Source-specific Exposure Assessment by Using Bayesian Spatial Multivariate Receptor Modeling

Eun Sug Park

Texas A&M Transportation Institute

Acknowledgements

- **Collaborators**
  - Philip K. Hopke, PhD, University of Rochester
  - Marco Benedetti, PhD, Nationwide Children's Hospital
  - Jae Young Kim, PhD, Seoul National University
  - Young Su Lee, Seoul National University
  - Hyeri Jo, Seoul National University
  - Seungmuk Yi, PhD, Seoul National University
  - Ho Kim, PhD, Seoul National University
- **Funding**
  - Health Effects Institute (HEI) R-82811201
  - Seoul National University, KT, Korea Agency for Infrastructure Technology Advancement

Scientific Questions

- Can we identify major pollution sources based on ambient concentrations of pollutants measured from monitoring stations?
- Can we assess source-specific exposures at any locations (e.g., unmonitored sites or residential locations)?



Outline

- Background
- Multivariate Receptor Modeling
- Model Identifiability and Model Uncertainty
- Bayesian Spatial Multivariate Receptor Modeling
- Real Applications
  - Houston Volatile Organic Compounds (VOCs) data
  - Korea $PM_{2.5}$ speciation data
- Summary and Discussions

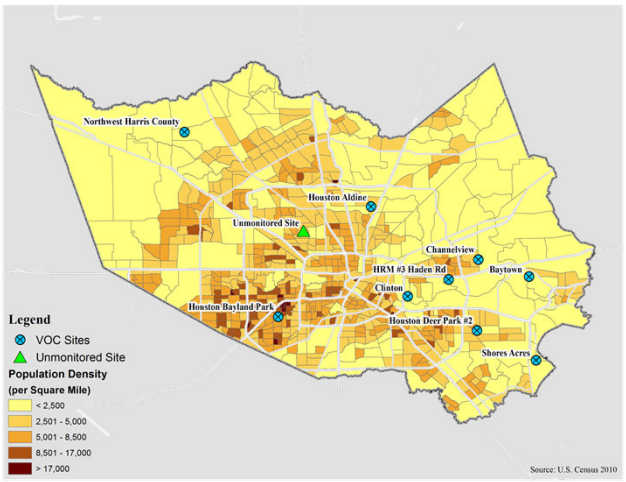
Background

- Increasing interest in assessing health effects of multiple pollutants.
- Quantify exposure to multiple pollutants by considering 'source-specific exposure'.
  - Identify important emission sources of "high risk" pollutants or harmful pollution sources.
  - More targeted enforcement strategies.
  - Source-specific exposures tend to be less highly correlated than pollutants.

Background

- Challenge: Source-specific exposures cannot be measured/observed directly.
  - Measured pollutants are mixture of contributions from several sources.
  - Source-specific exposures need to be estimated by source apportionment methods such as Multivariate Receptor Modeling.

Motivating Example: Houston Volatile Organic Compounds (VOCs) monitoring stations



- Growing availability of multipollutant data collected from multiple monitoring sites (multivariate space-time data).
- Source-specific exposures estimated from a single monitoring site are prone to the bias caused by spatial misalignment of health and exposure data.
- Method for incorporating spatial correlations in multisite multipollutant data into estimation of source-specific exposures is desired.

Multivariate Receptor Modeling

- A collection of methods used for source identification and apportionment.
  - Unfold the multivariate air pollution data into components associated with different sources.
  - Estimate the pollution source profiles (source composition profiles) and amounts of pollution (source contributions) based on the ambient concentrations of multiple air pollutants measured at the receptor.

Multivariate Receptor Modeling

Basic Assumption: Chemical Mass Balance

The total airborne particulate concentration (ng/m^3) measured at a site is the sum of contributions from independent source types. For example,

$$Zn_{Total} = Zn_{smelter} + Zn_{incinerator} + Zn_{auto} + \ldots$$

$$Zn_{smelter}\ (ng/m^3) = [\%\ Zn\ in\ smelter\ particulate\ emissions] \times [mass\ concentration\ (ng/m^3)\ of\ smelter\ particles\ in\ the\ atmosphere]$$
$$= P_{Zn,\ smelter}\, S_{smelter}\ (ng/m^3)$$

