시공간 베이지안 계층모형-미국 연기온 편차자료에 적용-

이의규[1], 문명상[2], Richard F. Gunst[3]

요 약

전형적인 시공간모형은 시공간 변이도(semivariogram) 또는 공분산 함수(covariance function)를 필요로 한다. 본 논문에서는 계산하기 어렵고 현실적이지 못한 결합 공분산 함수를 통한 고전적 모형 대신, 일련의 독립적인 조건분포를 이용하는 보다 현실적인 베이지안 계층모형을 이용한다. 미국 전 지역에 산재해 있는 138개 기온 관측소로부터 얻어진 61년(1920-1980) 동안의 연기온편차 자료에 시공간 베이지안 계층모형을 적용하고 순수시계열모형에서의 적합값과 제안된 모형의 적합값을 비교분석한다.

주요용어 : ARIMA models, Gibbs sampling, Hierarchical model, Semivariogram.

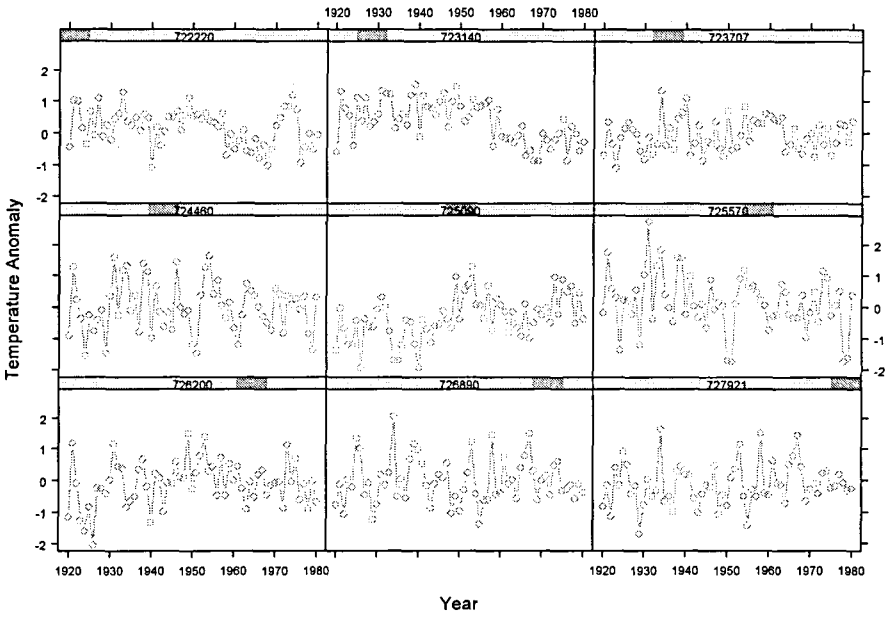
1. 서론

여러 지점에서 일정 기간동안 관측하여 얻어진 자료를 시공간자료(space-time data)라 한다. 이러한 시공간자료를 어떻게 모형화하고 어떻게 분석할 것인지에 대한 관심이 더해가고 있다. 실제로 시공간자료에 대한 모형분석은 환경, 기상, 지질, 역학(疫學), 임업, 수자원, 어업등 많은 분야에서 점차 사용범위가 증대되고 있다. 한편 전형적인 시공간모형은 시공간 변이도(semivariogram) 또는 공분산 함수(covariance function)를 필요로 한다. 본 논문에서는 계산하기 어렵고 현실적이지 못한 결합 공분산 함수를 통한 고전적 모형 대신, 일련의 독립적인 조건분포를 이용하는 보다 현실적인 베이지안 계층모형(Wikle(1998), 안윤기 외 다수(2001), 최일수 외 다수(2001), 이승천과 이덕환(2002), 오만숙과 박현진(2002))을 이용하려 한다. 즉, 시공간자료의 계층모형(hierarchical model)을 설정하고 베이지안(Bayesian) 분석틀에서 자료의 상관성을 효과적으로 이용하여 분석하고자 한다.

