

표 2: 결측값 추정결과

지역	관측값(2008년 5월자 료)	최우추정방법(MLE)	확률변수방법(RV)
경기	359	451	340
강원	20	31	19
충남	20	67	32
충북	35	18	22
전북	8	0	1
전남	9	6	10
경북	148	105	137
경남	52	38	47
전체	SSE	13201	871

리나라에서는 생후 12~15개월과 4~6세에 MMR 백신 (홍역-유행성이하선염(볼거리) -풍진 혼합백신) 접종을 추천하고 있다.

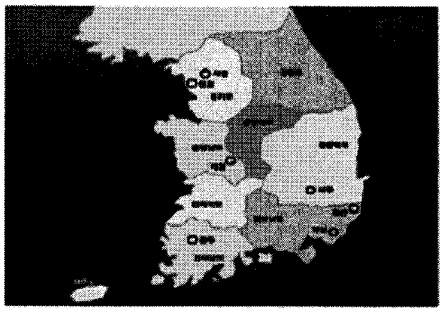
본 논문에서 사용된 Mumps는 한국질병관리본부(Korea Center for Disease Control and Prevention)에서 2001년 1월부터 2009년 11월까지 전염성 감시체계로 전산 보고된 16개 시도의 월별 Mumps 자료이다. 여기서 한국의 Mumps 자료는 8개도로 재분류하였다. 사용된 Mumps 자료는 연속성 자료가 아닌 빈도 자료로 Poisson분포를 따르는 변수이므로, 원래의 자료를 분산안정화변환(여기서, $\sqrt{Z}$)과 12차 계절차분을 취한 후 정상시계열이 되었고, 이 자료를 정규분포로 근사시키기 위해 각 도별로 평균과 표준편차를 구해 자료를 표준화 하였다. 여기서 2008년 5월 자료가 결측값이고 2001년 1월~2008년 12월 자료까지의 월별 자료만을 이용하여 모형을 추정, 결측값을 대체했고 나머지 자료(2009년 1월~2009년 11월)는 예측력을 비교하는데 사용했다.

4.2. 모수와 결측치 추정

(1) 자기회귀이동평균(ARMA)모형

ARMA모형의 식별은 자기상관함수(Autocorrelation Function; ACF)와 편차자기상관함수(Partial Autocorrelation Function; PACF) 등의 감소패턴을 보면서 차수를 식별하였고 AIC, BIC 등을 통해 모형을 선택하였다. 이에 8개도의 식별된 모형과 모수추정 결과는 표 1과 같다. 그리고 결측값인 (2008년 5월) 자료에 대해 위에서 선택된 모형에 따라 최대가능도방법과 확률변수방법으로 결측값을 추정한 결과가 표 2와 같다.

8개도의 결측값 추정결과 모든 지역에서 최대가능도방법에 비해 확률변수방법이 관측 값에 매우 가깝게 추정이 된 것을 확인 할 수 있었고, 이를 전체 지역에 대한 오차제곱합(SSE)을 구해 살펴본 결과



$$W^{(1)} = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & \frac{1}{5} & \frac{1}{5} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

그림 3: *8개도의 Mumps 대상지역과 1-차 가중치행렬*

표 3: 각 도에서 첫 번째 이웃하고 있는 도의 집합

Site	Province	Neighbor Province
1	KG	KW, CB, CB
2	KW	KG, CB, KB
3	CN	KG, CB, JB
4	CB	KG, KW, CB, JB, KB
5	JB	CN, CB, KB, KN, JN
6	JN	JB, KN
7	KB	KW, CB, JB, KN
8	KN	KB, JB, JN

확률변수방법의 결측값에 대한 추정의 정도가 매우 좋게 나타났다.

$$\text{SSE} = \sum_{i=1}^{8} [\text{observed value} - \text{estimated value}]^2.$$

(2) 공간시계열자기회귀(STAR)모형

공간시계열모형의 식별은 Pfeifer와 Deutsch (1980), 손건태와 백지선 (1997)에 의해 연구되었고 이를 통해 STAR모형의 식별은 자기상관함수(Space-Time Autocorrelation Function; STACF)와 공간시계열 편자기상관함수(Space-Time Partial Autocorrelation Function; STPACF) 등의 감소패턴을 보면서 차수를 식별하였고 AIC, BIC 등을 통해 모형을 선택한 결과 다음과 같은 STAR(1_1)모형으로 식별되었다.

$$\mathbf{z}(t) = \phi_0 I\mathbf{z}(t-1) + \phi_1 \mathbf{W}^{(1)}\mathbf{z}(t-1) + \mathbf{e}(t),$$

여기서, $I = W^{(0)}$, $W^{(1)} = w_{ij}$이다.