9

9

Multivariate Receptor Modeling

- **Basic Model:** $\mathbf{X} = \mathbf{SP} + \mathbf{E}$

where
X : $T \times J$ data matrix (concentrations of J pollutants measured over T days)
q : # of major pollution sources (often unknown)
P : $q \times J$ source composition profile matrix (**P** ≥ 0)
S : $T \times q$ source contribution matrix (**S** ≥ 0)
E : $T \times J$ error matrix

- Both **S** and **P** are unknown parameters to be estimated.
- **Nonnegative Factor Analysis Model**

10

10

Multivariate Receptor Modeling

- **Obstacles:**
  - Unknown number of sources (q).
  - Non-identifiability of model (without further constraining the parameters):

$$\mathbf{SP}=\mathbf{SRR}^{-1}\mathbf{P}=\mathbf{S}^{*}\mathbf{P}^{*}$$

for any $q \times q$ nonsingular matrix **R**.

- Factor indeterminacy/Rotational ambiguity
- Need additional constraints (called 'identifiability conditions') on the parameters to remove factor indeterminacy.

11

11

Identifiability Conditions Used in Multivariate Receptor Modeling

- **Constraints on P:**
  - Each source has a tracer element (too strong & often not satisfied).
  - Pre-specifications of zeros in **P**
    - Assume some pollutants are not contributed by a particular source. More general than the tracer element assumption.
- **Constraints on S:**
  - Pre-specifications of zeros in **S**, i.e., assume each source is missing on some days.
    - Park, Spiegelman, and Henry (2002).

12

12

Pre-specifications of Zeros in P

Conditions in confirmatory factor analysis (Anderson, 2003)

C1. There are at least q-1 zero elements in each row of **P**.

C2. For each k=1,…,q, the rank of $\mathbf{P}^{(k)}$ is q-1, where $\mathbf{P}^{(k)}$ is the matrix composed of the columns containing the assigned zeros in the kth row with those assigned zeros deleted.

C3. $\mathbf{P}_{kj}$=1 for some j (j=1,…,J) for each k (k=1,…,q) or $\sum_{j=1}^{J} \mathbf{P}_{kj} = 1.$

P	Species #								
Source #	1	2	3	4	5	6	7	8	9
1	0.1	0.05	0.25	0.1	**0**	**0**	0.3	0.1	0.1
2	**0**	0.4	**0**	0.1	0.1	0.05	0.1	0.05	0.2
3	0.1	**0**	0.05	**0**	0.1	0.4	0.05	0.2	0.1

13

13

Model Uncertainty in Multivariate Receptor Modeling

- Number of sources (q) is unknown.
- The identifiability conditions (e.g., pre-specified positions of zeros in **P**) that are appropriate for the data at hand is unknown.

→Each possible combination of q and identifiability conditions (e.g., pre-specified positions of zeros in **P** or **S**) defines a different model.

➢ A Bayesian approach (considering posterior probability of each model) to handle model uncertainty in multivariate receptor models was developed in Park, Oh, and Guttorp (2002).

14

14

Main Objectives

- Estimate source composition profiles (**P**) and source contributions (**S**) along with their uncertainties.
- Predict source contributions (source-specific exposures) at unmonitored sites and provide uncertainty estimates.
- Handle model uncertainty (unknown number of sources and rotational ambiguity) simultaneously.

15

15

Spatially Extended Bayesian Multivariate Receptor Modeling

- **Bayesian Spatial Multivariate Receptor Modeling (BSMRM)** -Park et al. (Technometrics, 2018)
  - Developed by adapting dynamic factor process convolution models (Calder, 2007).
  - Relaxes the assumption of the known number of sources and model identification conditions.
  - Incorporates spatial correlation in data into modeling.
  - Enables prediction of unobserved (latent) source contributions along with their uncertainty estimates at any locations.

