Journal of Environmental Science International
27(7); 499~507; July 2018

pISSN: 1225-4517 eISSN: 2287-3503
https://doi.org/10.5322/JESI.2018.27.7.499

ORIGINAL ARTICLE

농업기상 결측치 보정을 위한 통계적 시공간모형

박다인 · 윤상후[1)*]

대구대학교 통계학과, [1)]대구대학교 수리빅데이터 학부 통계·빅데이터 전공

A Missing Value Replacement Method for Agricultural Meteorological Data Using Bayesian Spatio-Temporal Model

Dain Park, Sanghoo Yoon[1)*]

Department of Statistics, Daegu University, Gyeongsan, 38453, Korea
[1)]Division of Mathematics and big data science, Daegu University, Gyeongsan 38453, Korea

Abstract

Agricultural meteorological information is an important resource that affects farmers' income, food security, and agricultural conditions. Thus, such data are used in various fields that are responsible for planning, enforcing, and evaluating agricultural policies. The meteorological information obtained from automatic weather observation systems operated by rural development agencies contains missing values owing to temporary mechanical or communication deficiencies. It is known that missing values lead to reduction in the reliability and validity of the model. In this study, the hierarchical Bayesian spatio–temporal model suggests replacements for missing values because the meteorological information includes spatio–temporal correlation. The prior distribution is very important in the Bayesian approach. However, we found a problem where the spatial decay parameter was not converged through the trace plot. A suitable spatial decay parameter, estimated on the bias of root–mean–square error (RMSE), which was determined to be the difference between the predicted and observed values. The latitude, longitude, and altitude were considered as covariates. The estimated spatial decay parameters were 0.041 and 0.039, for the spatio-temporal model with latitude and longitude and for latitude, longitude, and altitude, respectively. The posterior distributions were stable after the spatial decay parameter was fixed. root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and bias were calculated for model validation. Finally, the missing values were generated using the independent Gaussian process model.

Key words : Agricultural meteorological, Bayesian spatio–temporal Model, Missing value

1. 서 론

농업기상은 농가 수익, 식량 안보, 농업 재해에 많은 영향을 미치는 중요한 국가 자원이다. 때문에 농업 정책의 기획, 수립, 집행, 평가를 담당하는 다양한 분야에서 농업기상을 이용한 다양한 연구가 수행되고 있다(Lee, 2000). 농촌진흥청은 농작물의 생육과 병해충 발생예측, 기상재해방지 등에 활용하기 위해 고품질의 농업기상자료를 제공하는데 힘쓰고 있다.

특정 지역에서 재배하는 작물이나 품종은 그 지방의 기후에 맞게 진화되어 왔다. 날씨가 평년과 비슷할 경우 작물의 생육도 순조롭지만, 최근과 같이 기상의

Received 7 February, 2018; **Revised** 27 March, 2018;
Accepted 26 April, 2018
[*]**Corresponding author**: Sanghoo Yoon, Division of Mathematics and big data science, Daegu University, Gyeongsan 38453, Korea
Phone : +82-53-850-6421
E-mail : statstar@daegu.ac.kr

변화폭이 커지면 작물의 생육도 예년과 달라지며, 그 정도가 심할 때는 큰 피해를 받게 된다(Lee et al., 2012). 특히 지구온난화에 따른 겨울철 및 봄철의 이상기온은 양파 호르몬의 불균형을 초래하고 양파의 전반적인 생육단계를 빠르게 하여 품질을 저하시킨다(Jang et al., 2002). 이 외에도 양파재배와 기상에 관한 연구로 농작물 생산량과 기상요소의 상관관계 분석(Lee et al., 2012)과 온도에 따른 양파의 생리적 특성 및 수량 변화(Lee et al., 2014)등 의 연구가 진행되었다.

농촌진흥청에서 운용하는 자동기상관측장비인 AWS로 부터 얻는 기상정보는 일시적인 기계 장애나 통신 장애등으로 결측치가 포함되어 있다. 결측치가 포함된 자료를 이용한 자료 분석은 연구의 신뢰도와 타당도를 떨어뜨릴 수 있다. 이에 결측치가 보정된 농업기상정보로 연구를 수행한다면 더 좋은 결과를 얻을 수 있을 것이다. 기상정보의 결측치 보정방법으로 인공신경망을 이용한 방법(Min et al., 2016)과, 공간 시계열모형인 STAR 모형(Lee and Kim, 2010)을 이용하여 결측치를 보정하는 연구들이 국내에서 수행되었다.

결측치 처리 방법은 크게 세 가지가 있다. 첫 번째 방법은 결측치를 제거하여 연구를 수행하는 것이다. 결측치 제거 방법은 분석에 사용되는 표본의 수가 줄어들어 통계적 검정력이 감소하게 된다. 또한 연구의 내적 및 외적 타당도가 훼손될 수 있으며, 불완전한 모수추정 등의 문제가 발생한다. 두 번째 방법은 평균값이나 중앙값, 최빈값 등으로 대체하는 방법이다. 이러한 단일값 대체 방법은 손쉽게 완전한 데이터 셋을 구축할 수 있다는 점에서 좋지만, 대부분의 단일대체 값은 편향된 추정을 하는 것으로 알려져 있다(Baraldi and Enders, 2010). 또한, 하나의 값으로 대체되기 때문에 표준오차가 과소 추정되어 통계적 검정력의 관점에서 문제가 되며 해당 변수의 분산을 작게 하고 다른 변수와의 상관관계를 낮추는 등 연구결과의 편향을 가져온다(Ko and Tak, 2016). 결측치를 처리하는 세 번째 방법은 결측치를 예측하여 대체하는 방법이다.

