Estimation of ozone concentration using Hierarchical Bayesian Spatio-Temporal model

Dain, Park (Master's Degree, Department of Statistics, Daegu University)

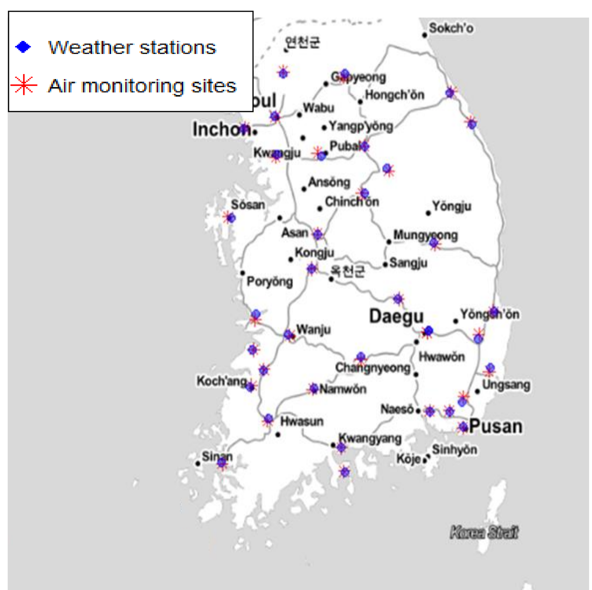
Sanghoo, Yoon(Assistant professor, Division of Mathematics and big data science, Daegu University)

Purpose

- **Ozone** causes **bronchial disease** and **lung function decrease** when it enters the human body(Tilton, 1989)
- Ozone concentration is closely related to **weather conditions** such as **temperature, wind speed, relative humidity, and sunlight**(Vukovich, 1995).
- The Korea has distinct **local weather conditions** and a large number of **pollutant emission sources** are **scattered**(Oh and Kim, 2002).
- In this study, we estimated the **ozone concentration levels using Bayesian spatio-temporal model.**

Research flow

Dependent variable: 8 hours ozone concentration levels
Independent variable: maximum temperature, wind speed, relative humidity, sunlight



2014~2016 Daily ozone data *(Air Korea)*	2014~2016 **Meteorological information** *(Korea Meteorological Administration, ASOS)*

Hierarchical Bayesian spatio-temporal model
- **Independent Gaussian process**
- **Autoregressive models**

- **Spatial heterogeneity occurs** because weather stations(ASOS) are located **far** from air monitoring sites
- **33 stations** are located within **5Km** from air monitoring sites.
- So that, that stations are **selected** for Hierarchical Bayesian Spatio-temporal model.

Hierarchical Bayesian Spatio-Temporal Model

- **Independent Gaussian process model(GP model)**

$\boldsymbol{Z_{lt} = O_{lt} + \epsilon_{lt}}$: True underlying Processes
$\boldsymbol{O_{lt} = X_{lt}\beta + \eta_{lt}}$: Spatio-temporal effect

Assume that ϵ_{lt} and η_{lt} follow **independence** and **normality**.

Covariance matrix= $\boldsymbol{\sigma_\eta^{\ 2}\ S_\eta}$
$\boldsymbol{\sigma_\eta^{\ 2}}$= Variance that does not change with space
$\boldsymbol{S_\eta}$= Assume exponential correlation function as spatial correlation matrix.
ϕ controls the **spatial decay rate** over distance.
Let $\boldsymbol{O}$ denote all random effects, and $\boldsymbol{O_{lt}}$ and $\boldsymbol{\theta = (\beta,\ \sigma_\epsilon^{\ 2},\ \sigma_\eta^{\ 2},\ \phi)}$ denote all the parameters of this model.
The logarithm of posterior distribution is defined by Bakar and Sahu(2015).

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

The temporal effect is independent in GP model.
The spatial correlation is equally applied at each time by ϕ.

- **Autoregressive model(AR model)**

$\boldsymbol{Z_{lt} = O_{lt} + \epsilon_{lt}}$: True underlying Processes
$\boldsymbol{O_{lt} = \rho O_{lt-1} + X_{lt}\beta + \eta_{lt}}$: Spatio-temporal effect

$\boldsymbol{\rho}$ is the **autocorrelation parameter** between (-1,1) (if $\boldsymbol{\rho}$=0 then GP model)