16

16

Spatially Extended Multivariate Receptor Models Based on Discrete Process Convolution Models

$$\mathbf{X}(s_r,t)=\mathbf{K}(s_r)\mathbf{G_t}\mathbf{P}+\mathbf{E}(s_r,t),\quad t=1,\ldots,T$$

where

$\mathbf{X}(s_r,t)$: multivariate spatial-temporal process of air pollutants

s_r: spatial location of the rth receptor ($r=1,\ldots,N$)

$\mathbf{P}$: $q \times J$ source composition matrix ($\mathbf{P}>0$)

$\mathbf{G_t}$: q underlying latent processes defined at L spatial locations $\{w_1,\ldots,w_L\}$ on a coarse grid, $\mathbf{G_t} \sim N_q(\xi_0, \mathrm{I}_L, \Omega_0)\mathbf{I}(\mathbf{G_t} \geq \mathbf{0})$,

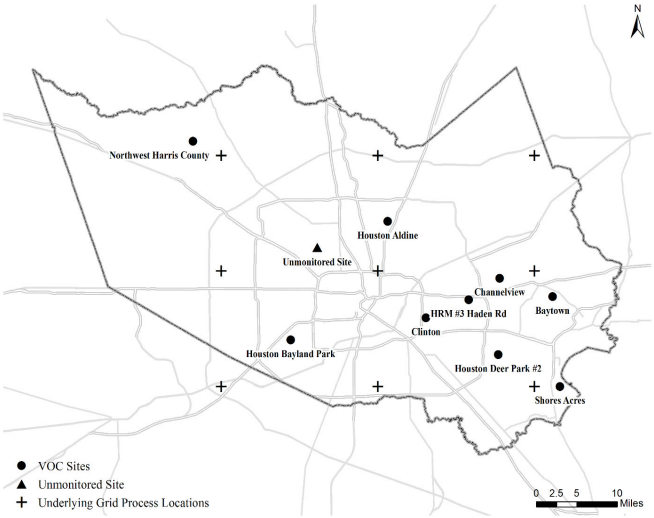
$\mathbf{K}(s_r)=[k(w_1-s_r),\cdots,k(w_L-s_r)]$, k :smoothing kernel,

$\mathbf{E}(s_r,t)$: iid, mean zero, Gaussian process on (s_r,t) with variance $\Sigma_J = diag(\sigma_1^2,\cdots,\sigma_J^2)$.

17

17

Example: Locations of VOC monitoring sites in Houston and underlying grid process locations (N=9, L=9)



18

18

Bayesian Spatial Multivariate Receptor Modeling with unknown *q* and/or identifiability conditions

- Consider a range of plausible models (built by exploratory data analysis or prior knowledge) instead of a single model and estimate model uncertainty by posterior model probabilities (or marginal likelihoods).
- Parameter estimation (for $\mathbf{G}$, $\mathbf{P}$, $\boldsymbol{\Sigma}$) and model uncertainty estimation (estimation of marginal likelihood for each model) can be performed simultaneously by Markov chain Monte Carlo (MCMC).
  - Implementation details: Park et al. (Technometrics, 2018)
  - An efficient computation method of Oh (1999) was utilized for marginal likelihood computation.

Spatial Prediction of Latent Source Contributions

- Source contributions at any location s can be predicted by

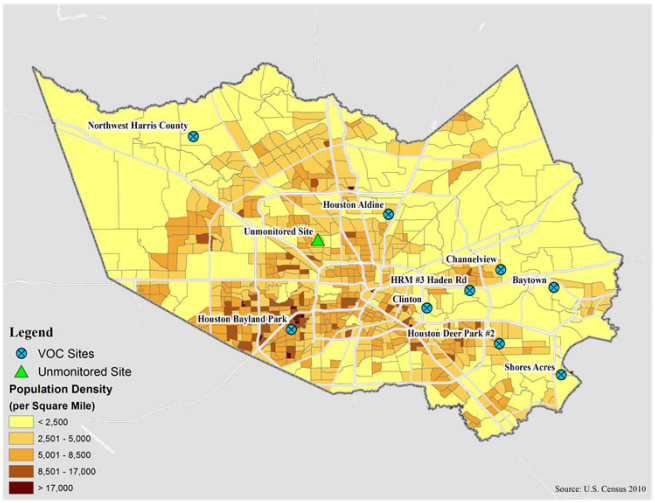
$$\mathbf{S}(s) = \mathbf{K}(s)\mathbf{G}_t$$

where

$$\mathbf{K}(s) = \left[\kappa(w_1 - s) \quad \kappa(w_2 - s) \quad \cdots \quad \kappa(w_L - s)\right].$$

Real Application I: Houston VOCs data

- 24-hour canister Volatile Organic Compounds (VOCs) data collected every 6 days from monitoring sites in Harris County during 2003-2005 and 2010-2012.