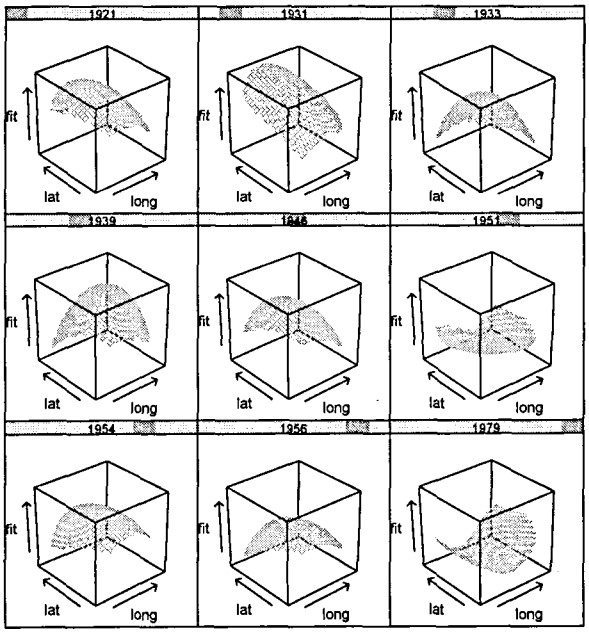
본 논문에서는 미국 전 지역에 산재해 있는 138개 기온 관측소로부터 얻어진 61년(1920-1980) 동안의 연기온편차(관측지점에서 일정기간동안 관측된 평균기온과의 기온차) 자료에 시공간 베이지안 계층모형을 적용한다. 일반적으로 시계열자료는 시간적으로 순서화 되어 있고, 따라서 각 시점에서의 시계열추세와 관찰값간의 자기상관관계를 갖게 되는 것이 일반적이다. 반면에 공간자료에는 공간추세와 서로 가까운 관측지점에서는 유사한 관찰값이 예상되는 공간적 상관관계가 존재할 수 있다. 즉, 시공간자료에는 시간적 성분과 공간적 성분이 동시에 포함되어 있을 수 있기 때문에 어떻게 시간과 공간성분을 동시에 모형화시킬 것인가가 관건이라 하겠다.

[그림 1.1]는 각 관측소(138개의 관측소 중 9개의 관측소를 선택)의 시계열을 나타낸 것인데 각 관측지점에서 자기상관이 존재하는 것으로 생각된다. [그림 1.2]은 각 연도(61개의 연도중 9개의 연도를 선택)에서 위도와 경도에 대한 이차 회귀적합값들을 3차원으로 도식한 것이다. [그림 1.2]에서 볼 수 있듯이 공간성분은 각 연도마다 다른 형태로서 적합된다. Cressie(1993)는 공

1) 건국대학교 상경대학 응용통계학과 강의교수
2) 연세대학교 자연과학부 정보통계학과 부교수
3) Professor, Department of Statistical Science, Southern Methodist University, USA



[그림 1.1] 특정 9개 관측소에서의 기온편차의 시계열



[그림 1.2] 특정 9개 년도에서의 기온편차의 추정된 공간추세함수

간적 추세를 제거한 잔차를 이용하여 각 변이도(semivariogram)들을 추정한 후 이를 이용하여 관측되지 않은 지점에서의 값을 예측하는데, 이는 시계열에서 추세를 제거한 후 상관도(correlogram)를 추정하는 것과 같은 맥락으로 볼 수 있다(실제로 추정된 61개의 변이도를 합성하여 구하여지는 종합 변이도(ensemble semivariogram)를 본 논문에서 이용하는데 이는 후에 다시 언급하기로 한다). 그러므로 각 지점, 또는 각 시점마다 동일한 추세나 변동을 갖지 않는다는 것은 이러한 시공간자료에 대해 단순한 형태로 모형화하는 것이 부적합함을 암시한다. Hartfield와 Gunst(1999)는 시공간적으로 상관이 있는 자료에 대하여 여러 모형족을 일반화하고 이러한 모형들에서 시간과 공간성분을 식별하는 방법을 제안한다. 그들은 이 기온편차자료에 대해 위도와 경도에 대한 이차함수형태의 추세와 공간적으로 상관된 오차의 AR(1) 시계열 구조로 결론지었다. 실제로 138개의 시계열분석에서 Akaike 정보판단기준(Akaike's information criterion)에 근거하면 96개가 백색잡음과정(white noise process)이거나 AR(1)으로 식별된다.

