Journal of the Korean Data & Information Science Society
2016, **27**(5), 1155–1168

http://dx.doi.org/10.7465/jkdi.2016.27.5.1155
한국데이터정보과학회지

혼합 조건부 종추출모형을 이용한 여름철 한국지역 극한기온의 위치별 밀도함수 추정[†]

조성일[1] · 이재용[2]

[1]싱가포르 국립대학교 통계 및 응용확률학과 · [2]서울대학교 통계학과

접수 2016년 7월 20일, 수정 2016년 8월 24일, 게재확정 2016년 9월 2일

요 약

기상 자료의 경우 한 지역의 기후가 인접지역의 기후와 비슷한 양상을 띄고 각 지역의 확률 밀도 함수 (probability density function)가 잘 알려진 확률 모형을 따르지 않는다는 것이 알려져 있다. 본 논문에서는 이러한 특성을 고려하여 이상 기후 현상이 뚜렷히 나타나는 여름철 평균 극한 기온 (extreme temperature)의 확률 밀도 함수를 추정하고자 한다. 이를 위하여 공간적 상관관계 (spatial correlation)를 고려하는 비모수 베이지안 (nonparametric Bayesian) 모형인 조건부 자기회귀 종추출 혼합모형 (mixtures of conditional autoregression species sampling model)을 이용하였다. 자료는 이스트앵글리아 대학교 (University of East Anglia)에서 제공하는 전 지구의 최대 기온과 최소 기온자료 중 우리나라에 해당하는 지역의 자료를 사용하였다.

주요용어: 공간적 상관관계, 머서 조건부 자기회귀모형, 밀도함수추정, 여름철 극한 기온, 조건부 자기회귀 종추출 혼합모형.

1. 서론

National Institute of Meteorological Sciences (2009)에 따르면 해마다 거듭되는 이상기후 현상으로 기후 변화에 대한 관심이 전 세계적으로 확대 되고 있으며 2007년 Intergovernmental Panel on Climate Change (IPCC) 4차 평가보고서에서 기후변화로 부터 파생되는 자연재해, 환경, 보건 등에 대한 영향의 과학적 증거들이 제시되면서 국내에서도 관심이 고조 되고 있다.

최근 들어 전 세계적으로 극한기후변화에 의해 자연재해 피해가 빈번히 발생하고 있다. 미국에서는 기상이변으로 폭설이 내려 한 도시의 모든 기능이 정지되고 인도네시아의 수도에서는 집중호우로 인한 홍수 피해로 국가적 엄청난 손실이 생겼다. 특히, 우리나라의 경우 기후변화는 다른 국가에 비해 더 뚜렷히 나타나고 있다. 지난 수십년 동안 우리나라의 평균 기온 상승률은 전 지구 평균기온 상승률에 비해 높게 나타나고 있고 이러한 평균 기온의 변화는 극한기온현상의 큰 변화를 야기하여 (Mearns 등, 1984; Meehl 등, 2000; Griffiths 등, 2005) 여름철 고온에 의한 사망자수가 급격히 증가하고 있으며 (Jo 등, 2012), 전력수요도 큰 폭으로 증가하고 있다. 온실효과의 영향으로 강수량에서도 이상 현상이 뚜렷하게 나타나고 있다. 해가 거듭될 수록 연 강수량은 증가하고 있으나 강수일수는 줄어들고 강수일의 강수 강

[†] 이 논문은 2015년도 정부 (미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2011-0030811).
[1] 교신저자: (119077) 싱가포르 로워 켄트리지 로드 21, 싱가포르 국립대학교 통계 및 응용확률학과, 리서치 펠로우. E-mail: joseongil@gmail.com
[2] (08826) 서울특별시 관악구 관악로 1, 서울대학교 통계학과, 교수.

도가 증가하고 있다. 특히 여름철 집중호우로 인한 피해는 해마다 많은 인명 및 재산피해로 이어져 사회적, 경제적으로 엄청난 국가의 손실이 발생하고 있다. 또한 Im과 Kwon (2007)과 National Institute of Meteorological Sciences (2009)에 따르면 미래의 우리나라 기온은 저온일이 감소하고 고온일이 증가하여 고온으로 인한 극한기온현상이 더 두드러질 가능성이 높다고 예측하고 있어 이에 대한 적절한 예방과 대비가 절실히 필요한 실정이다.

우리나라 기상청에서는 기후 예측에 주로 사용하는 방법 중 하나로 과거의 자료를 이용해 각 기상 변수의 확률밀도 함수를 추정하고 추정된 밀도 함수를 세 개의 범주로 나누어 기온이나 강수량이 평년보다 높은지 같은지 낮은지에 대해 예측값을 제공한다 (Jo 등 (2016)). 이와 마찬가지로 만약 극한 기온의 확률 밀도 함수를 정확하게 추정할 수 있다면 추정된 밀도 함수로 부터 극한기후변화의 특성을 파악할 수 있고 더 나아가 과거의 극한기후현상에 비해 어떻게 바뀌는지에 대한 정확한 예측을 통해 적절한 예방과 대비할 수 있어 국가적인 손실을 줄일 수 있을 것이다. 따라서 본 논문에서는 극한기후변화가 가장 뚜렷하게 나타나는 여름철의 극한 기온에 대한 확률밀도함수 추정을 시행하고자 한다. 추정 모형으로는 Jo 등 (2016)에서 제시된 혼합 조건부 자기회귀 종추출모형 (mixtures of conditional autoregressive species sampling models)을 이용한다. 혼합 조건부 자기회귀 종추출모형은 공간적인 상관관계가 있을 때 정확한 추정이 가능한 모형으로 특히 기상자료의 경우 공간적인 상관관계가 뚜렷하게 나타난다는 것이 알려져 있기 때문에 정확한 추정이 가능할 것이다.

본 논문은 다음과 같이 구성된다. 2절에서는 분석에서 사용되는 자료를 자세히 살펴보고, 3.1절에서는 밀도함수 추정에 사용된 베이지안 모형을 설명하고 각 모수의 사후분포로 부터 사후표본을 추출하는 방법 그리고 사후 예측 밀도함수의 추정량을 3.2절과 3.3절에서 각각 서술하였다. 마지막으로 모형을 적용한 자료분석 결과를 4절에 담았다.

2. 자료의 개요

본 논문에서는 이스트앵글리아 대학교 (University of East Anglia)의 Climate Research Unit (CRU)에 의해 공개된 자료를 사용하였다. 자료는 `http://badc.nerc.ac.uk/data/cru`에서 받을 수 있다. 이 자료는 1901년부터 2009년까지 108년 동안의 기간을 대상으로 전 지구의 위도와 경도를 $0.5^\circ \times 0.5^\circ$로 나누고 (즉, 격자점 사이의 간격이 0.5°로 나누어져 있다.) 각 격자점 위에서 관측된 일별 최고 온도와 최저 온도의 월 평균 값으로 구성되어 있다. 자료에 대한 더 자세한 사항은 Harris 등 (2013)에 설명되어 있다.