한반도 기상은 계절(시간)과 지역(공간)에 따라 큰 편차를 보이는 시공간적 특성을 내포하고 있기 때문에 관심지역의 기상자료를 예측하기 위해선 통계적 시공간 모형이 필요하다. 시공간 자료 분석의 주된 관심은 특정 공간에서 어떤 반응 변수와 변화를 시간에 따라 부드럽게 예측하는 것이다. 기상관측소 위치와 주산지 농작물의 재배지의 위치는 상이하므로 통계적 시공간 모형을 이용해 주산지의 기상정보를 예측할 수 있다. 본 연구에서는 베이지안 시공간 모델을 이용하여 양파주산지인 전라남도에 대한 농업기상정보의 결측치를 보정하는 방법에 대해 알아보고자 한다.

2. 연구 방법

본 연구에서는 농업기상 데이터의 결측치를 보정하기 위해 계층적 베이지안 시공간 모델을 이용하였다. 베이지안 시공간 모델의 식은 공간 선형 모델과 표현이 유사하다(Yoon and Kim, 2016). 계층적 베이지안 시공간 모형의 사전분포와 예측방법 및 모델평가에 대한 상세한 방법론을 이번 장에서 다룬다.

2.1. 베이지안 시공간모형

베이지안 시공간 모델은 두 단계의 계층 구조로 설명 된다. 첫 번째 단계는 실제 자료 기반 확률과정(true underlying process)을 나타내고, 두 번째 단계는 시공간 랜덤 효과(spatio-temporal random effect)를 나타낸다(Gelfand et al., 2005). 실제 자료 기반 확률과정에서 장기기간 l(year, $l,\cdots,r$)과 단기기간 t(day, $t=1,\cdots\ T_l$)에 따른 종속변수 $\boldsymbol{Z_{lt}}$는 관측값 $\boldsymbol{O_{lt}}$와 오차항 ϵ_{lt}로 구성된다. 위치 $s_i(i=1,\cdots,n)$와 장·단기 기간(l, t)에 따른 종속변수를 $\boldsymbol{Z_l}(s_i,t)$로 표기하며,

$$\boldsymbol{Z_l}=(Z_l(s_1,t),\cdots,Z_l(s_n,t))',$$
$$\boldsymbol{O_{lt}}=O(O_l(s_1,t),\cdots,O_l(s_n,t))',$$
$$\mathrm{N}=n\sum_{l=1}^{r}T_l$$

이다. N은 관측치들의 총 개수이다. 잡음효과인 너겟효과(nugget effet)나 순수오차기간(pure error term)은 $\boldsymbol{\epsilon_{lt}}=(\epsilon_l(s_1,t),\cdots,\epsilon_l(s_n,t))'$이다. σ_ϵ^2이 알려지지 않은 순수오차에 대한 분산이고, I_n은 n차원 단위행렬일 때, $\boldsymbol{\epsilon_{lt}}$ 독립정규분포 $N(0,\sigma_\epsilon^2 I_n)$을 따른다. $\boldsymbol{\eta_{lt}}=$

$(\eta_l(s_1,t),\cdots,\eta_l(s_n,t))'$는 시공간 랜덤효과로 시간에 따른 독립 $N(0,\sum_{\eta})$을 가정하고 분산 공분산행렬 $(\sum_{\eta})$은 $\sigma_\eta^2 S_\eta$이다. σ_η^2는 규모모수이고, S_η는 공간 상관행렬로 다음과 같이 정의되는 지수형 상관함수(exponential function)으로 얻어진다.

$$k(s_i,s_j,\phi)=\exp(-\phi|s_i-s_j|),\ \phi>0$$

모수 ϕ는 거리 $|s_i-s_j|$에 따라 상관관계의 공간 감쇠율을 제어한다.(Cressie, 1994).

2.2. 독립 가우시안 회귀모형(Independent Gaussian process model, GP model)

$\boldsymbol{Z_{lt}}$는 실제자료의 확률과정이고 $\boldsymbol{O_{lt}}$는 시공간적 확률과정이라 하면 독립 가우시안 회귀모형은 다음과 같이 정의된다.

$$\boldsymbol{Z_{lt}}=\boldsymbol{O_{lt}}+\epsilon_{lt},\ \boldsymbol{O_{lt}}=\boldsymbol{X_{lt}}\beta+\eta_{lt}$$

여기서 ϵ_{lt}와 η_{lt}는 독립이고 $n\times p$ 행렬 $\boldsymbol{X_{lt}}$는 절편을 포함한 공변량 p로 구성되어 있다. 즉 $\beta=(\beta_0,\beta_1,\cdots,\beta_{p-1})$는 p개의 회귀계수들로 구성된 $p\times 1$ 벡터이다. $\theta=(\beta,\sigma_\epsilon^2,\sigma_\eta^2,\phi)$는 GP모델의 모수의 집합이고, 사전분포 $\pi(\theta)$를 통해 결측치가 고려된 사후분포를 구할 수 있다. 관측치(z)가 주어졌을 때 결측치(z^*)가 고려된 로그우도함수는 다음과 같다.