위에서 언급한 바와 같이, 각 기후관측소에서 나타나는 상이한 경향들은 고전적 ARIMA 모델(Box, G., Jenkins, G. M. and Reinsel, G. C., 1994)을 통하여 기온편차의 변동을 종합적으로 설명하기 어렵다는 것을 말해준다. 따라서 본 논문에서는 연 기온 편차 자료에 공간적인 구성요소와 전시점에서 관측된 기온변화의 영향을 고려한 자기회귀 구성요소를 포함하는 시공간 모형을 적용하여 순수한 시계열 모형을 적용했을 때보다 더 좋은 적합 결과를 도출해보고자 한다. 다음의 2장에서는 베이지안 계층모형을 설정하며, 3장에서는 순수한 시계열모형과 베이지안 계층모형에서의 적합값을 비교하고 4장에서 결론을 맺는다.

2. 베이지안 계층 시공간모형

베이지안 계층모형의 설정과 적합은 독립적인 조건분포의 사용을 통하여 상관된 자료의 효과적인 분석을 가능하게 한다. 즉, 독립적인 조건분포를 이용함으로서 전통적인 모형적합의 계산상 어려움을 줄일 수 있게 되어 많은 분야에서 빈번하게 사용되고 있다. 깁스샘플링은 완전조건분포로부터 생성된 깁스샘플의 평균이 사후분포의 평균으로 접근한다는 사실에 기초한다. 이는 다변량의 샘플을 생성하는 대신 완전조건분포를 이용하여 일변량 샘플을 생성함으로써 분석을 단순화하는 것이다. 본 논문에서는 관심모수의 완전조건분포를 유도하고 이를 바탕으로 한 프로그래밍(C와 IMSL)을 통하여 분석하였다. 그러면 이제 모형을 설정하기로 하자.

2.1 모형설정

1장에서 제시된 자료의 탐색적 분석기법에 의한 결과를 적용하여 연 기온편차 자료에 다음과 같이 주어지는 시공간 모형을 설정한다.

$$Z(s_i,t)=\mu(s_i,t)+W(s_i,t)+e(s_i,t),\quad i=1,2,\cdots,138,\quad t=1,2,\cdots,61. \tag{2.1}$$

여기서 $\mu(s_i,t)$는 시공간추세함수이고 $W(s_i,t)$는 추세함수주변의 변동을 설명하는 시공간 확률과정이며 $e(s_i,t)$는 오차항이다. 이를 벡터로 나타내면

$$Z_t=\mu_t+W_t+e_t,\quad t=1,2,\cdots,61. \tag{2.2}$$

와 같이 표현할 수 있다. 여기서 각 항은 138×1 열벡터이다. 그리고 위 모형은 다음 절차에 의해 계층모형으로 나타낼 수 있다.

(1) $Z(s,t)$는 조건부 독립변수이며 정규분포를 가정한다. 즉,

$$Z_t \mid \mu_t,\ W_t,\ \sigma_e^2 \sim N(Z_{0t},\sigma_e^2 I),\qquad Z_{0t}=\mu_t+W_t \tag{2.3}$$

(2) 조건부 독립적인 공간구조와 시간구조는 각각 정규분포를 가정한다. 즉,

$$\boldsymbol{\mu}_t \mid \boldsymbol{\mu}_{0t}, \sigma_\tau^2 \sim N(\boldsymbol{\mu}_{0t}, \sigma_\tau^2 \boldsymbol{I}), \quad \boldsymbol{\mu}_{0t} = \boldsymbol{P}\,\boldsymbol{\alpha}_t \tag{2.4}$$

$$\boldsymbol{W}_t \mid \boldsymbol{W}_{t-1}, \boldsymbol{H}, \sigma_\eta^2 \sim N(\boldsymbol{\omega}_{0t}, \sigma_\eta^2 \boldsymbol{I}), \quad \boldsymbol{\omega}_{0t} = \boldsymbol{H}\,\boldsymbol{W}_{t-1}. \tag{2.5}$$