본 연구의 분석에서는 우리나라를 포함하는 지역의 여름철 평균 자료만을 사용하였다. 여름철 자료를 구성하기 위해 매년 6월에서 8월까지 3개월치 자료의 산술평균 값을 사용하였으며 우리나라를 포함하는 지역으로 북위 34°에서 북위 39°까지, 동경 126°에서 동경 129.5°까지의 범위를 이용하였다. 총 격자점의 갯수는 70개이다. Figure 2.1은 본 논문에서 분석을 위해 사용한 지역을 나타낸 것이다.

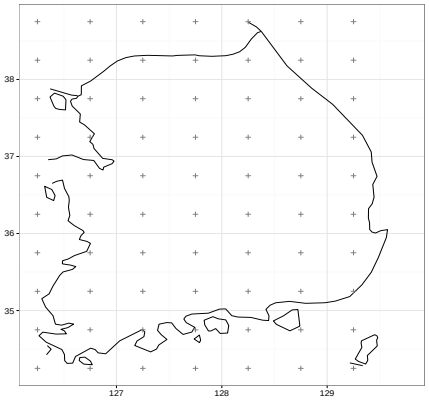

Figure 2.1 Location and grid points for analysis

3. 공간 확률 밀도 함수 추정 모형

3.1. 혼합 조건부 자기회귀 종추출모형

본 절에서는 혼합 조건부 종추출모형에 대해 설명한다. 이를 위해, $\mathcal{D} = \{s_1, \ldots, s_n\} \subset \mathbb{R}^2$, $s_i =$ (위도, 경도)를 한국 지역을 포함하는 격자점들의 집합이라 하고, 각 격자점 (또는 위치) s_i에서 관측된 최고온도와 최저온도 값들의 집합을 $\mathcal{Y} = \{y_1(s_i), \ldots, y_{n_i}(s_i),\ s_i \in \mathcal{D}\}$라 하자. 이 때, 각 위치에서 주변지역의 정보를 활용한 확률 밀도 함수 추정모형은 다음과 같다. 각 위치 s_i마다 정의된 랜덤확률측도 (random probability measure)들의 집합 $\{G_{s_i},\ i = 1, \ldots, n\}$은 머서 조건부 자기회귀 종추출모형 (mercer conditional autoregressive species sampling model, 약자로 MCS라 쓴다)을 따르고, $y_j(s_i)$들은 G_{s_i}의 정규 혼합모형 (mixtures of normal distributions)을 따른다고 가정한다. 즉,

$$
\begin{aligned}
y_j(s_i) &\overset{\text{iid}}{\sim} \int \mathrm{N}(y|\mu, V) dG_{s_i}(\mu, V), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, n, \\
(G_{s_i},\ i = 1, \ldots, n) &\sim MCS(G_0, \alpha, \beta, \tau, \rho),
\end{aligned} \tag{3.1}
$$

$\mathrm{N}(\cdot \mid \mu, V)$는 평균이 μ, 분산이 V인 정규분포 (Normal distribution)를 나타내고, MCS는 머서 조건부 자기회귀 종추출모형 (Jo 등, 2016)을, G_0와 $\alpha, \beta, \tau, \rho$는 각각 기반 측도 (base measure)와 초모수 (hyper-parameters)들을 나타낸다. 참고로, 위에서 제시한 무한차원의 정규혼합모형은 임의의 절대 연속 분포 (absolutely continuous distribution)를 추정가능한 것으로 알려져 있다 (Lo, 1984).

머서 조건부 자기회귀 종추출모형은 종추출모형을 기반으로 정의된 종속 랜덤확률측도들의 집합이다. 머서 조건부 자기회귀 종추출모형을 설명하기에 앞서 종추출모형을 설명하고자 한다. 랜덤확률측도 G가

$$
G = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}
$$

의 형태를 갖고, G의 원자 (Atom) θ_h와 가중치 (Weight) p_h들은 서로 독립이면서, $\theta_h \overset{\text{iid}}{\sim} G_0$이고, 확률 1로 $\sum p_h = 1,\ p_h \geq 0, \forall h$이면, 즉, $Pr\left(\sum p_h = 1\right) = 1$이면, G를 종추출모형을 따른다고 한다. 여기서, G_0은 G의 기반측도이고 δ_{θ_h}는 크로네커 델타 (Kronecker delta)함수이다. 종추출모형은 디리크레 과정 (Dirichlet process; Ferguson, 1973)과 피트만-요 과정 (Pitman-Yor process; Pitman과 Yor, 1997)을 포함하는 일반적인 랜덤확률측도들의 집합니다. Jo 등 (2016)은 종추출모형을 기반으로하여 공간적으로 각 위치마다 정의된 랜덤확률측도들의 분포인 조건부 자기회귀 종추출모형을 정의하였다. 특히, Jo 등 (2016)은 두 개의 조건부 자기회귀 종추출모형, 머서 (Mercer)와 클레이튼-칼더 (Clayton-Kaldor) 모형을 이용하여 조건부 자기회귀 종추출모형을 정의하였다. 여기서는 머서 조건부 자기회귀 종추출모형을 이용한다. MCS를 현재의 자료에 맞는 형태로 서술하면 다음과 같다.

$$
\begin{aligned}
G_{s_i} &= \sum_{h=1}^{\infty} p_h(s_i) \delta_{(\mu_h, V_h)}, \\
p_h(s_i) &= \frac{e^{u_h(s_i)}}{\sum_{k=1}^{\infty} e^{u_k(s_i)}}, \quad (\mu_h, V_h) \overset{\text{iid}}{\sim} G_0(\cdot \mid \mu_0, \alpha, \nu, \psi),\ h = 1, 2, \ldots,
\end{aligned} \tag{3.2}
$$

여기서 가중치는 머서 조건부 자기회귀 모형을 따른다. 구체적으로 머서 조건부 자기회귀 모형의 형태는 다음과 같다.

$$
u_h(s_i) | u_h(s_\ell),\ \ell \neq i \sim \mathtt{N}\Big(m_h(s_i) - \sum_{\ell=1}^{n} \xi_{i\ell}\big(u_h(s_\ell) - m_h(s_\ell)\big), \tau^2\Big), \quad i = 1, \ldots, n,
$$

위의 머서 조건부 가우시안 모형에서 $\tau^2 > 0$이고 각 위치의 평균 $m_h(s_i)$은 가중치의 합이 1이 되도록 하기 위해, 즉, $\sum_{k=1}^{\infty} e^{u_k(s_i)} < \infty$이 만족되도록 하기 위해 Lee 등 (2013)에서 제시된 것처럼 다음과 같이 정의하고

$$m_h(s_i) = \log\{1 - (1 + e^{\beta - \alpha h})^{-1}\}, \quad i = 1, \ldots, n, \tag{3.3}$$