$$\log\ \pi(\theta,O,z^*|z)\propto -\frac{N}{2}log\ \sigma_\epsilon^2-$$
$$\frac{1}{2\sigma_\epsilon^2}\sum_{l=1}^{r}\sum_{t=1}^{T_l}(Z_{lt}-O_{lt})'(Z_{lt}-O_{lt})-\frac{\sum_{l=1}^{r}T_l}{2}log|\sigma_\eta^2 S_\eta|$$
$$-\frac{1}{2\sigma_\eta^2}\sum_{l=1}^{r}\sum_{t=1}^{T_l}(O_{lt}-X_{lt}\beta)S_\eta^{-1}(O_{lt}-X_{lt}\beta)+\log\pi(\theta)$$

모수는 깁스샘플링을 통해 추정된다. 깁스샘플링(Gelfand and Smith, 1990)은 두 개 이상의 확률변수의 결합 확률분포로부터 일련의 표본을 생성하는 확률적 알고리즘으로 모수들의 완전 조건부 분포에 필요하다(Banerjee et al., 2014). 사후분포의 완전 조건부 분포는 Bakar and Sahu(2015)를 참고 바란다. 모수 추정을 위해 본 연구에서는 13,000회의 모의를 실시하여 이중 초기값에 영향을 받는 번인 구간(burn-in period) 3,000개를 제거한 10,000개의 표본을 분석에 사용하였다. β의 완전 조건부 분포는 $\triangle^{-1}=\sum_{l=1}^{r}\sum_{t=1}^{T_l}\boldsymbol{X_{lt}}+\boldsymbol{I_p}/\delta_\beta^2$ 이고, $\chi=\sum_{l=1}^{r}\sum_{t=1}^{T_l}\boldsymbol{X_{lt}}'\sum_{\eta}^{-1}\boldsymbol{O_{lt}}$ 일 때, $\pi(\beta|...,z)\sim N(\triangle\chi,\triangle)$에 의해 얻어진다. σ_ϵ^2와 σ_η^2의 조건부분포는 아래와 같다.

$$\pi(1/\sigma_\epsilon^2|...,z)\sim$$
$$G\left(\frac{2}{N}+a,b+\frac{1}{2}\sum_{l=1}^{r}\sum_{t=1}^{T_l}(\boldsymbol{Z_{lt}}-\boldsymbol{O_{lt}})'(\boldsymbol{Z_{lt}}-\boldsymbol{O_{lt}})\right),$$
$$\pi(1/\sigma_\eta^2|...,z)\sim$$
$$G\left(\frac{2}{N}+a,b+\frac{1}{2}\sum_{l=1}^{r}\sum_{t=1}^{T_l}(\boldsymbol{O_{lt}}-\boldsymbol{X_{lt}}\beta)'(\boldsymbol{O_{lt}}-\boldsymbol{X_{lt}}\beta)\right).$$

또한, ϕ의 완전 조건부 분포는 아래와 같다.

$$\pi(\phi|\cdots,z)\propto\pi(\phi)\times|S_\eta|^{-\sum_{l=1}^{r}T_l/2}\times$$
$$\exp\left[-\frac{1}{2\sigma_\eta^2}\sum_{l=1}^{r}\sum_{t=1}^{T_l}(O_{lt}-X_{lt}\beta)'S_\eta^{-1}(O_{lt}-X_{lt}\beta)\right].$$

2.3. 사전분포

베이지안 모형에서는 적절한 사전 분포의 설정이 중요하다. 모수에 대한 사전정보가 없는 경우 사전분포는 우도함수에 비해 상대적으로 평평(flat)한 균등분포를 가정한다. 본 연구에서 β의 사전분포로 정규분포를, 시공간적 효과의 오차항 σ_ϵ^2와 랜덤효과의 오차항σ_η^2의 사전분포로 역감마 분포를 고려하였다. 공간모수 ϕ의 사전분포는 감마분포와 균등분포를 고려했으나 모수의 trace plot이 수렴하지 않았다. 이에 안정적인 ϕ값을 찾기 위해 기상청 기상자료로 농촌진흥청 기상자료를 예측한 값에 대한 RMSE를 최소화 하는 ϕ를 최적공간모수로 설정하였다.

2.4. 예측 방법

본 연구에서 종속변수는 일 평균기온이고, 독립변수는 위도, 경도, 해발고도이다. 위치 s'와 시간 t'에서 공간 예측은 종속변수 $Z_l(s',t')$에 대한 사후분포를 통해 진행된다. 종속변수 $Z_l(s_0,t')$에 대한 사후분포는 추정된 모수를 적분함으로써 얻어진다.

$$\pi(Z_l(s',t')|z) = \int \pi(Z_l(s',t')|O_l(s_0,t'),\sigma_\epsilon^2)\pi(O_l(s',t')|\theta,z^*)\pi(\theta|z)dO_l(s',t')d\theta\,dz^*$$

깁스샘플링을 이용하여 사후분포 $\pi(\theta,z^*|z)$로부터 랜덤샘플 $\theta^{(j)}$을 추출한다. 베이지안 크리깅을 적용하여 $O_l(s_1,t'),\cdots,O_l(s_n,t')$가 주어진 $O_l(s',t')$의 조건부 분포에서 샘플 $O_l^{(j)}(s',t')$을 얻는다. 베이지안 크리깅은 관심 있는 위치의 특성치를 알기 위해 이미 알고 있는 주위의 값들의 선형 조합으로 그 값을 예측하는 지구 통계학적 기법이다. 깁스샘플링의 마지막 단계에서 결측값 $Z_l^{(j)}(s',t'), j=1,...,J$을 예측한다. 깁스샘플링 예측값은 95% 신용구간(Credible Interval)과 중앙값 및 평균을 사용하여 요약한다(Bakar and Sahu, 2015).