여기서 $\boldsymbol{P}$ 는 138×6 계획행렬(design matrix)이고 $\boldsymbol{\alpha}_t$ 는 각 연도 t에 대한 공간다항추세식의 계수행렬이며 $\boldsymbol{H} = Diag(\boldsymbol{\alpha})$는 각 지점 s에 대한 일차 자기회귀계수 대각행렬이다. 즉,

$$\mu_0(s,t) = \alpha_{1(t)} + \alpha_{2(t)}\text{위도}(s) + \alpha_{3(t)}\text{경도}(s) + \alpha_{4(t)}\text{위도}(s)^2 + \alpha_{5(t)}\text{경도}(s)^2 + \alpha_{6(t)}\text{위도}(s)\text{경도}(s)$$

로서 각 연도에서 이차함수식으로 표현되고 $\omega_0(s,t) = a(s)W(s,t-1)$로 1차자기회귀식이다. 식 (2.4)-(2.5) 은 보다 간단한 모형으로 표현될 수 있는데, 예를 들면 시간에 의존하지 않고 공통 공간결정함수를 갖는 모형 ($\boldsymbol{\alpha}_t \rightarrow \boldsymbol{\alpha}$)이거나 지점에 의존하지않고 공통된 1차 자기상관계수를 갖는 모형 ($\boldsymbol{\alpha} \rightarrow a$)으로 축소할 수 있다.

(3) 사전분포의 설정

사전분포는 과거의 분석이나 조사자의 경험에 의해 제시되어야 하나 이 자료에서는 모형에 근거한 사전분포를 사용하였다. 회귀계수들에 대한 분포는 정규분포 그리고 이에 대응되는 분산은 역감마분포로 가정할 수 있다. 일반적으로 위치모수(location parameter)는 정규분포를, 척도모수(scale parameter)는 역감마분포(inverse gamma distribution)를 가정한다. 요약하면 다음과 같은 사전분포를 가정할 수 있으며 각 사전분포의 모수는 생략한다.

$$\begin{aligned}
&\boldsymbol{\alpha} \sim N(\tilde{a}_0 \mathbf{1}, \tilde{\sigma}_{a_0} \boldsymbol{I}) && \sigma_e^2 \sim IG(\tilde{q}_e, \tilde{r}_e) \\
&\sigma_\eta^2 \sim IG(\tilde{q}_\eta, \tilde{r}_\eta) && \sigma_\tau^2 \sim IG(\tilde{q}_\tau, \tilde{r}_\tau) \\
&\boldsymbol{\alpha}_t \sim N(\tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\Sigma}}_\alpha) && \boldsymbol{W}_0 \sim N(\tilde{\boldsymbol{\mu}}_{W_0}, \tilde{\boldsymbol{\Sigma}}_{W_0}).
\end{aligned} \tag{2.6}$$

2.2 완전조건분포(Full conditional distribution)와 초기값

깁스샘플링 절차는 각 모수에 대한 완전조건분포로부터의 연차적 샘플생성에 기초한다. 한편 조건부 독립성 가정에 의해 본 논문에서 제시된 모형에서 결합분포는 다음과 같다.

$$\begin{aligned}
P(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{W}_0, \sigma_e^2, \sigma_\eta^2, \sigma_\tau^2, \boldsymbol{\alpha}) = &\Big[\prod_{t=1}^{T}\{P(\boldsymbol{Z}_t \mid \boldsymbol{\mu}_t, \boldsymbol{W}_t, \sigma_e^2) P(\boldsymbol{\mu}_t \mid \boldsymbol{\mu}_{0t}, \sigma_\tau^2) \\
&P(\boldsymbol{W}_t \mid \boldsymbol{W}_{t-1}, \boldsymbol{H}, \sigma_\eta^2) P(\boldsymbol{\alpha}_t)\}\Big] \\
&\times P(\boldsymbol{W}_0) P(\sigma_e^2) P(\sigma_\eta^2) P(\sigma_\tau^2) P(\boldsymbol{\alpha})
\end{aligned}$$

여기서 $\boldsymbol{H} = Diag(\boldsymbol{\alpha})$, $\boldsymbol{\mu}_{0t} = \boldsymbol{P}\,\boldsymbol{\alpha}_t$ 이다. 모수에 대한 완전조건분포는 우도함수와 사전분포의 곱에 비례하므로 결합밀도함수로부터 완전조건분포를 유도할 수 있다.