주변지역과의 상관성을 나타내는 계수 $\xi_{i\ell}$는 머서 커널 (Mercer kernel)을 이용하여 아래와 같이 정의한다.

$$\xi_{i\ell} = \begin{cases} \exp(-\rho||s_i - s_\ell||^2) & if \;\; i \neq \ell \\ 0 & if \;\; i = \ell \end{cases} \tag{3.4}$$

단, α, β, ρ는 양의 값을 가지는 실수 값이다. 또한, 기반측도 G_0는 아래와 같이 켤레사전분포 (conjugate prior)로 알려진 정규분포와 감마분포 (gamma distribution)를 이용하고

$$G_0(\cdot \mid \mu_0, \alpha, \nu, \psi) = \mathtt{N}(\mu; \mu_0, g_0 V)\mathtt{Gamma}(1/V; \nu_0/2, \nu_0\psi_0/2) \tag{3.5}$$

초모수의 사전분포는 다음과 같이 가정하였다.

$$\mu_0 \sim \mathtt{N}\big(\mu_0; \mu_{00}, \kappa^2\big), \tag{3.6}$$

$$g_0^{-1} \sim \mathtt{Gamma}\big(g_0^{-1}; a_g, b_g\big), \tag{3.7}$$

$$\psi_0 \sim \mathtt{Gamma}\big(\psi_0; \nu_{00}/2, \nu_{00}\psi_{00}^{-1}/2\big), \tag{3.8}$$

여기서 초모수 사전분포의 인수들에 대해 가중치를 구성하는 모수는 Lee 등 (2013)에서 제시된 것처럼 $\alpha = 0.5, \beta = 20$ 그리고 $\rho = 4, \tau = 1$을 사용하였으며, 기저분포의 초모수는 무정보적 사전분포 (noninformative prior)를 나타내기 위해 $\mu_{00} = 0, \kappa^2 = 1000, a_g = 1, b_g = 100, \nu_0 = 4, \nu_{00} = 4, \psi_{00} = 1/2, a_\tau = 0.01$, 그리고 $b_\tau = 0.01$의 값으로 고정 하였다. 초모수를 설정하는 다른 방법에 대해서는 Jo 등 (2016)을 참고하기 바란다.

3.2. 사후분포의 계산

사후 분포로 부터로 표본을 추출하기 위해 본 논문에서는 Ishwaran과 James (2001)에서 제시된 블럭 깁스 샘플링 알고리즘 (Block Gibbs sampling algorithm)과 Scott (2011)에서 제시된 다항 로짓 모형 (multinomial logit model)에 대한 알고리즘을 기반으로 하여 구성하였다. 특히, 블럭 깁스 샘플링 알고리즘을 구현하기 위해 혼합모형의 구성성분은 $K = 100$으로 고정하였다. 알고리즘을 설명하기 위해, 먼저 c_{ji}, $j = 1, \ldots, n_i$, $i = 1, \ldots, n$를 각 위치 s_i에서의 j번째 관측치 $y_j(s_i)$를 혼합모형의 h번째 성분에 할당하는 잠재 변수 (latent variables)라 하고 $\boldsymbol{m}_h = (m_h(s_1), \ldots, m_h(s_n))^T$, $\boldsymbol{\Gamma} = \boldsymbol{I} - (\xi_{i\ell})_{i,\ell=1}^n$라 하자. (참고. 잠재 변수 c_{ji}는 보다 효율적으로 사후표본 추출이 가능하도록 도와준다. 이와 관련된 설명은 Neal (2000)과 Seo와 Kim (2014)에 나타나 있다.) 이 때 추정모형 (3.1)는 아래와 같이 계층적 모형으로 표현 할 수 있고

$$\begin{aligned} y_j(s_i) \mid \mu_{c_{ji}}, V_{c_{ji}} &\overset{\text{iid}}{\sim} \mathtt{N}(\mu_{c_{ji}}, V_{c_{ji}}), \;\; j = 1, \ldots, n_i \\ c_{ji} \mid u_h(s_i), h = 1, \ldots, K &\overset{\text{iid}}{\sim} \sum_{h=1}^{K} p_h(s_i)\delta_h, \\ \boldsymbol{u}_h = \{u_h(s_1), \ldots u_h(s_n)\} &\sim \mathtt{N}_n\big(\boldsymbol{m}_h, \tau^2\boldsymbol{\Gamma}^{-1}\big), \\ \mu_h \mid \mu_0, g_0, V_h &\sim \mathtt{N}(\mu_0, g_0 V_h), \; h = 1, \ldots, K \\ V_h^{-1} \mid \nu_0, \psi_0 &\sim \mathtt{Gamma}(\nu_0/2, \nu_0\psi_0/2), \; h = 1, \ldots, K \end{aligned} \tag{3.9}$$

이 모형으로 부터 계산된 다음의 결합 확률 사후분포 (joint posterior)를 이용하여 블럭 깁스 샘플링 알고리즘을 구성하였다.

$$\begin{aligned}
&\pi(\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{V},\boldsymbol{u},\mu_0,g_0,\psi_0,\alpha,\beta,\tau^2,\rho\,|\,\boldsymbol{y})\\
&\quad\propto\prod_{i=1}^{n}\prod_{j=1}^{n_i}\mathtt{N}\big(y_j(s_i);\mu_{c_{ji}},V_{c_{ji}}\big)\prod_{i=1}^{n}\prod_{j=1}^{n_i}\prod_{h=1}^{K}\left(\frac{e^{u_h(s_i)}}{\sum_{k=1}^{K}e^{u_k(s_i)}}\right)^{I(c_{ji}=h)}\prod_{h=1}^{K}\mathtt{N}_n\big(\boldsymbol{m}_h,\tau^2\boldsymbol{\Gamma}^{-1}\big)\\
&\qquad\times\prod_{h=1}^{K}\mathtt{N}(\mu_h;\mu_0,g_0V_h)\mathtt{Gamma}\Big(V_h^{-1};\nu_0/2,\nu_0\psi_0/2\Big)\\
&\qquad\times\mathtt{N}(\mu_0;\mu_{00},\kappa^2)\mathtt{Gamma}\Big(g_0^{-1};a_g,b_g\Big)\mathtt{Gamma}\Big(\psi_0;\nu_{00}/2,\nu_{00}\psi_{00}^{-1}/2\Big)
\end{aligned}$$

여기서 $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_K),\boldsymbol{V}=(V_1,\ldots,V_K),\boldsymbol{u}=(\boldsymbol{u}_h,h=1,\ldots,K)$, $\boldsymbol{c}=(c_{ji},\ j=1,\ldots,n_i,\ i=1,\ldots,n)$이고 $\boldsymbol{y}=\{y_j(s_i),\ j=1,\ldots,n_i,\ i=1,\ldots,n\}$이다. 먼저 사후분포로 부터 추출하고자 하는 모수 $\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{V},\boldsymbol{u},\mu_0,g_0,\psi_0$의 초기값을 설정한다. 그 때 반복 $b=1,\ldots,B$에 대하여 다음의 단계를 통해 표본을 추출한다.