$\boldsymbol{\mu_l}$ and $\boldsymbol{\sigma_l^{\ 2}}$ are the mean and standard deviation of O_{l0}.
Let $\boldsymbol{O}$ denote all random effects, and $\boldsymbol{O_{lt}}$ and $\boldsymbol{\theta = (\beta, \rho, \sigma_\epsilon^{\ 2},\ \sigma_\eta^{\ 2},\ \phi, \mu_1, \sigma_l^{\ 2})}$ denote all the parameters of this model.
The logarithm of posterior distribution is defined by Sahu and Mardia (2005)

$$log\pi(\theta, \boldsymbol{O}, z^*|z) \propto -\frac{N}{2} log\ \sigma_\epsilon^{\ 2} - \frac{1}{2\sigma_\epsilon^{\ 2}}\sum_{l=1}^{r}\sum_{t=1}^{T_l}(Z_{lt}-O_{lt})'(Z_{lt}-O_{lt}) - \frac{\sum_{l=1}^{r} T_l}{2}\log|\ \sigma_\eta^{\ 2} S_\eta|$$
$$-\frac{1}{2\sigma^2{}_\eta}\sum_{l=1}^{r}\sum_{t=1}^{T_l}(O_{lt}-\rho O_{lt-1}-X_{lt}\beta)'S_\eta^{\ -1}(O_{lt}-\rho O_{lt-1}-X_{lt}\beta)$$
$$-\frac{1}{2}\sum_{l=1}^{r}\log|\ \sigma_l^{\ 2}S_0| - \frac{1}{2}\sum_{l=1}^{r}\frac{1}{\sigma_l^{\ 2}}(O_{l0}-\mu_l)'S_0^{\ -1}(O_{l0}-\mu_1) + log\pi(\theta)$$

The **full conditional distribution of parameters** is obtained by **Gibbs sampling**(Sahu and Baker, 2012)

- **Prior distribution**

The **prior distribution** was referred from the **previous study**(Sahu and Baker(2012), Yoon and Kim(2016)).

$\boldsymbol{\beta, \rho, \mu_l}$: Normal distribution ~ $N(0,10^4)$
$\boldsymbol{\sigma_\epsilon^{\ 2}, \sigma_\eta^{\ 2}, \sigma_l^{\ 2}}$: Inverse Gamma ~ $IG(2,1)$
$\boldsymbol{\phi}$: Gamma, Uniform, Normal

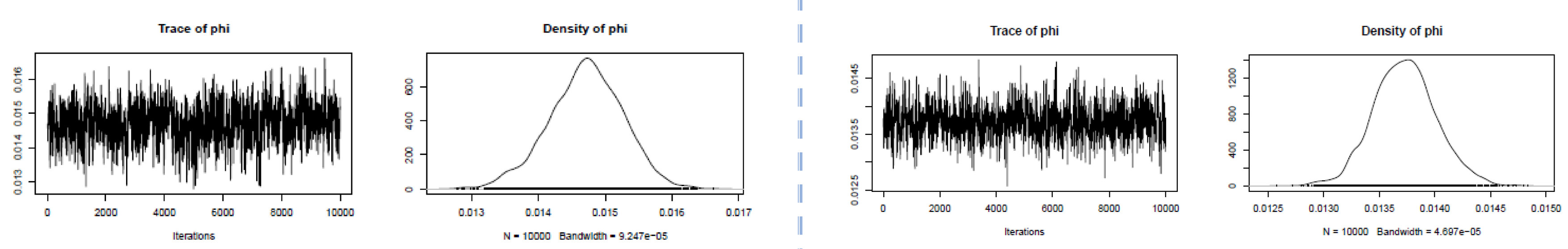
- **Validation methods**

- RMSE = $\sqrt{\frac{1}{n}\sum_1^n(\hat{y}_i - y_i)^2}$
- MAE = $\sqrt{\frac{1}{n}\sum_1^n|\hat{y}_i - y_i|}$
- MAPE=$\frac{100}{n}\sum_1^n|\hat{y}_i - y_i/\ y_i|$
- BIAS=$\hat{y}_i - y_i$