Real Application: Houston VOCs data analysis

- # of VOC species originally measured=100+.
- 2003-2005: # of monitoring sites (N)=9, # of days (T)=235.
- 2010-2012: N =12, T =198.
- Based on previous studies (e.g., Buzcu and Fraser, 2006), refineries, petrochemical production facilities, unburned gasoline, natural gas, vehicle exhaust, and aromatics were presumed to be potential candidate sources affecting the region.
- This prior knowledge was utilized in selecting the appropriate subset of species (17 species, i.e., J=17) for source apportionment and in building a set of candidate models to be compared.

Real Application: Houston VOCs data analysis

- Seven candidate models with q = 4, 5, 6, or 7 and different identifiability conditions (pre-specified zero elements in **P**) that vary with q were considered.
- A model with 5 sources resulted in the highest marginal likelihood among those seven models.

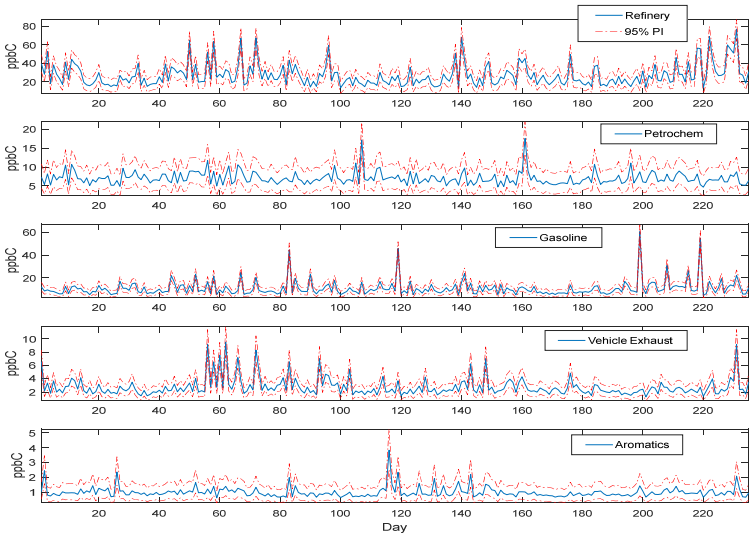
Real Application: Houston VOCs data analysis

- Estimated source composition profile matrix (**P**′)

Species #	Species name	Source 1 (Refinery)	Source 2 (Petrochem)	Source 3 (Gasoline)	Source 4 (Vehicle Exhaust)	Source 5 (Aromatics)
1	1,2,4-Trimethylbenzene	0	0	0.23	1.15	0.32
2	1,3-Butadiene	0.07	1.44	0.61	3.23	0
3	2,2,4-Trimethylpentane	0	0	1.43	0	0.24
4	Acetylene	0.30	0.09	0	55.85	0.78
5	Benzene	0.11	3.30	5.34	0.96	2.02
6	Ethane	45.29	0.72	0	2.24	4.94
7	Ethylbenzene	0.01	0.02	0.20	0.12	10.36
8	Ethylene	2.35	35.81	5.24	20.97	3.82
9	Isobutane	7.27	2.24	17.03	3.49	0
10	Isopentane	0.93	0.54	31.48	1.30	3.78
11	Propane	31.30	0.57	0.69	1.46	3.21
12	Propylene	0	54.11	0	4.57	0
13	Toluene	0.42	0.82	1.20	3.53	37.62
14	n-Butane	11.59	0	19.42	0	2.50
15	n-Hexane	0.32	0.28	3.35	0	0.55
16	n-Pentane	0.04	0.08	13.78	0	0
17	p-Xylene+m-Xylene	0	0	0	1.12	29.85

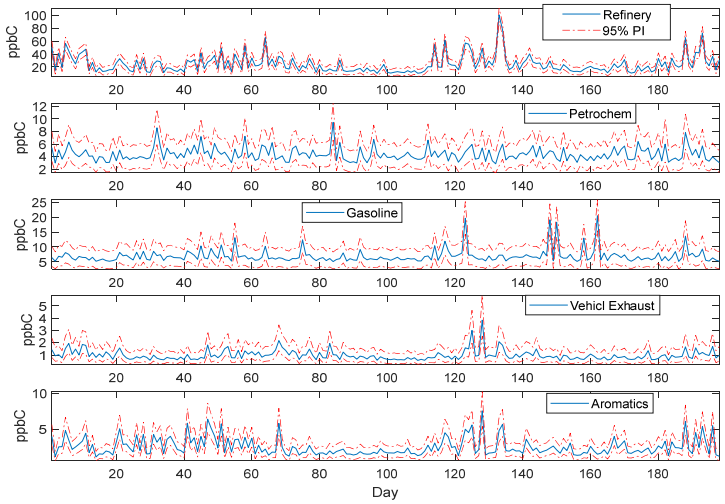
Real Application: Houston VOCs data

- **Predicted source contributions along with uncertainty estimates at an unmonitored site (2003-2005)**