본 논문의 관심영역과 가중치행렬(weight matrix)은 그림 3과 같고 표 3은 각 도의 첫 번째 이웃하고 있는 집합을 나타낸다.

공간시계열모형의 모수추정과 예측은 Alcroft와 Glasbey (2005), Billard와 Dai (1998, 2003) 등에 의해 연구되었고, 공간시계열모형에 대하여 Billard와 Dai가 제안한 Newton-Raphson의 반복최대가능도추정(IMLE)과 Kalman-Filter방법 중 Kalman-Filter방법을 이용하여 모수를 추정한 결과와 추정된 모수를 이용해 최대가능도방법과 확률변수방법을 이용하여 결측치에 대해 추정한 결과가 표 4와 같다.

8개도의 각각의 결측값 추정결과 전남을 제외한 모든 지역에서 최대가능도추정법에 비해 확률변수방법이 관측 값에 매우 가깝게 추정이 된 것을 확인 할 수 있다. 이를 전체 지역에 대한 오차제곱

표 4: 각 지역별로 2008년 5월에 결측치가 발생한 경우에 추정결과

지역(결측치)	모수추정	관측값	MLE	RV
경기(2008.5)	$\phi_0 = 0.5019,\ \phi_1 = 0.0074$	359	584	315
강원(2008.5)	$\phi_0 = 0.5089,\ \phi_1 = 0.0113$	20	27	19
충남(2008.5)	$\phi_0 = 0.5098,\ \phi_1 = 0.0099$	20	57	33
충북(2008.5)	$\phi_0 = 0.5108,\ \phi_1 = 0.0090$	35	17	22
전북(2008.5)	$\phi_0 = 0.5117,\ \phi_1 = 0.0106$	8	0	1
전남(2008.5)	$\phi_0 = 0.5088,\ \phi_1 = 0.0084$	9	11	19
경북(2008.5)	$\phi_0 = 0.5047,\ \phi_1 = 0.0113$	148	11	175
경남(2008.5)	$\phi_0 = 0.5114,\ \phi_1 = 0.0087$	52	109	85
전체		SSE	74453	4242

표 5: 8개도 월별 Mumps 자료에 대한 예측오차제곱합의 비교

지역	ARMA-MLE (SSF)	ARMA-RV (SSF)	STAR-MLE (SSF)	STAR-RV (SSF)
경기	462361	436996	525793	512408
강원	3505	3910	3557	3913
충남	6666	3740	5464	3675
충북	352	312	351	418
전북	361	346	382	360
전남	7118	6998	5515	5241
경북	12745	11252	12718	23985
경남	3490	3110	6290	3995

합(SSE)을 구해 살펴본 결과 확률변수방법의 결측값에 대한 추정의 정도가 매우 좋게 나타났다. 미래 값 예측을 위해 최대가능도추정방법과 확률변수방법에 의한 추정된 결측치를 대체한 후 Kalman-Filter를 이용한 공간시계열모형의 모수추정결과는 다음과 같다.

$$\text{최대가능도추정방법(MLE)}: \ \phi_0 = 0.5193,\ \phi_1 = 0.0072,\ \sigma^2 = 0.7279,$$
$$\text{확률변수방법(RV)}: \ \phi_0 = 0.5171,\ \phi_1 = 0.0052,\ \sigma^2 = 0.7306.$$

5. 예측결과

본 논문에 사용된 Mumps 자료는 추정에 2001년 1월~2008년 12월 자료까지의 월별자료를 활용하였고 나머지 자료(2009년 1월~2009년 11월)는 예측력을 파악하는데 사용하였다. 최대가능도추정방법과 확률변수방법에 의해 결측값을 대체한 후의 모형의 예측력 비교는 아래의 예측오차 제곱합(Sum of square for forecasting error; SSF)을 사용하였고 8개도의 Mumps 월별자료에 대한 관측값과 예측값의 비교결과는 표 5와 같다.

$$\text{SSF} = \sum_{i=1}^{11} [\text{observed value} - \text{predicted value}]^2.$$

예측오차 제곱합(SSF)를 비교해 보면 자기회귀이동평균(ARMA)모형의 경우 강원을 제외한 나머지 지역에 대해 확률변수방법이 최대가능도방법에 비해 더 좋은 예측력을 보여주고 있으며, 공간시계열자기회귀(STAR)모형의 경우 강원, 충북, 경북을 제외한 나머지 지역에 대해 확률변수방법이 최대가능도방법에 비해 더 좋은 예측력을 보임을 알 수 있다.