1. $j=1,\ldots,n_i,\ i=1,\ldots,n$에 대하여

$$\begin{aligned}
\pi(c_{ji}^{(b)}|\boldsymbol{y},\boldsymbol{\mu}^{(b-1)},\boldsymbol{V}^{(b-1)},\boldsymbol{u}^{(b-1)})&\sim\sum_{h=1}^{K}w_h(s_i)^{(b-1)}\delta_h(c_{ji}),\\
w_h(s_i)^{(b-1)}&=\frac{p_h(s_i)^{(b-1)}\mathtt{N}(y_j(s_i);\mu_h^{(b-1)},V_h^{(b-1)})}{\sum_{k=1}^{K}p_k(s_i)^{(b-1)}\mathtt{N}(y_j(s_i);\mu_k^{(b-1)},V_k^{(b-1)})},\\
p_h(s_i)^{(b-1)}&=\frac{e^{u_h(s_i)^{(b-1)}}}{\sum_{\ell=1}^{K}e^{u_\ell(s_i)^{(b-1)}}},\ h=1,\ldots,K.
\end{aligned}$$

2. $h=1,\ldots,K$에 대하여

$$\begin{aligned}
\pi(\mu_h^{(b)}\mid\boldsymbol{y},g_0^{(b-1)},V_h^{(b-1)})&\sim\mathtt{N}(\hat{\mu}_h,\hat{V}_h),\\
\hat{\mu}_h&=\frac{n_h\bar{y}_h+\mu_0^{(b-1)}/g_0^{(b-1)}}{n_h+1/g_0^{(b-1)}},\\
\hat{V}_h&=\frac{V_h^{(b-1)}}{1/g_0^{(b-1)}+n_h},
\end{aligned}$$

여기서 $n_h=\sum_{i=1}^{n}\sum_{j=1}^{n_i}I(c_{ji}^{(b-1)}=h)$이고 $\bar{y}_h=\sum_{\{(j,i):c_{ji}^{(b-1)}=h\}}y_j(s_i)/n_h$이다.

3. $h=1,\ldots,K$에 대하여

$$\begin{aligned}
\pi(V_h^{(b)}\mid\boldsymbol{y},\mu_h^{(b-1)},\psi_0^{(b-1)})&\sim\mathtt{Gamma}\Big\{(\nu_0+n_h)/2,\hat{\psi}_h/2\Big\},\\
\hat{\psi}_h&=\nu_0\psi_0^{(b-1)}+\sum_{\{(i,j):c_{ji}=h\}}(y_j(s_i)-\mu_h^{(b-1)})^2.
\end{aligned}$$

4. $\pi(\mu_0^{(b)}|\boldsymbol{\mu}^{(b-1)},\boldsymbol{V}^{(b-1)})\sim N(\hat{\mu}_0,\hat{V}_0)$,

$$\begin{aligned}
\hat{V}_0&=\Big\{(1/g_0^{(b-1)})\sum_{h=1}^{K}(1/V_h^{(b-1)})+1/\kappa^2\Big\}^{-1},\\
\hat{\mu}_0&=\hat{V}_0\Big\{(1/g_0^{(b-1)})\sum_{h=1}^{K}\big(\mu_h^{(b-1)}/V_h^{(b-1)}\big)+\mu_{00}/\kappa^2\Big\}.
\end{aligned}$$

5. $\pi(g_0^{-1(b)}|\boldsymbol{\mu}^{(b-1)}, \boldsymbol{V}^{(b-1)}) \sim \texttt{Gamma}\left\{a_g + K/2, b_g + \left(\sum_{h=1}^{K}(\mu_0^{(b-1)} - \mu_h^{(b-1)})^2/V_h^{(b-1)}\right)/2\right\}$.
6. $\pi(\psi_0^{(b)}|\psi_{00}^{(b-1)}, \boldsymbol{V}^{(b-1)}) \sim \texttt{Gamma}\left\{(\nu_{00} + \nu_0 K)/2, \left(\nu_{00}\psi_{00}^{-1} + \nu_0 \sum_{h=1}^{K}\left(1/V_h^{(b-1)}\right)\right)/2\right\}$.
7. $i = 1, \ldots, n$에 대하여 $\{u_h(s_i),\, h = 1, \ldots, K\}$의 사후표본은 평균이 $e^{u_h(s_i)}$인 지수분포를 따르는 잠재변수 λ_{jih}를 이용하여 추출한다. 특히, 본 논문은 Roberts와 Rosenthal (2009)의 적응 메트로폴리스-깁스 (Adaptive Metropolis-within-Gibbs)방법을 사용한다. 이외에 다른 추출방법으로는 Kim 등 (2015)에 설명된 적응 기각 추출 (Adaptive rejection sampling)방법이 있다.

   (a) $j = 1, \ldots, n_i$, $h = 1, \ldots, K$에 대하여

$$\pi(\lambda_{ji,c_{ji}}^{(b)}|u_h(s_i)^{(b-1)}, h = 1, \ldots, K) \sim Exp\left(\sum_{k=1}^{K} e^{u_k(s_i)^{(b-1)}}\right),$$
$$\lambda_{jih'}^{(b)} = \lambda_{ji,c_{ij}}^{(b)} + \lambda_{jih'}^{*},\ \ \lambda_{jih'}^{*} \sim Exp(e^{u_h(s_i)^{(b-1)}}),\ h' \neq c_{ji}$$

   (b) $h = 1, \ldots, K$에 대하여

   i. $\texttt{N}(u_h(s_i)^{(b-1)}, v_u^2)$부터 $u_h(s_i)'$를 뽑는다.

   ii. 다음의 식을 이용하여 채택확률을 계산한다.

$$\tilde{\alpha}\{u_h(s_i)', u_h(s_i)^{(b-1)}\} = \min\left\{1, \frac{f(u_h(s_i)')}{f(u_h(s_i)^{(b-1)})}\right\},$$

   여기서

$$f\big(u_h(s_i)\big) = \pi(u_h(s_i) \mid \alpha, \beta, \tau^2, \rho, \boldsymbol{\lambda}_{ih})$$
$$\propto \texttt{N}\left(m_h(s_i) - \sum_{\ell=1}^{n}\xi_{i\ell}(u_h(s_i) - m_h(s_\ell)), \tau^2\right)\prod_{j=1}^{n_i} e^{u_h(s_i)}e^{-\lambda_{ijh}e^{u_h(s_i)}}$$

   iii. $\texttt{Unif}(0,1)$으로 부터 생성된 값보다 채택확률 $\tilde{\alpha}\{u_h(s_i)', u_h(s_i)^{(b-1)}\}$이 크면 새로운 값 $u_h(s_i)'$을 반복 b에서의 표본 $u_h(s_i)^{(b)}$의 값으로 선택한다.