Result

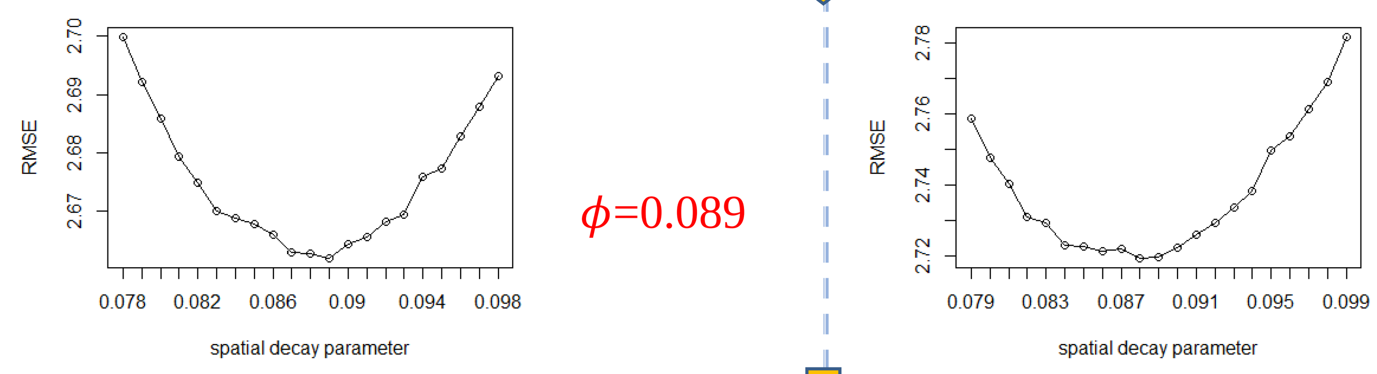
GP model | **AR model**

- **Trace plot of spatial decay parameter ϕ**



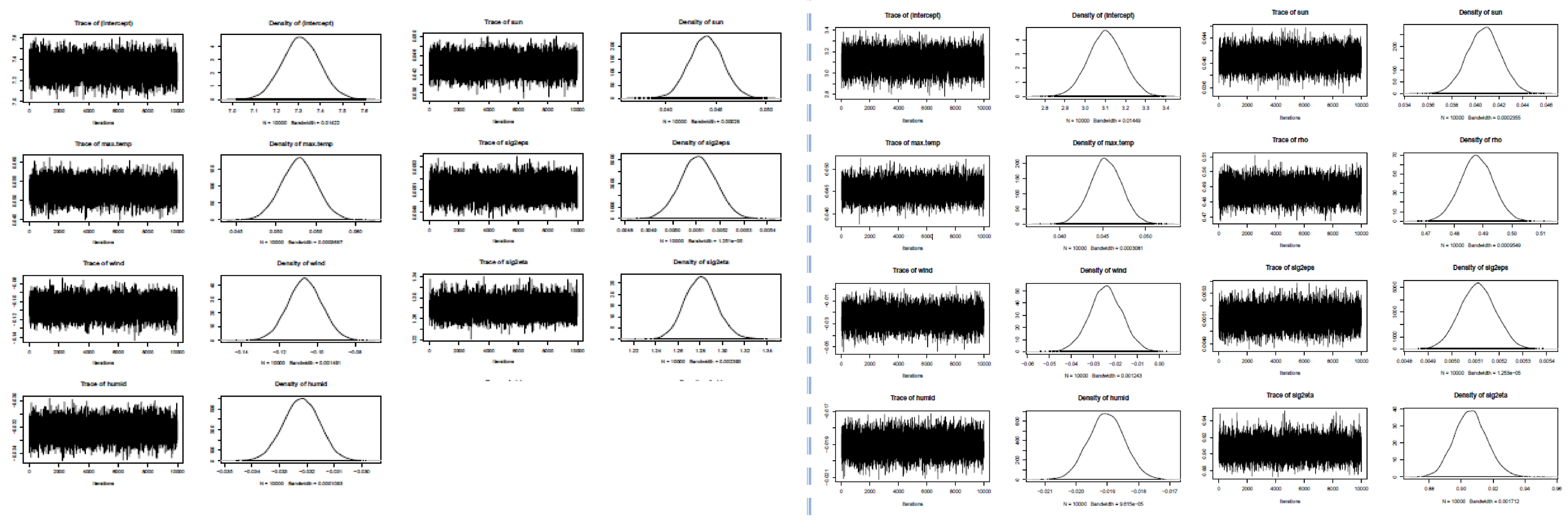
Spatial decay parameter ϕ is not converged.

- **Select optimal spatial decay parameter ϕ**



ϕ=0.089 | ϕ=0.088

- **Trace plot after spatial decay ϕ fix**



Coefficients were normally converged.

- **The result of parameter estimation**

GP model	Mean	Median	SD	Credible Interval	
				Low2.5p	Up97.5p
(Intercept)	7.312	7.312	0.085	7.148	7.480
Max temp	0.053	0.053	0.002	0.049	0.057
Wind speed	-0.107	-0.107	0.009	-0.124	-0.089
Humidity	-0.032	-0.032	0.001	-0.034	-0.031
Sunlight	0.044	0.044	0.002	0.041	0.047
$\sigma_\epsilon^{\ 2}$	0.005	0.005	0.000	0.005	0.005
$\sigma_\eta^{\ 2}$	1.281	1.280	0.014	1.254	1.310