2.5. 평가방법

모형 평가 척도로 평균제곱근편차(RMSE), 평균절대오차(MAE), 평균절대비율오차(MAPE), 편향(BIAS)을 이용하였다. 평가방법의 값이 작을수록 예측정확도가 높음을 나타낸다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2},$$

$$MAE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)},$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|(\hat{y}_i - y_i)/y_i|,\ BIAS = (\hat{y}_i - y_i)$$

여기서 n은 전체 관측치의 수이고, y_i는 i번째 관측값이며, $\hat{y}_i$은 i번째 예측값을 나타낸다.

3. 연구 자료 및 지역 특성

양파는 월동작물로 주산지가 전라남도, 경상남도, 경상북도로 남부지방에 집중율이 높다(Lee and Lee, 1995). 특히 전라남도가 2003년 이후 양파 재배면적의 50% 이상을 차지하고 있다(Yoon et al., 2014). 양파는 온대지방에서 잘 자라는데, 전라남도가 겨울철(12~2월) 일평균기온이 0℃ 이상이고, 여름철(7~8월) 일평균 기온이 25℃ 정도인 온대지방이기 때문이다(Yoon et al., 2014).

연구를 위해 양파의 주재배지인 전라남도에 설치 운영 중인 기상관측자료를 이용하였다. 기상청은 95개의 자동기상관측소(AWS, ASOS)를 운영하고 있고 농촌진흥청은 25개소의 농업 기상관측소를 운영하고 있다. 분석을 위해 수집된 자료의 기간은 2013년 1월 1일부터 2016년 12월 31일까지이다. 수집된 관측소의 위치는 Fig. 1.이다. 파란색으로 표시된 곳은 기상청의 기상관측소이며 빨간색으로 표시된 곳은 농촌진흥청의 기상관측소이다.

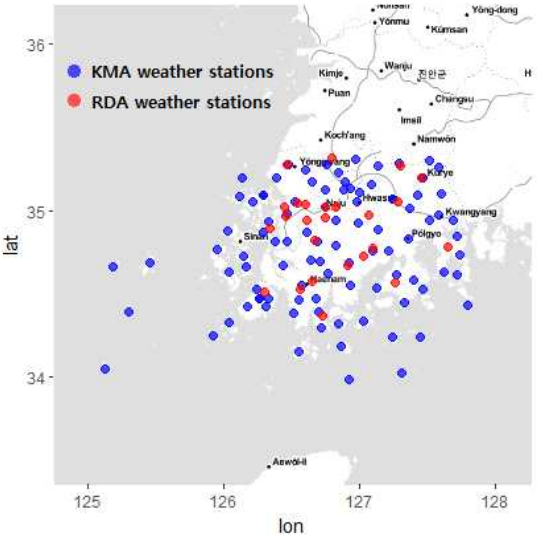

Fig. 1. Weather stations(KMA, RDA).

농업기상 정보를 분석한 결과 일시적인 기계결함이나 통신 장애로 인한 결측치가 포함되어 있다. 관측치는 365일*4년 동안 25개 기상관측소로부터 총 36,500개이며, 이 중 결측값이 3,639개로 나타났다.

전체 데이터에 대한 결측률은 9.97%이다(Table 1).

Table 1. Missing value status

Agricultural area	Observation (n)	Missing number	percentage (%)
onion	36,500	3,639	9.97

4. 자료 분석

결측치는 자료가 수집되지 않아 실제값을 확인하기 어렵다. 이에 본 연구에서는 기상청에서 운영하는 기상자료로 농업기상관측소의 기상정보를 예측하였다. 농업기상관측소는 관측된 기상정보가 있으므로 예측값과 관측값 사이의 차이를 통해 베이지안 시공간모형이 결측치 보정에 적절한지 평가할 수 있다. 본 연구에서는 베이지안 시공간모형에서 고려된 공변량으로 위도, 경도(위도+경도) 그리고 위도, 경도, 해발고도(위도+경도+고도)가 고려된 2개의 모형을 고려하였다.

베이지안 시공간모형의 모수추정을 위한 사전분포에 따른 모수들의 trace plot은 Fig. 2~Fig. 3이다. Fig. 2와 Fig. 3에는 각 모델에 대한 독립변수와 시공간효과의 표준편차 σ_ϵ^2와 통계적 랜덤효과 σ_η^2, 그리고 공간모수 ϕ에 대한 trace plot이다. 분석결과 공간을 표현하는 모수 ϕ는 값이 수렴하지 않은 문제점이 있다.

베이지안 시공간 모형에서 공간모수 ϕ는 다른 모수들의 추정치에 영향을 미친다. Trace plot에서 확인되듯이 ϕ가 수렴되지 않으므로 ϕ를 어떻게 추정할지에 관한 문제가 발생한다. 본 연구에서는 ϕ값을 0.01에서 0.05까지 0.001씩 증가시키며 농업기상정보의 관측값과 예측값의 RMSE를 계산하였다. RMSE는 평균 제곱근 편차로 예측값과 관찰값의 차이를 다룰 때 쓰는 측도로 값이 낮을수록 좋다. 그 결과 위도+경도 모형에서는 ϕ가 0.041일 때, 위도+경도+고도 모형에서는 ϕ가 0.039 일 때 RMSE 값이 가장 낮았다(Fig. 4, Fig. 5).