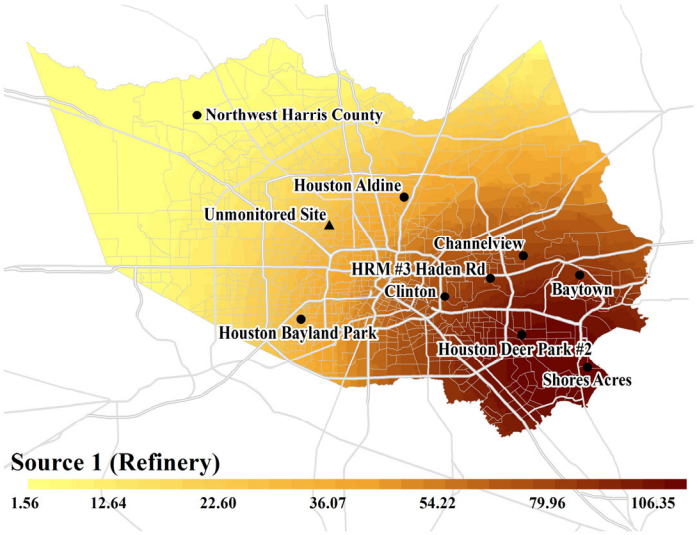
Real Application: Houston VOCs data

- **Predicted source contributions along with uncertainty estimates at an unmonitored site (2010-2012)**



Source contribution surface (Refinery)

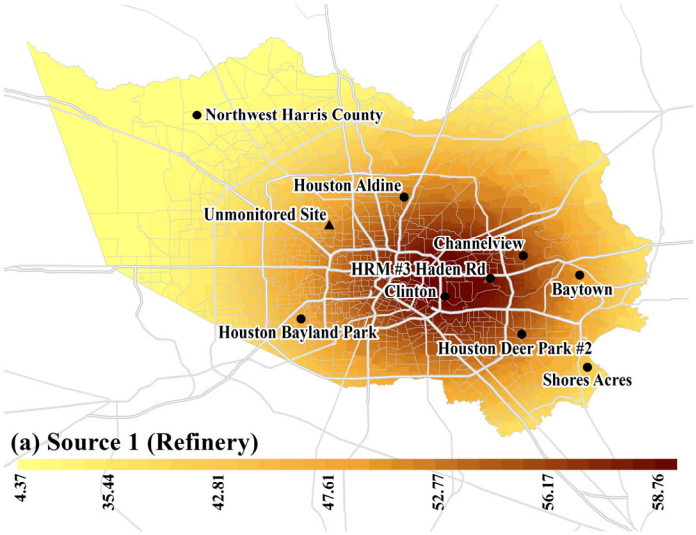
- Predicted refinery source contributions for Houston, TX on June 8, 2003



27

Source contribution surface (Refinery)

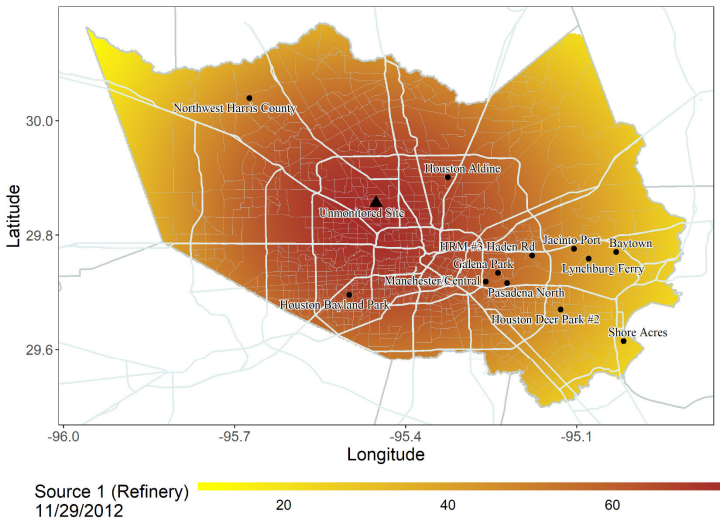
- Predicted refinery source contributions for Houston, TX on December 30, 2004