   iv. 채택확률이 44%에 가깝도록 반복 50번째 마다 b번째 까지의 채택율 (acceptance rate) r을 계산하여 v_u^2의 값을 다음과 같이 갱신한다.

$$\log(v_u) = \begin{cases} \log(v_u) + \min(0.01, 1/\sqrt{b}) & \text{if } r > 0.44 \\ \log(v_u) - \min(0.01, 1/\sqrt{b}) & \text{if } r < 0.44. \end{cases}$$

3.3. 라오-블랙웰 추정량

각 위치 s_i, $i = 1, \ldots, n$에서의 사후 예측 밀도함수 (posterior predictive density)를 추정하기 위해 본 논문에서는 라오-블랙웰 추정량 (Rao-Blackwellized estimator)을 이용한다. 라오-블랙웰 추정량은 사후분포가 해석적으로 계산이 불가능 할 때 사용되는 방법으로 효율적이면서 안정적인 추정량으로 알려져 있다. 또한 깁스 샘플링 알고리즘에서 얻어지는 마르코프 연쇄 (Markov chain)로 부터 계산이 가능하여 추가적인 계산을 필요로 하지 않는다. 라오-블랙웰 추정량을 구하기 위해 $\{\Psi^{(b)}, b = 1, \ldots, B\}$를 3.2절에서 설명된 깁스 샘플링 알고리즘으로 부터 얻어진 관심 모수의 사후 분포 표본이라 하자. 그 때 사후 예측 밀도함수의 라오-블랙웰 추정량은 아래와 같다.

$$\hat{f}(y_{n+1}(s_i)|\boldsymbol{y}) \approx \frac{1}{B}\sum_{b=1}^{B} f(y_{n+1}(s_i) \mid \boldsymbol{y}, \Psi^{(b)}),$$

여기서

$$f(y_{n+1}(s_i) \mid \boldsymbol{y}, \Psi^{(b)}) = \sum_{h=1}^{K} p_h(s_i)^{(b)} \mathtt{N}(y_{n+1}(s_i) \mid \mu_h^{(b)}, V_h^{(b)})$$

$$p_h(s_i)^{(b)} = \frac{e^{u_h(s_i)^{(b)}}}{\sum_{k=1}^{K} e^{u_k(s_i)^{(b)}}}.$$

4. 자료 분석 결과

본 절에서는 3.1절에서 기술된 위치 종속 혼합모형과 각 모수에 할당된 사전분포를 바탕으로 하여 여름철 평균 최고 온도 자료와 여름철 평균 최저 온도자료에 대해 베이지안 분석을 실시한다. 자료적합을 위해 본 논문에서는 포트란 (`http://www.gnu.org`) 과 R (`http://r-project.org`) 프로그램을 이용하여 3.2절에서 구성된 블럭 깁스 샘플링 방법을 통해 사후표본을 추출하였다. 서로 다른 초기값을 이용하여 두 개의 마르코프 연쇄 (Markov chain)를 구성하였다. 각 연쇄마다 총 60,000번의 반복 (iteration) 추출하였고 그 중에서 초기의 10,000번의 추출과정을 소각 (burn-in)하고 그 후 10번의 반복마다 하나의 표본을 채택하여 5,000개의 사후표본을 얻었다. 두 연쇄으로 부터 나온 총 10,000개의 표본을 이용하여 모형적합에 사용하였다. 분석에 앞서 추출된 사후표본의 수렴성을 판단하기 위해 본 논문에서는 Gelman과 Rubin (1992)에 의해 제안된 Gelman-Rubin 검정방법을 사용하였다. Gelman-Rubin 검정은 전통적인 분산분석 (Analysis of Variance) 검정을 이용하여 여러 마르코프 연쇄 간의 분산과 각 마르코프 연쇄 내의 분산을 비교하여 수렴성을 판단하는 방법으로 Gelman-Rubin 통계량 값이 1에 가까우면 사후표본들이 정상분포 (Stationary distribution)로 잘 수렴되었음을 나타낸다. 구체적인 식과 자세한 설명은 Gelman 등 (2004)에서 살펴볼 수 있다. 참고로 Gelman-Rubin 검정은 R의 `coda`패키지를 통해 쉽게 사용할 수 있다.

Table 4.1은 두 개의 마르코프 연쇄로 부터 얻어진 10,000개의 사후표본에 대한 요약된 Gelman-Rubin 통계량을 나타낸다. Table 4.1로부터 Gelman-Rubin 통계량의 값은 대부분 1에 매우 가까움을 알 수 있었고 이는 10,000번 이후의 사후표본들이 정상분포로 잘 수렴됨을 나타내고 있다.

Table 4.1 Gelman-Rubin statistics.

Parameters	Maximum temperature		Minimum temperature	
	Point est.	Upper C.I.	Point est.	Upper C.I.
μ_0	1.00	1.01	1.00	1.02
g_0	1.01	1.01	1.00	1.01
ψ_0	1.01	1.06	1.01	1.03

추가적으로 서로 다른 두 개의 마르코프 연쇄에서 나온 기저분포의 각 모수에 대한 사후표본의 자취그림 (trace plot)과 자기상관함수 그림 (autocorrelation plot)이 Figure 4.1에 주어져 있다. Figure 4.1은 여름철 최고 기온에 대한 그림을 나타낸다. 이러한 그림에서도 역시 주기성이나 추세가 없이 특정한 범위 안에서 흩어져 있음을 알 수 있으며 따라서 사후 표본들이 정상분포로 적절하게 잘 수렴하는 것을 확인할 수 있다. Figure 4.2는 기반측도의 각 모수들에 대한 사후 분포를 나타낸다. 빨간색 선은 최저 온도에 대한 기반측도 모수들의 사후분포를 나타내고 검은색 선은 최고 온도의 기반측도에 대한 모수들의 사후분포를 나타낸다.

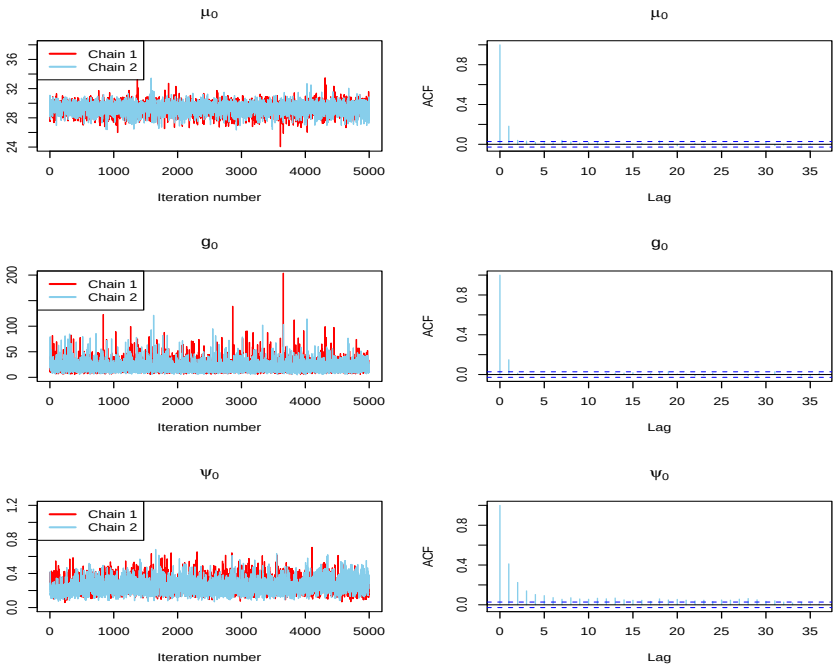

Figure 4.1 Trace plots and autocorrelation plots for parameters of base measure of summer maximum temperature.