가장 낮은 RMSE를 공간모수 ϕ의 추정치로 사용하여 trace plot을 그리면 Fig. 6과 Fig. 7이다. 공간모수

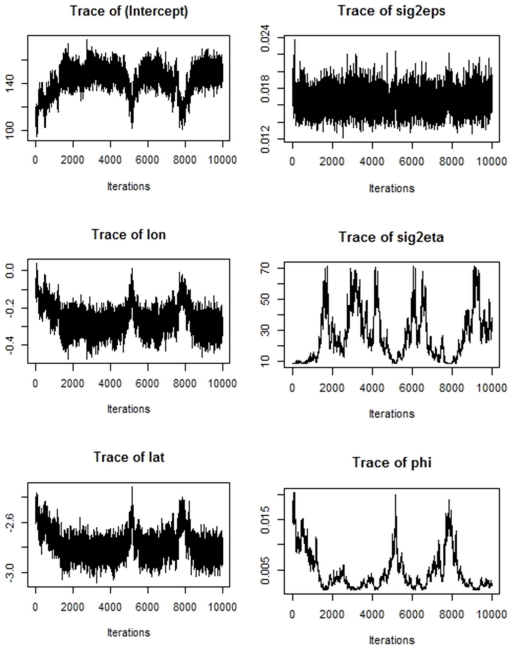

Fig. 2. Trace plot about parameters of prior distribution in lon+lat model.

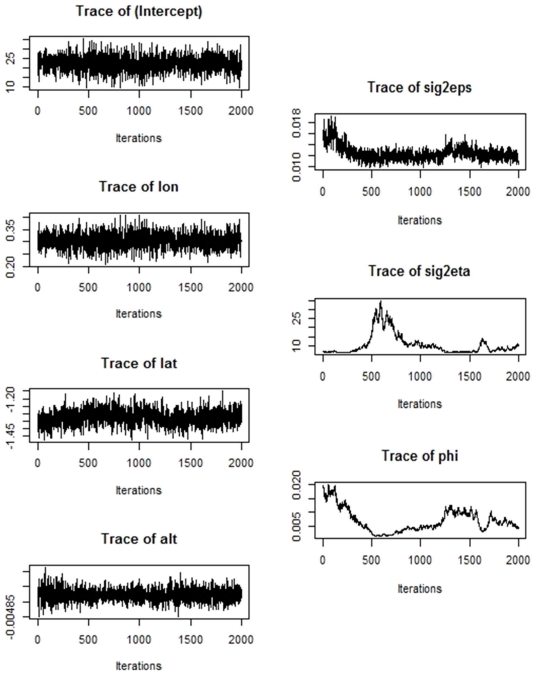

Fig. 3. Trace plot about parameters of prior distribution in lon+lat+alt model.

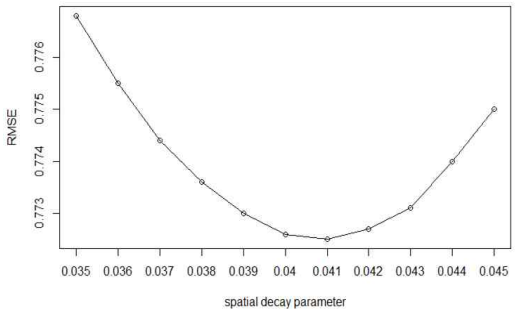

Fig. 4. Select optimal spatial decay ϕ (lon+lat model).

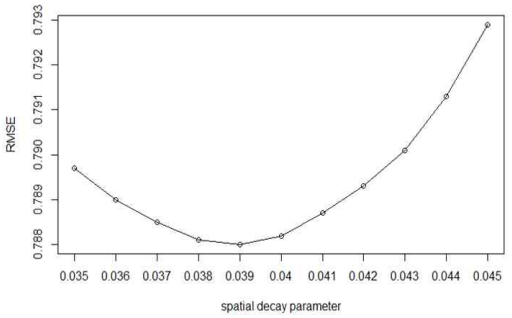

Fig. 5. Select optimal spatial decay ϕ (lon+lat+alt model).

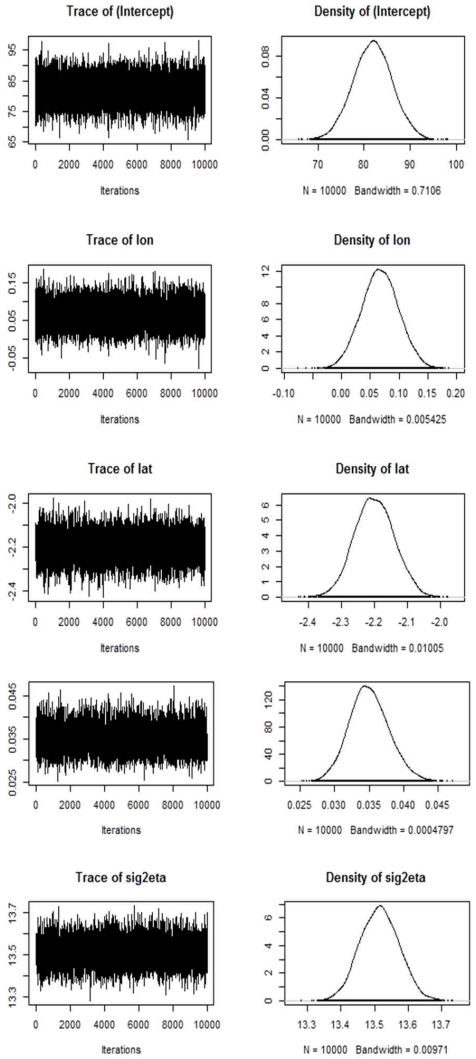

Fig. 6. trace plot of lon+lat model(ϕ=0.041).