28

Source contribution surface (Refinery)

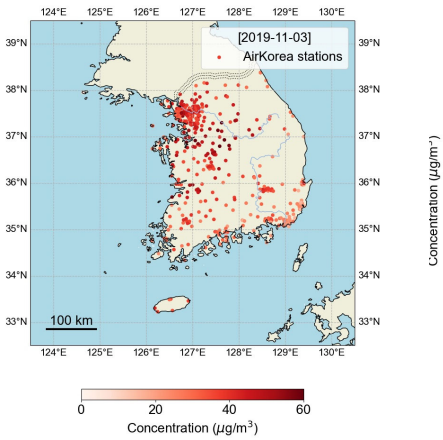
- Predicted **refinery** source contributions for Houston, TX, **on November 29, 2012**



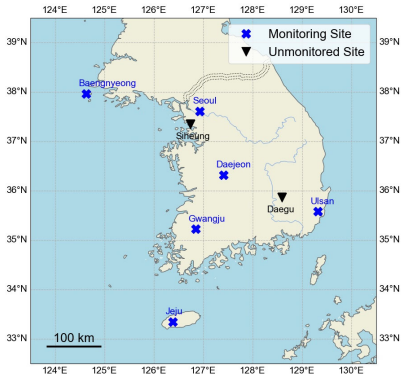
29

Real Application II: Korea $PM_{2.5}$ Speciation Data

$PM_{2.5}$ mass concentration stations



$PM_{2.5}$ speciation monitoring stations



30

Real Application: Korea $PM_{2.5}$ speciation data

- # of $PM_{2.5}$ speciation monitoring sites (M)=6
  - Seoul, Daejun, Gwangu, Ulsan, Jeju, Baengnyeong
- # of $PM_{2.5}$ species originally measured=28.

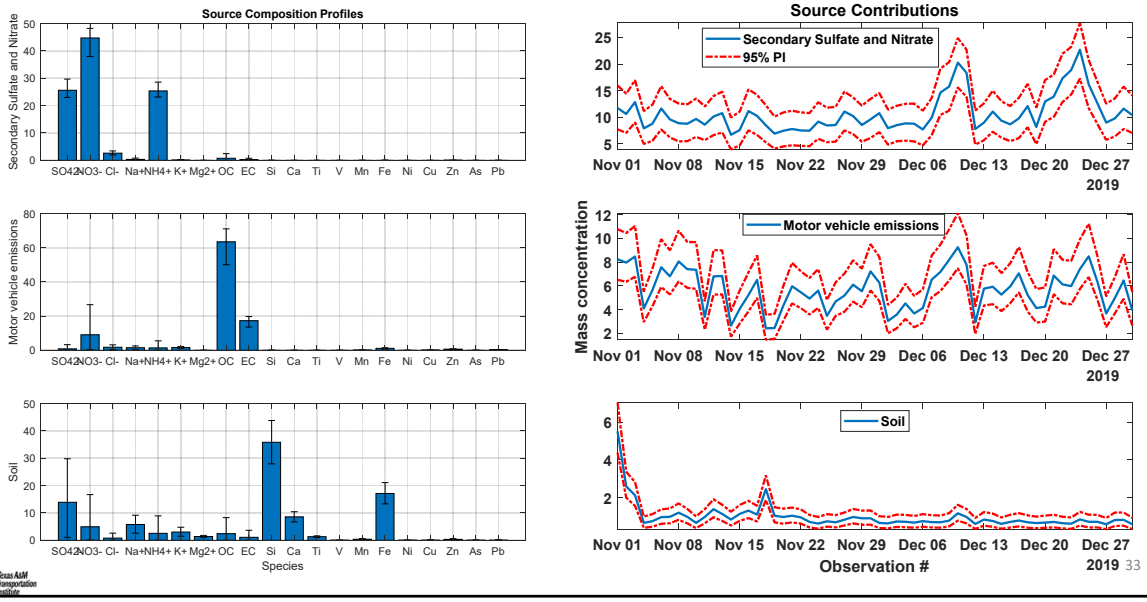
Chemical Species	$PM_{2.5}$	Ions (SO_4^{2-}, NO_3^-, Cl^-, Na^+, NH_4^+, K^+, Mg^{2+}, Ca^{2+})	Organic carbon, Elemental carbon	Trace elementals (Si, S, K, Ca, Ti, V, Cr, Mn, Fe, Ni, Cu, Zn, As, Se, Br, Ba, Pb)
Measurement method	Beta attenuation monitoring	Ion chromatography	Thermal-optical transmission method	X-ray fluorescence