AR model	Mean	Median	SD	Credible Interval	
				Low2.5p	Up97.5p
(Intercept)	3.104	3.103	0.086	2.935	3.274
Max temp	0.045	0.045	0.002	0.042	0.049
Wind speed	-0.024	-0.024	0.008	-0.039	-0.010
Humidity	-0.019	-0.019	0.001	-0.020	-0.018
Sunlight	0.041	0.041	0.001	0.038	0.043
ρ	0.488	0.488	0.006	0.477	0.499
$\sigma_\epsilon^{\ 2}$	0.005	0.005	0.000	0.005	0.005
$\sigma_\eta^{\ 2}$	0.906	0.906	0.010	0.887	0.927

$\boldsymbol{O_{l0}}$		Mean	SD	Credible Interval	
				Low	up
$\mu_l^{\ 2}$	2014	8.268	0.815	7.314	8.913
	2015	8.312	0.797	7.391	8.945
	2016	8.314	0.841	7.389	8.961
$\sigma_l^{\ 2}$	2014	0.059	0.015	0.049	0.068
	2015	0.059	0.015	0.049	0.067
	2016	0.059	0.015	0.049	0.067

- The **maximum temperature** and **sunlight** are **positively related** with ozone concentration.
- **Wind speed** and **relative humidity** have a **negative relationship**.
- **Spatio-temporal effects** $\sigma_\eta^{\ 2}$ are **larger** than statistical random effects $\sigma_\epsilon^{\ 2}$.

- **The result of model validation**

GP model:

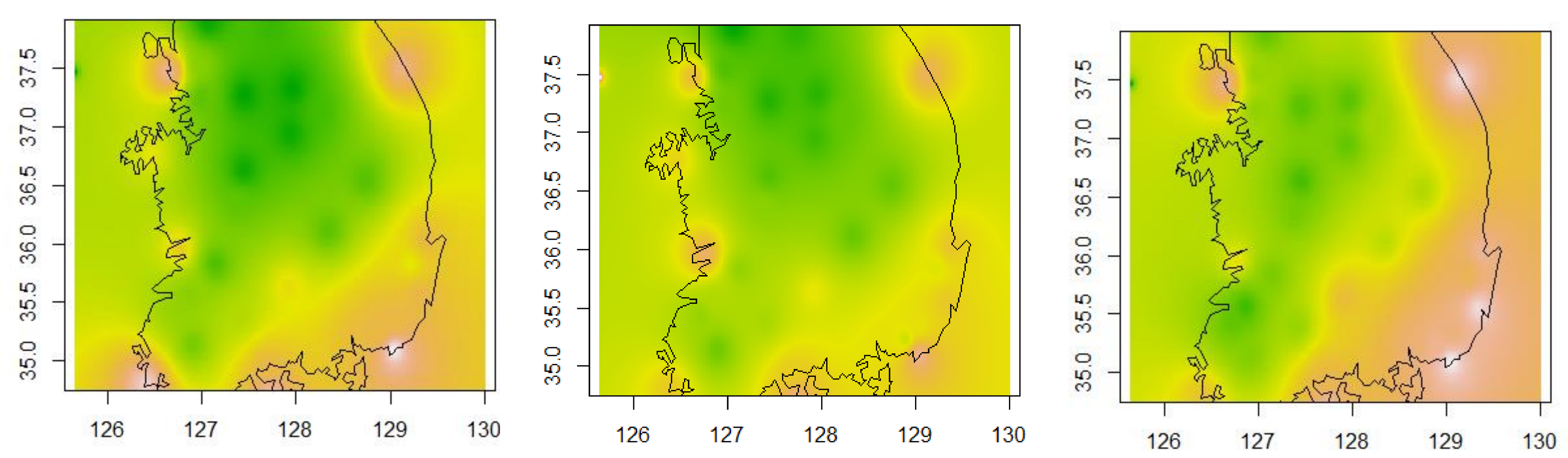
RMSE		MAE		MAPE		BIAS	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
15.334	1.107	**12.149**	0.828	**31.728**	2.480	-1.354	2.756

AR model:

RMSE		MAE		MAPE		BIAS	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
15.484	1.250	12.183	0.982	32.044	2.876	-0.964	3.064

- We **divided 33 meteorological stations and air monitoring sites** into 27 fitted sites and 6 validation site and it runs **100 times.**
- **Each validation model** was **fitted** and the **mean of validation scale** was **compared.**
- The **GP model was better** than AR model in the **sense of RMSE, MAE and MAPE**.

- **Model based interpolation of annual average ozone levels**



- **Interpolate annual average ozone concentration** predicted by 33 ASOS stations.
- It will be generated using about **130 ASOS** stations.(Further study)
- **Urban** and **inland areas** have **higher** ozone concentrations than **Coastal areas**.
- The **eastern and southern east areas** show that they have **lower** ozone concentration.
- Metropolitan cities(**Seoul, Daegu**) and Satellite **region(Kyunggi**, **Gyeongbuk**) are located in **danger** of ozone concentration.