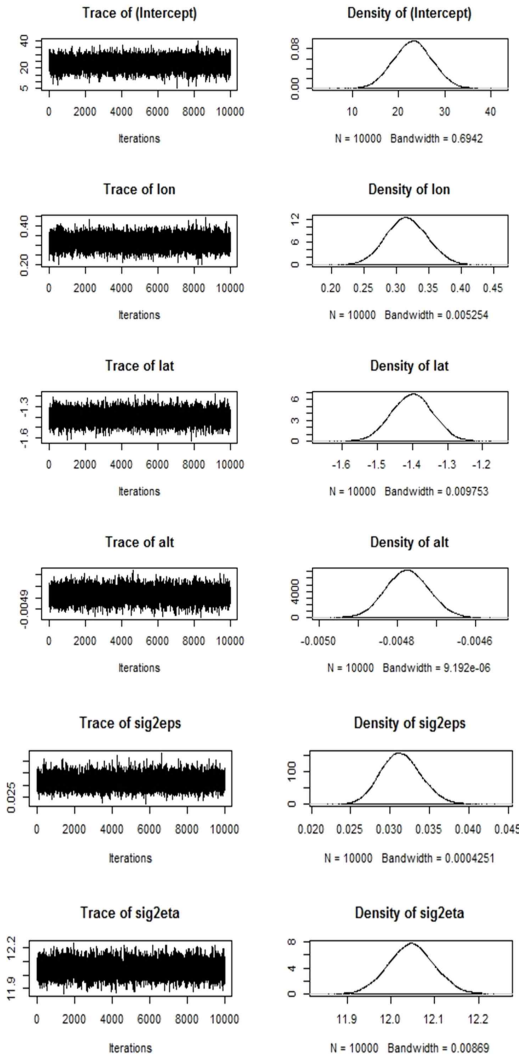

Fig. 7. trace plot of lon+lat+alt model(ϕ=0.039).

ϕ가 고정되었을 때, 다른 모수들의 trace plot의 연쇄는 Fig. 2와 Fig. 3과 달리 임의적으로 변동하며 정규분포에 근사하므로 수렴한다고 보인다.

최적 공간모수로 고정하여, 각 모델의 모수를 추정하면 Table 2와 Table 3이다. 시공간모형이므로 공간성이 반영되어 관측소 인근지역은 관측소에서 관측된 값과 유사한 값을 보인다. Table 2의 위도와 경도만 고려된 베이지안 시공간모형의 모수추정 결과를 살펴보면, 랜덤효과인 σ_ϵ^2는 0.035이고 시공간효과인 σ_η^2은 13.515이다. 시공간효과의 분산이 랜덤효과 보다 크므로 시공간 모형에 의해 자료가 잘 설명되고 있다. 공간적 상관성인 ϕ는 0.041로 0에 가까우므로 공간적 상관성은 높다. 위도와 경도의 추정된 계수를 통해 관측소에서 멀리 떨어질수록 위도 1도 증가하면 일 평균기온이 2.2도 감소하고 경도가 1도 증가하면 일 평균기온이 0.067씩 증가한다.

Table 2. Estimated parameter (lon+alt model)

	mean	median	std.	95% Credible Interval	
				Lower	Upper
Int.	81.846	81.879	4.230	73.554	90.218
lat.	-2.200	-2.201	0.060	-2.316	-2.084
lon.	0.067	0.067	0.032	0.004	0.131
σ_ϵ^2	0.035	0.035	0.003	0.030	0.041
σ_η^2	13.515	13.515	0.058	13.403	13.628
ϕ	0.041	0.041	0.000	0.041	0.041

Table 3. Estimated parameter (lon+alt+alt model)

	mean	median	std.	95% Credible Interval	
				Lower	Upper
Int.	23.042	23.055	4.133	15.104	31.052
lat.	-1.401	-1.400	0.058	-1.513	-1.289
lon.	0.316	0.315	0.031	0.255	0.378
alt.	-0.005	-0.005	0.000	-0.005	-0.005
σ_ϵ^2	0.031	0.031	0.003	0.027	0.037
σ_η^2	12.047	12.047	0.052	11.945	12.151
ϕ	0.039	0.039	0.000	0.039	0.039

위도와 경도 외에 해발고도까지 고려된 모형에서 추정된 모수는 Table 3이다. 위도와 경도만 고려된 모형과 마찬가지로 랜덤효과 (σ_ϵ^2=0.031)가 시공간효과 (σ_η^2=12.047)보다 작으므로 시공간모형의 의미가 있었다. 95% 신용구간을 통해 위도, 경도 그리고 고도 모두 통계적으로 유의미한 결과가 얻어졌다.