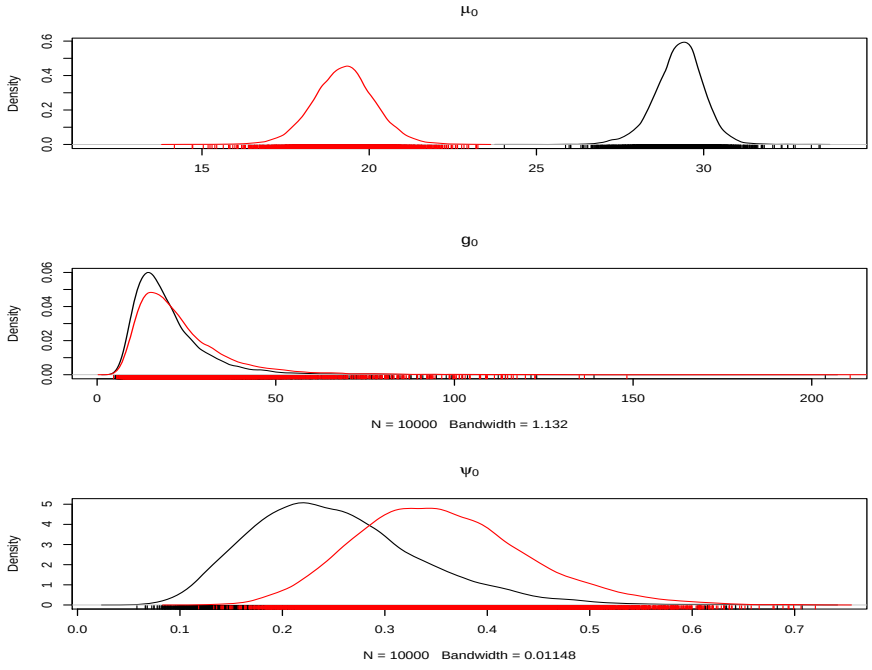

Figure 4.2 Posterior distributions for parameters of base measure: summer minimum temperature (red line) and summer maximum temperature (black line).

Figure 4.3는 위치 종속 종추출모형으로 부터 각 위치마다 추정된 여름철 최대온도, 최저온도의 확률밀도함수의 33%와 66%에 해당하는 분위수 (quantile)의 임계값 (critical value)을 나타낸다. 여기서, 33% 분위수는 각각의 위치에서 추정된 확률밀도함수로 부터 누적확률 (cumulative probability)이 33%가 되는 임계값을 나타내고, 66% 분위수는 각 위치의 추정된 확률밀도함수에서 누적확률값이 66%가 되는 임계값을 나타낸다. 그림으로 부터 알 수 있듯이 인접지역간 유사한 패턴을 보이고 있다. 따라서, 본 논문에서 고려한 위치관계를 고려하는 종추출모형이 적절하다고 판단할 수 있다. 또한 Figure 4.2로 부터 한국지역 전체의 평균 최저 기온과 평균 최고 기온의 차이가 약 10도 이상 차이나는 것을 알 수 있다. 이는 여름철 온도 변화가 크다는 것을 암시하고 따라서 급격한 기온 상승에 대비해야 할 것으로 보인다. 왜냐하면 Jo 등 (2012)에서도 언급 되었듯이 급격한 온도 상승은 사망율을 증가시킬 수 있기 때문이다.

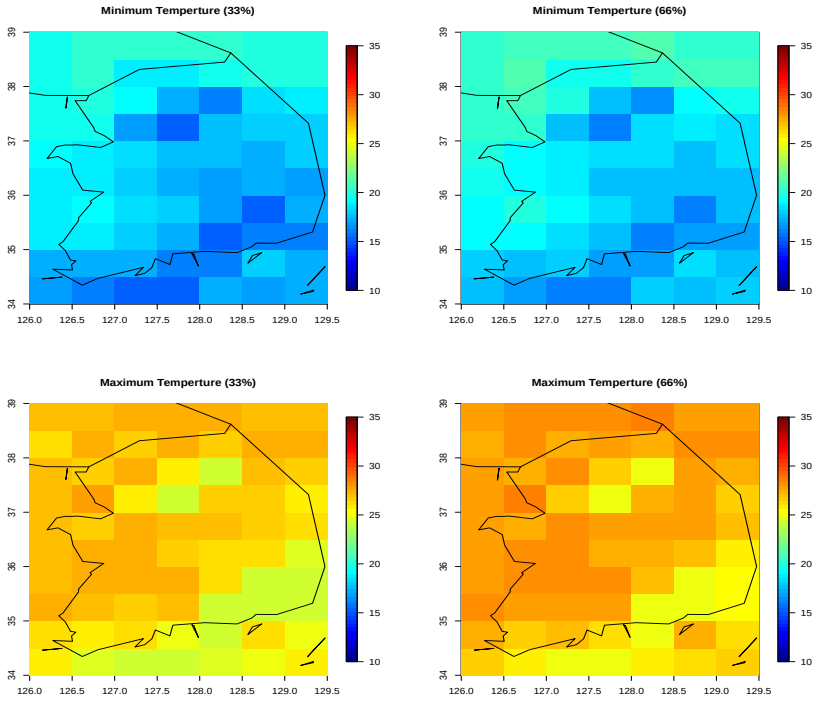

Figure 4.3 Tercile based estimates of summer minimum temperature (top panels) and summer maximum temperature (bottom panels).

Table 4.2는 우리나라 주요도시 (서울, 부산, 인천, 대구, 대전, 광주, 울산)의 위도와 경도를 나타내고 Figure 4.4와 4.5는 주요도시의 여름철 극한 기온에 대한 혼합 조건부 종추출모형으로부터 추정된 확률밀도함수를 보여준다. Figure 4.4은 여름철 최저 기온의 밀도함수 추정치를 나타내고, Figure 4.5는 여름철 최고 기온의 밀도함수 추정치를 나타낸다. Figure 4.4와 4.5에서 볼 수 있듯이 기온의 변화는 최저 기온보다 최고 기온에서 더 뚜렷이 나타나고 있고, 부산과 울산 지역이 다른 지역에 비해 여름철 최고 온도의 변화가 크게 나타나고 있다. 따라서 부산, 울산 지역의 경우 타 지역에 비해 여름철 급격히 변하는 최고 온도에 대비해야 할 것으로 보인다. 서울의 경우 다른 지역에 비해 최고, 최저 기온의 밀도함수가 일정하게 나타나고 있음을 알 수 있고, 또한 전체적으로 남쪽에 비해 위쪽지역의 온도가 높게 나타남을 알 수 있다.