6. 결론 및 향후과제

본 논문에서는 시계열자료에서 존재하는 결측값을 대체하는 방법으로 최대가능도방법과 베이지안 방법의 확률변수방법을 이용해 8개도의 Mumps 자료에 대해 각각 1개의 결측값을 만들고 두 가지 추정방법에 따른 결측값을 추정하여 그 추정의 정도를 비교한 결과 ARMA모형의 경우 모든 지역에서 확률변수방법이 추정의 정도가 더 좋았으며, STAR모형의 경우역시 1개 지역을 제외하고는 확률변수방법이 추정의 정도가 매우 좋았다. 그리고 두 가지 방법에 의해 추정된 결측값을 각각 대체한 후 추가적으로 미래 11개월에 대한 예측력을 예측오차제곱합(SSF)을 구하여 비교한 결과 ARMA모형과 STAR모형에서 확률변수방법이 대체로 좋은 예측력을 보임을 알 수 있었다. 본 논문에서는 기존의 선형시계열 모형인 ARMA모형과 공간과 시간을 동시에 고려한 STAR모형에 대해서 결측치 대체방법을 비교 연구하였고, 추후 연구계획으로 비선형시계열 모형인 중선형모형(Bilinear Model)과 시간과 공간을 동시에 고려한 공간시계열 중선형모형(STBL; Space-Time Bilinear Model)으로 확장하여 보다 포괄적이면서 복잡한 모형에서의 결측치의 대체방법에 대해 연구할 계획이다.

참고 문헌

손건태, 백지선 (1997). 공간자기회귀모형의 식별, <응용통계학회>, **10**, 121–136.

이성덕, 김덕기 (2009). 시계열자료에서 결측치 추정방법의 비교, <한국통계학회>, **16**, 723–730.

Alcroft, D. J. and Glasbey, C. A. (2005). STARMA process applied to Solar Radiation, *Biomathmatics and Statistics Scotland,* 1–24.

Bayarri, M. J., DeGroot, M. H. and Kadane, J. B. (1986). What is the likelihood function?, In: *Statistical Decision Theory and Related Topics IV, Volume 1.,* Springer-Verlag, New York.

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

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

Brubacher, S. R. and Wilson, T. (1976). Interpolating time series with application to the estimation of holiday effects on electricity demand, *Applied statistics,* **25**, 107–116.

Dunsmuir, W. and Robinson, P. M. (1981). Estimation of time series models in the presence of missing data, *Journal of the American Statistical Association,* **76**, 560–568.

Horton, N. J. and Kleinman, K. P. (2007). A comparison of missing data methods and software to fit incomplete data regression models, *The American Statistician,* **61**, 79–90.

Kim, J. (2005). Parameter estimation in stochastic volatility models with missing data using particle methods and the EM algorithm, *PhD thesis, University of pittsburgh.*

Pena, D. (1987). Measuring the importance of outliers in ARIMA models, In *New Perspectives in Theoretical and Applied Statistics,* M. L. Puri (ed.) Wiley, New York, 109–118.

Pena, D. and Tiao, G. C. (1991). A note on likelihood estimation of missing values in time series, *The American Statistician,* **45**, 212–213.

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

Rubin, D. B. (1994). Missing data, imputation, and the bootstrap: Comment, *Journal of the American Statistical Association,* **89**, 475–478.

2010년 1월 접수; 2010년 3월 채택

The Comparison of Imputation Methods in Space Time Series Data with Missing Values

SungDuck Lee[1,a], DuckKi Kim[a]

[a]Department of Information and Statistics, Chungbuk National University

Abstract

Missing values in time series can be treated as unknown parameters and estimated by maximum likelihood or as random variables and predicted by the conditional expectation of the unknown values given the data. The purpose of this study is to impute missing values which are regarded as the maximum likelihood estimator and random variable in incomplete data and to compare with two methods using ARMA and STAR model. For illustration, the Mumps data reported from the national capital region monthly over the years 2001~2009 are used, and estimate precision of missing values and forecast precision of future data are compared with two methods.

Keywords: MLE, RV, ARMA, STAR, Mumps data, SSF, STACF, STPACF, Kalman-Filter.

This Paper was supported by the Chungbuk National University in 2008.

[1] Corresponding author: Professor, Department of Information and Statistics, Chungbuk National University, Chungbuk 361-763, Korea. E-mail: sdlee@cbnu.ac.kr