기상청에서 얻은 일 평균기온으로 농업기상관측소의 일 평균기온을 예측하여 베이지안 시공간 모형의 예측성능을 평가한 결과는 Table 4이다. 고도가 높아질수록 온도가 내려가는 물리적 특성 때문에 고도를 고려한 모델이 예측성능이 더 좋을 것으로 기대했으나, 위도와 경도만 고려한 모델이 RMSE, MAE, MAPE 기준으로 좋았다.

Table 4. The result of evaluation

	RMSE	MAE	MAPE	BIAS
lon+lat	1.184	0.769	7.757	0.222
lon+lat+alt	1.187	0.785	7.840	0.157

기온의 공간적 분포는 계절에 영향을 받을 수 있다. 계절효과를 살펴보기 위해 최적 모형인 위도와 경도 모형의 계절별 예측성능을 살펴보면 Table 5이다. 전체 모델의 RMSE가 1.184와 비교하면 가을이 봄, 여름, 겨울에 비해 베이지안 시공간모형의 예측성능이 부족하였다. RMSE, MAE, MAPE, BIAS를 종합하면 여름철 일평균 자료가 베이지안 시공간 모형으로 가장 잘 설명된다.

Table 5. The result of evaluation by season

	RMSE	MAE	MAPE	BIAS
spring	1.040	0.761	10.115	0.251
summer	1.040	0.573	2.369	0.028
fall	1.410	0.841	7.164	0.282
winter	1.112	0.814	11.319	0.336

계절 외에 베이지안 시공간 모형의 예측성능을 공간적으로 살펴보기 위한 지역별 RMSE 결과는 Table 6이다. 보성군에 위치한 지역의 계절별 RMSE 평균이 1.7이상으로 다른 지역에 비해 예측정확도가 가장

낮았다. 이외에 베이지안 시공간모형으로 일 평균기온이 잘 설명되지 않은 지역은 여수(1.39), 구례(1.27), 무안(1.25) 순이다.

Table 6. RMSE by station and month

station	spring	summer	fall	winter	mean
Jindo	1.14	1.10	1.30	1.04	1.14
Yeosu	2.04	0.90	1.27	1.36	1.39
Muan	1.19	2.44	0.58	0.79	1.25
Mua	0.72	1.26	1.92	0.86	1.19
Yeonggwang	0.95	0.51	0.90	0.83	0.80
Hampyeong	0.93	0.75	0.74	0.76	0.80
Haenam	0.55	0.34	0.81	0.67	0.59
Naju	1.02	1.09	0.86	0.87	0.96
Naju	0.67	0.50	0.87	0.79	0.71
Haenam	0.82	0.84	0.71	1.62	1.00
Yeongam	0.65	0.48	0.88	1.00	0.76
Wando	0.88	0.60	0.74	1.56	0.94
Naju	0.62	0.76	1.41	0.70	0.87
Naju	0.62	0.59	0.46	0.57	0.56
Jangseong	0.91	0.58	0.86	0.91	0.81
Naju	0.66	0.82	0.96	1.27	0.93
Jangheung	0.55	0.37	0.61	0.65	0.55
Boseong	1.60	1.62	2.18	1.43	1.71
Hwasun	0.73	0.59	0.91	0.95	0.80
Boseong	1.04	0.88	3.63	1.39	1.73
Goheung	0.53	0.37	0.61	0.63	0.53
Suncheon	0.76	0.59	1.17	1.21	0.93
Gokseong	0.72	0.68	1.02	1.22	0.91
Gurye	1.20	1.62	1.00	1.27	1.27
Sinan	1.78	0.76	0.90	1.13	1.14

Table 6의 결과를 공간적으로 살피면 Fig. 8이다. 보성군은 주변 지역보다 표고가 높아 일교차가 크며 해양성 기후의 영향을 받아 기후 변화가 다른 지역에 비해 상대적으로 높다. 무안군도 해안지역으로 해양성 기후의 영향을 받는다. 지역적 기후 특성이 강한 지역은 베이지안 시공간 모형으로 예측하는데 어려움이 있다는 사실을 알 수 있다. 예측정확도가 높은 지역은 고흥군 풍양면, 장흥군 장흥읍, 나주시 금천면, 해남군 삼산면 순으로 기상청 기상관측소와의 거리는 5.91 km, 2.16 km, 0.38 km, 2.96 km 순이다. 즉, 베이지안 시공간모형의 특성상 기상관측소와 농업기상관측소 간 거리가 멀지않으면 예측 정확도가 높았다.

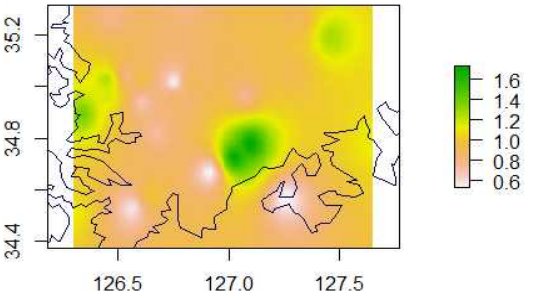

Fig. 8. Map of RMSE by stations.

마지막으로 기상청과 농업기상관측소에서 얻어진 모든 일평균자료를 활용한 베이지안 시공간모형으로 전라남도 보성군 웅치면의 일 평균기온의 결측치를 보정하였다(Table 7).