Table 4.2 Major cities over Korea

City	Longitude	Latitude
Seoul	127.25	37.75
Busan	129.25	35.25
Incheon	126.25	37.25
Daegu	128.25	35.75
Daejeon	127.25	36.25
Gwangju	126.75	35.25
Ulsan	128.25	35.25

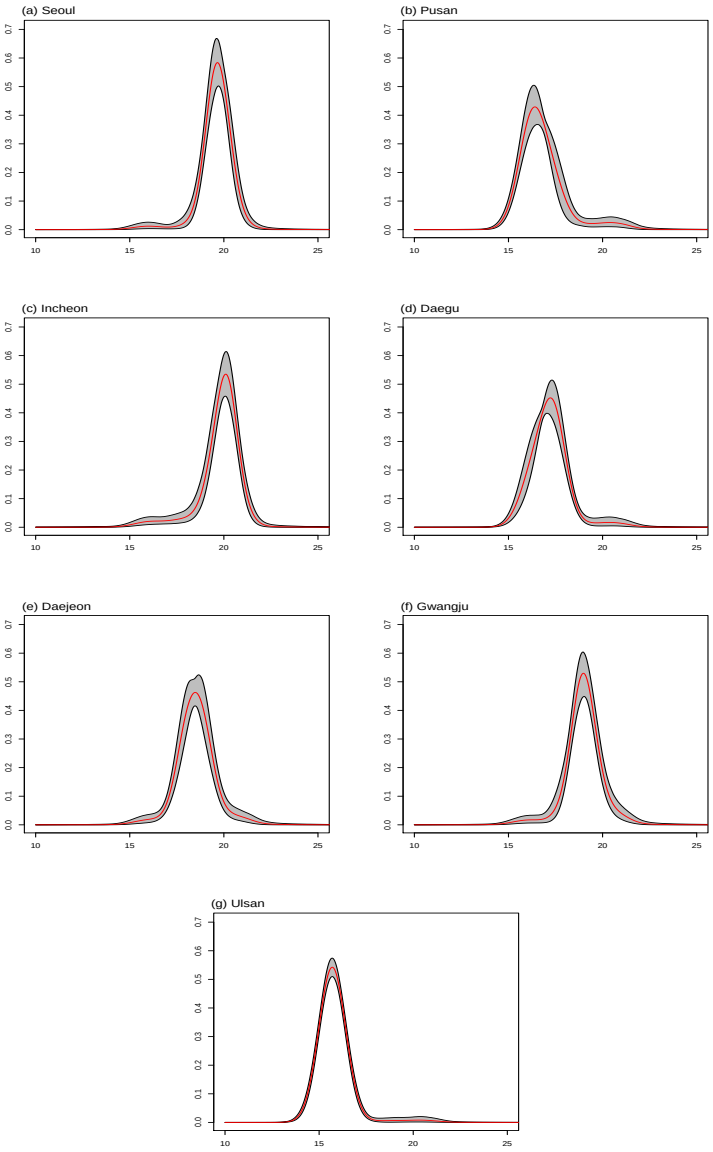

Figure 4.4 Density estimates of summer minimum temperature for the major cities of South Korea.

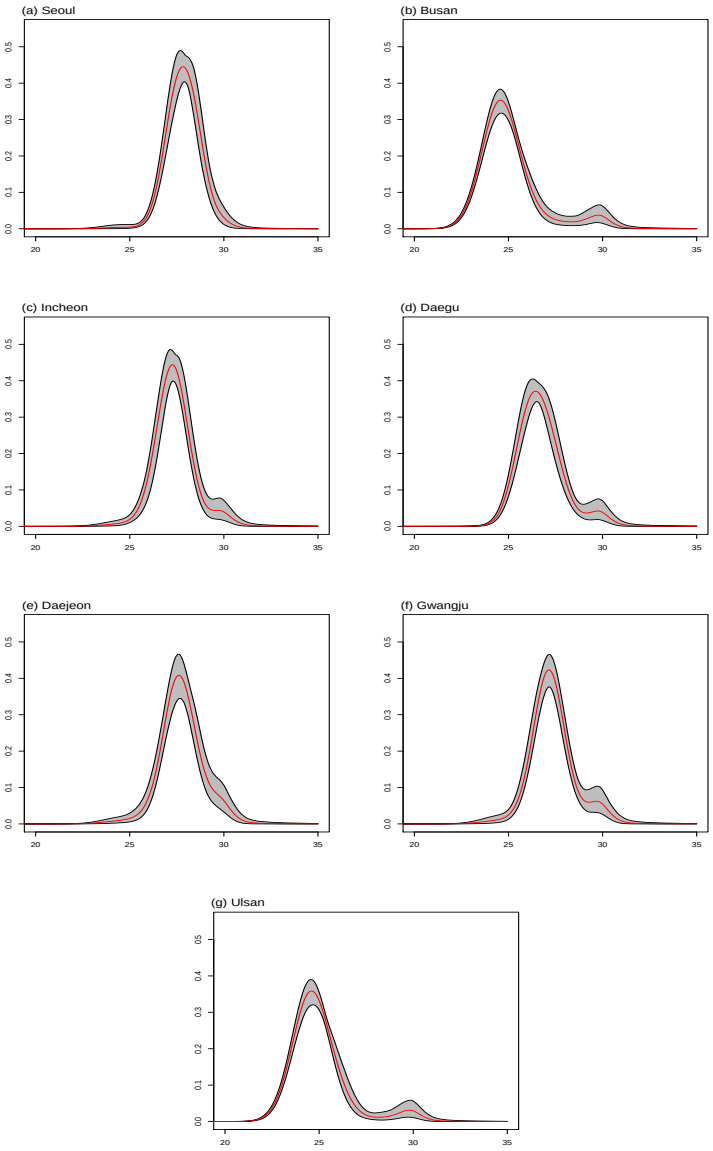

Figure 4.5 Density estimates of summer maximum temperature for the major cities of South Korea.