깁스샘플링은 모든 모수에 대해 초기값을 필요로 한다. 일단 초기값이 결정되면 한 모수의 완전조건분포로부터 깁스샘플이 생성되고 생성된 깁스샘플은 다음 모수의 완전조건분포에 최신화되어 깁스샘플을 연차적으로 생성한다. 반복횟수가 적당히 크게 되면 샘플들은 초기값과는 독립적으로 수렴하나 초기값의 선택은 수렴속도에 영향을 주게 된다. 이 모형에서의 초기값은 역시 모형에 근거하여 얻어진 사전분포의 평균을 사용하였다.

3. 모형의 적합값 비교

이 장에서 우리는 제안된 모형의 모수를 추정하고 시공간모형과 개별적인 시계열모형의 적합값을 비교한다. 특히 하나의 관측소에서의 각각의 적합값을 그래프를 통하여 비교한다. 깁스샘

플은 2가지의 모형으로부터 생성되었다. 그 중 하나는 연도에 상관없이 공통된 공간추세와 1차 자기회귀를 갖는 시계열모형(모형 I), 또 하나는 연도에 따라 다른 공간추세와 1차 자기회귀를 갖는 모형(모형 II) 두가지를 고려했다. 두 모형에서 자기회귀계수는 지점에 상관없이 공통된 것으로 가정하였다.

3.1 베이지안 모형의 모수추정

분석의 정확도를 위해 각 모형에 대해 깁스샘플 4,000개를 생성하였고 첫 2,000개는 준비기(burn-in period)로 간주하고 나머지 2,000개의 샘플을 사용하여 사후평균을 구했다. 즉, $\hat{\theta}_k^{(i)}$를 θ_k의 i 번째 샘플이라 할 때 깁스 추정량은 다음과 같다.

$$\hat{E}(\hat{\theta}_k) = \frac{1}{n} \sum_{i=b+1}^{b+n} \hat{\theta}_k^{(i)}, \qquad b=2,000,\ n=2,000 \tag{3.1}$$

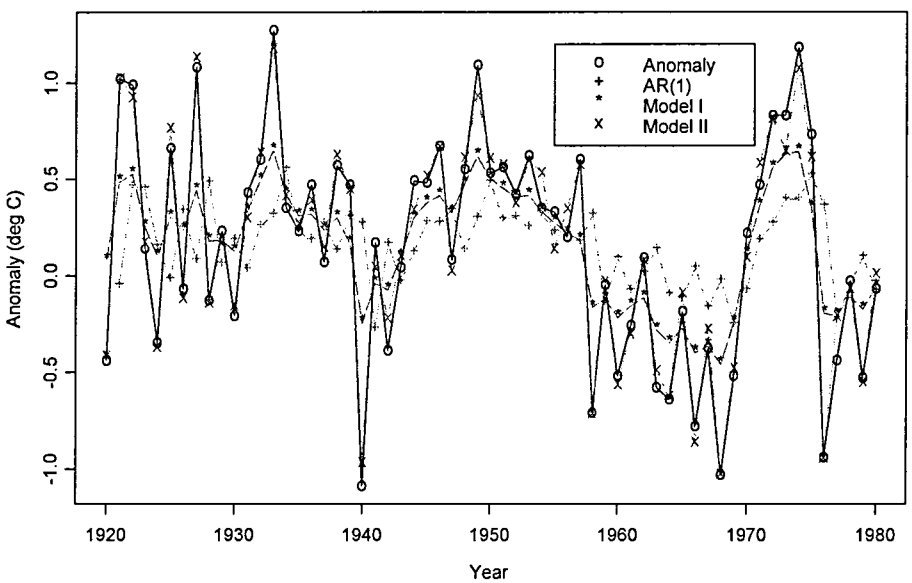
또한 수치적으로 비교하기 위하여 평균절대예측오차(mean absolute prediction errors)를 다음과 같이 정의한다.

$$MAPE = \frac{\sum_{s,t} | z(s,t) - \hat{z}(s,t) |}{n-p-1}, \quad n=8418, \quad p=6 \tag{3.2}$$

그러나 모형 II에서는 (3.2)식에서 분모가 $n-(p\times T)-1$ ($T=61$)로 조정된다.

3.2 베이지안 모형과 순수시계열모형의 적합값 비교

[그림 3.1]는 AIC에 의한 시계열모형의 적합값과 베이지안 계층모형의 적합값을 선택된 하나의 관측지점에서 비교한 것이다.