Table 7. The examples of replacement missing values

date	median
2013-03-27	10.058
2013-03-28	10.028
2013-03-29	7.660
2013-03-30	8.186
2013-03-31	8.255
2013-04-01	9.930
2014-08-21	24.037
2014-08-22	24.086
2014-08-23	23.790
2014-08-24	23.348
2014-08-25	24.648
2014-12-26	0.167
2015-02-20	4.940
⋮	⋮

5. 결 론

기상은 시공간적 특성을 내포하고 있어 계절(시간)과 지역(공간)에 따라 큰 편차를 보인다. 기존의 연구는

기상자료의 결측치 보정을 위해 평균치와 조건부 기대값을 최대화시키는 값을 이용하였다. 본 연구에서는 기상자료의 시공간적 특성을 반영한 베이지안 시공간 모형을 이용하여 결측치를 보정하는 방법을 다루었다. 최적 시공간모형을 선정하기 위해 기상청 기상관측소에서 관측된 일평균기온으로 농업기상관측소의 일평균기온을 예측하여 RMSE, MAE, MAPE와 BIAS로 평가하였다. 이 후에는 기상청과 농업기상관측소에서 관측된 모든 일평균자료로 베이지안 시공간 모형을 세워 농업기상관측소의 결측치를 보정하였다.

베이지안 시공간모형을 세우기 위해선 사전분포가 중요하다. 사전분포 중 공간모수 ϕ의 trace plot이 수렴하지 못하는 문제점이 제기되었다. RMSE를 기준으로 ϕ를 조금씩 증가시켜 최적 공간모수 값을 추정하여 베이지안 시공간모형을 세웠다. 분석결과 시공간효과 모수가 랜덤효과 모수보다 크므로 시공간모형으로 연구 자료를 잘 설명됨을 확인하였다. 하지만 지형의 특성이 강해 날씨의 변화가 심한 지역은 시공간모형으로 잘 설명되지 못한다는 한계점도 존재한다.

본 연구는 기상의 시공간적 특성을 반영한 베이지안 시공간 모형으로 농업기상 결측치를 보정해보자는 방법론적인 논지에서 시작 된 연구이다. 따라서 평균치 보정, EM 알고리즘을 이용한 보정 등의 다른 결측치 보정 방법과 결과를 비교분석은 수행되지 않았다. 또한 위도, 경도 그리고 해발고도외에 일평균기온에 영향을 미치는 외부 요인을 반영하지 못한 한계점이 존재한다. 따라서 향후 선행 결측치 보정 방법과의 비교분석과 농업기상을 위한 중요 요인을 파악하기 위한 추가 연구가 필요하다.

REFERENCES

Bakar, K. S., Sahu, S. K., 2015, spTimer: Spatio-temporal Bayesian modelling using R, J. Stat. Softw, 63, 1-32.

Banerjee, S., Carlin, B. P., Gelfand, A. E., 2014, Hierarchical modeling and analysis for spatial data, Crc Press.

Baraldi, A. N., Enders, C. K., 2010, An Introduction to modern missing data analyses, J. Sch. Psychol., 48(1): 5-37.

Cressie, N., 1994, An Approach to statistical spatial -temporal modeling of meteorological elds: Comment, J. Am. Stat. Assoc, 89, 379-382.

Gelfand, A. E., Banerjee, S., Gamerman, D., 2005, Spatial process modeling for univariate and multivariate dynamic spatial data, Environmetrics, 16, 465-479.

Gelfand, A. E., Smith, A. F. M., 1990, Sampling-based approaches to calculating marginal densities, J. Am. Stat. Assoc, 85, 398-409.

Jang, H. I., Seo, H. H., Park, S. J., 2002, Strategy for fruit cultivation research under the changing climate, Korean J. Hort. Sci. Technol, 20, 270-275.

Ko, K., Tak, H., 2016, The treatment of missing values using the integrated multiple imputation and callback method, Korean Journal of Pol. Stud., 54(4), 291-319.

Lee, B. L., 2000, Prospects on agrometeorological information for agricultural applications. Korean J. Agric. For. Meteorol., 2(1), 24-30.

Lee, H. J., Han, H. S., Chon, S. U., Kim, D. K., Kwon, H., Lee, K., 2014, Physiological characteristics and yield of onion affected by rapid temperature changes, Korean J. Environ Agric, 33(4), 364-371.

Lee, J., Lee, Y., 1995, Determinant Factors of Planted Area and Crop Situation of Red Pepper, Garlic, and Onions, Korea Rural Econ. INST.

Lee, K. K., Ko, K. K., Lee, J. W., 2012, Correlation analysis between meteorological factors and crop products, J. Environ. Sci. Int., 21(4), 461-470.

Lee, S., Kim, D., 2010, The comparison of imputation methods in space time series data with missing values, Commun Stat Appl Methods, 17, 263-273.

Min, J. S., Lee, M. H., Jee, J. B., Jang, M., 2016, A Study of the method for estimating the missing data from weather measurement instruments. J. Digital Convergence, 14, 245-252.

Yoon, D. K., Oh, S. Y., Nam, K. W., Eom, K. C., Jung, P. K., 2014, Changes of cultivation areas and major disease for spicy vegetables by the change of meteorological factors, J. Climate Change Res, 5(1), 47-59.

Yoon, S., Kim, M., 2016, Spatio-temporal models for generating a map of high resolution NO_2 level, J. Korea Data Info. Sci, 27(3), 803-814.