5. 결론

본 논문에서는 한국지역의 극한 온도에 대한 확률밀도 함수를 추정하였다. 특히, 6월부터 8월까지 여름철 평균 최고온도와 평균 최저온도의 확률밀도함수를 추정하였다. 기상자료가 가지고 있는 공간적 상관관계를 고려하기 위해 위치 정보를 사용하는 가우시안 조건부 자기회귀모형을 기반으로 하는 종속 종추출 혼합모형을 이용하였으며 모형으로 부터 사후표본을 추출하기 위해 블럭 깁스 샘플링 알고리즘과

다항 로짓 알고리즘을 기반으로 하는 깁스 샘플링 알고리즘을 설명하였다. 추출된 사후 표본으로 부터 확률밀도함수를 추정하기 위해 라오-블랙웰 추정량을 이용하였다. 추정된 확률밀도 함수로 부터 33%와 66%에 해당하는 분위수값과 우리나라 주요도시 (서울, 부산, 인천, 대구, 대전, 광주, 울산)에 대한 확률밀도함수 추정치를 제시하였으며, 이 결과로 부터 극한 기온의 분포와 인접지역의 유사패턴을 보이는 것을 알 수 있었다. 또한 추정된 결과로 부터 한국지역의 여름철 온도변화가 10도 이상 나는 것을 알 수 있었으며 남쪽지역에 비해 북쪽지역이 최고, 최저 온도가 더 높은 것을 알 수 있었다. 부산, 울산지역의 경우 최고 온도의 변화가 다른 지역에 비해 더 뚜렷이 나타남을 또한 알 수 있었다.

본 논문에서 사용된 머서 조건부 자기회귀 종추출모형은 두 가지 장점을 가지고 있다. 첫째는 공간적 상관관계를 가중치를 통해서 이루어지기 때문에 원자를 통해 상관관계를 고려하는 모형에 비해 더 유연하고 정확한 밀도함수 추정이 가능하고 가우시안 조건부 자기회귀 모형을 사용하기 때문에 직관적으로 공간적 상관관계를 고려한다 (Jo 등, 2016). 둘째는 무한차원의 혼합모형이기 때문에 혼합하는 분포의 갯수를 미리 정하지 않고 자료로 부터 자연스럽게 추정할 수 있다. 이는 비모수 베이지안 모형이 가지는 가장 큰 장점중 하나이다. 그러나, 본 논문에서는 극한 기온 밀도함수 추정이 가능한 기존의 모형과 비교를 실시하지 않았다. 따라서 추후에 이와 관련된 연구를 진행할 예정이다.

References

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209-230.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, **100**, 1021-1035.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457-511.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian data analysis*, 2nd Ed., Chapman & Hall/CRC, Boca Raton, Florida.

Griffiths, G. M., Chambers, L. E., Haylock, M. R., Manton, M. J., Nicholls, N., Baek, H., Choi, Y., Della-marta, P. M., Gosai, A., Iga, N., Lata, R., Laurent, V., Maitrepierre, L., Nakamigawa, H., Ouprasitwong, N., Solofa, D., Tahani, L., Thuy, D. T., Tibig, L., Trewin, B., Vediapan, K., and Zhai, P. (2005). Change in mean temperature as a predictor of extreme temperature change in the Asia-Pacific region. *International Journal of Climatology*, **25**, 1301-1330.

Harris, I. Jones, P. D., Osborn, T. J. and Lister, D. H. (2013). Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset. *International Journal of Climatology*, **34**, 623-642, doi:10.1002/joc.3711.

Im, E. and Kwon, W. (2007). Characteristics of extreme climate sequences over Korea using a regional climate change scenario. *SOLA*, **3**, 17-20.

Intergovernmental Panel on Climate Change (2007). *Climate changes: The physical science basis*, Cambridge University Press, United Kingdom and New York.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161-137.

Jo, S., Lee, J., Müller, P., Quintana, F. A. and Trippa, L. (2016). Dependent species sampling models for spatial density estimation. *Bayesian Analysis*, 1-28, doi:10.1214/16-BA1006.

Jo, Y., Lim, Y., Kim, H. and Lee, J. (2012). Bayesian analysis for heat effects on mortality. *Communications of the Korean Statistical Society*, **19**, 705-720.

Kim, H., Jo, S. and Choi, T. (2015). Performance comparison of random generators based on adaptive rejection sampling. *Journal of the Korean Data & Information Science Society*, **26**, 593-610.

Lee, J., Quintana, F., Müller, P. and Trippa, L. (2013). Defining predictive probability functions for species sampling models. *Statistical Science*, **28**, 209-222.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, **12**, 351-357.

Mearns, L. O., Katz, R. W. and Schneider, S. H. (1984). Extreme high-temperature events: changes in their probabilities with changes in mean temperature. *Journal of Climate and Applied Meteorology*, **23**, 1601-1613.

Meehl, G. A., Karl, T., Easterling, D. R., Changnon, S., Pielke Jr., R., Changnon, D., Evans, J., Groisman, P. Y., Knutson, T. R., Kunkel, K. E., Mearns, L. O., Parmesan, C., Pulwarty, R., Root, T., Sylves, R. T., Whetton, P. and Zwiersl, F. (2000). An introduction to trends in extreme weather and climate events: observations, socioeconomic impacts, terrestrial ecological impacts, and model projections. *Bulletin of the American Meteorological Society*, **81**, 413-416.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249-265.

National Institute of Meteorological Research (2009). Understaning climate changes, *Technical Report*, Available from `http://climate.go.kr/home/bbs/dw_proc.php`.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, **25**, 855-900.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, **18**, 349-367.

Scott, S. L. (2011). Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models. *Statistical Papers*, **52**, 639-650.

Seo, J. and Kim, Y. (2014). Nonparametric Bayesian estimation on the exponentiated inverse Weibull distribution with record values. *Journal of the Korean Data & Information Science Society*, **25**, 611-622.

Journal of the Korean Data & Information Science Society
2016, **27**(5), 1155–1168

http://dx.doi.org/10.7465/jkdi.2016.27.5.1155
한국데이터정보과학회지

Density estimation of summer extreme temperature over South Korea using mixtures of conditional autoregressive species sampling model[†]

Seongil Jo[1] · Jaeyong Lee[2]

[1]Department of Statistics and Applied Probability, National University of Singapore
[2]Department of Statistics, Seoul National University

Received 20 July 2016, revised 24 August 2016, accepted 2 September 2016

Abstract

This paper considers a probability density estimation problem of climate values. In particular, we focus on estimating probability densities of summer extreme temperature over South Korea. It is known that the probability density of climate values at one location is similar to those at near by locations and one doesn't follow well known parametric distributions. To accommodate these properties, we use a mixture of conditional autoregressive species sampling model, which is a nonparametric Bayesian model with a spatial dependency. We apply the model to a dataset consisting of summer maximum temperature and minimum temperature over South Korea. The dataset is obtained from University of East Anglia.

Keywords: Conditional autoregressive species sampling model, density estimation, extreme temperature, mercer conditional autoregressive model, spatial correlation

[†] This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0030811)
[1] Corresponding author: Research fellow, Department of Statistics and Applied Probability, National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077, Singapore.
E-mail: joseongil@gmail.com
[2] Professor, Department of Statistics, Seoul National University, Seoul 08826, Korea.