[그림 3.1] AR(1)모형과 베이지안 계층 시공간모형에서의 적합값 비교

이는 [그림 1.1]에서 나타난 첫 관측지점에서의 기온편차의 시계열인데 AIC를 근거로 AR(1)모형이 선택되었다. 각 시점에서 베이지안 계층모형으로부터 계산된 적합값이 실제 자료에 보다 가깝게 대응하는 것을 알 수 있다. 실제로 모든 지점에서 이와 같은 결과를 쉽게 확인할 수 있다.

수치적인 비교를 위하여 평균절대예측오차(mean absolute prediction error)를 계산하면 ARIMA (p,d,0)에서는 0.58로 나타난 반면 베이지안 계층모형은 0.18 (모형 I에서는 0.44)로 계산되었다. 이러한 차이는 베이지안 계층모형이 주어진 시점에서의 모든 공간자료를 이차함수를 통하여 조정하는 반면에 개별 시계열모형분석은 단지 그 지점에서의 평균으로 조정한 후 시계열의 랜덤성분만을 주로 예측하기 때문이다.

4. 결론

통계모형은 잡음자료(noisy data)로부터 신호(signal)를 효과적으로 잡아내는 것을 목적으로 하는 바, 자료가 시간과 공간상에서 관측되어졌다면 모형은 시간적 신호와 공간적 신호를 동시에 잡아내야 마땅할 것이다. 따라서, 일반적으로 시공간자료에 고전적인 시계열모형과 고전적인 공간모형을 각각 적용하는 것은 바람직하지 않을 것이다. 본 논문에서 제시된 베이지안 계층 시공간모형은 비현실적인 가정없이 공간과 시간의 두 정보를 효과적으로 포함한다.

본 논문에서는 시공간상 이차함수와 일차자기상관을 갖는 베이지안 계층 시공간모형이 순수한 시계열 모형을 압도하는 것을 확인하였다. 모형의 모수를 추정하기위해 우리는 베이지안 분석틀에서 깁스샘플링이라는 도구를 이용하였는데 이는 결측치를 포함한 경우일지라도 그 추정이 용이하다. 그러나 이러한 추정방법은 얼마만큼의 샘플을 얻고 얼마만큼을 버릴지, 초기치와 사전분포를 어떻게 결정할 것인가 등의 주의깊은 고찰이 필요하다. 끝으로 본 논문에서 제시된 모형에 여러가지가 추가된 모형을 고려할 수 있을 것이다. 예를 들면 자기회귀적 영향에 더하여 전시점의 인근지점들에서 관측된 평균값의 영향을 고려할 수도 있을 것이다.

참고문헌

[1] Seung-Chun Lee and Deukhwan Lee(2002), Bayesian Analysis of Multivariate Threshold Animal Models, 「통계학연구」, 31권 2호, 177-198,

[2] Yun Kee Ahn, IlSu Choi, Sung Suk Rhee(2001), A Bayesian Wavlet Threshold Approach for Image Denoising, 「한국통계학회 논문집」, 8권 1호, 109-116

[3] Il Su Choi, Sung Suk Rhee, Yun Kee Ahn (2001), Noise-free Distribution Comparison of Bayesian Wavlet Threshold for Image Denoise, 「한국통계학회 논문집」, 8권 2호, 573-579,

[4] Man-Suk Oh, Hyun-Jin Park (2002), Hierarchical Bayesian Analysis of Smoking and Lung Cancer Data, 「한국통계학회 논문집」, 9권 1호, 115-128,

[5] Box, G., Jenkins, G. M. and Reinsel, G. C.(1994), *Time Series Analysis. Forecasting and Control*, Eaglewood Cliffs: Prentice Hall.

[6] Cressie, N.(1993). *Statistics for Spatial Data*, revised ed. New York: John Wiley & Sons.

[7] Hartfield, M. I. and Gunst, R. F. (1999). Spatiotemporal Modeling of Continuous Space-Time Processes, Technical Report SMU-TR-290, Department of Statistical Science, Southern Methodist University, Dallas, TX.

[8] Wikle, C. K.(1998), Hierarchical Bayesian Space-Time Models, *Environmental Ecological Statistics*, Vol 5, 117-154.