- 20 species were selected after QA/QC for modeling.
- T(# of days)=61 (Nov 1- Dec 31, 2019).
- Based on prior knowledge, secondary sulfate and nitrate, motor vehicle emissions, oil combustion, industrial sources, biomass burning, soil, and sea salt were presumed to be potential candidate sources.
- Note that not all of those sources are major regional sources. Some are local (city-specific) sources rather than regional sources.

Real Application: Korea $PM_{2.5}$ speciation data

- Prior knowledge on potential source types and site-by-site analysis performed by Positive Matrix Factorization (PMF, Paatero and Tapper, 1994) and Bayesian Nonnegative Factor Analysis (BNFA) software (Park et al., CHEMOLAB 2021) were utilized in building a set of candidate models to be compared in BSMRM.
- Twelve candidate models with varying q (= 3, 4, or 5) along with different identifiability conditions (prespecified zero elements in **P**) for each q were considered for BSMRM.
- A BSMRM model with 3 sources resulted in the highest marginal likelihood among those 12 models regardless of the number of underlying process locations (L=9, 11, 12, 16) used for modeling fitting.

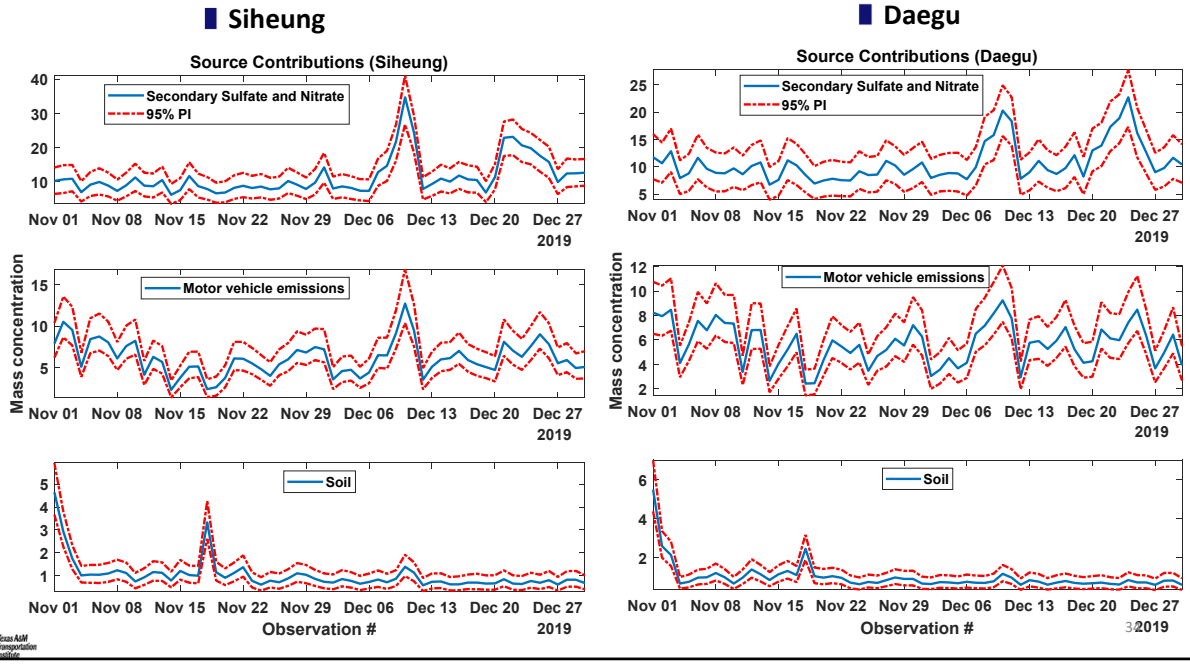
Estimated Source Compositions and Contributions with uncertainty estimates at an unmonitored site (Daegu)



33

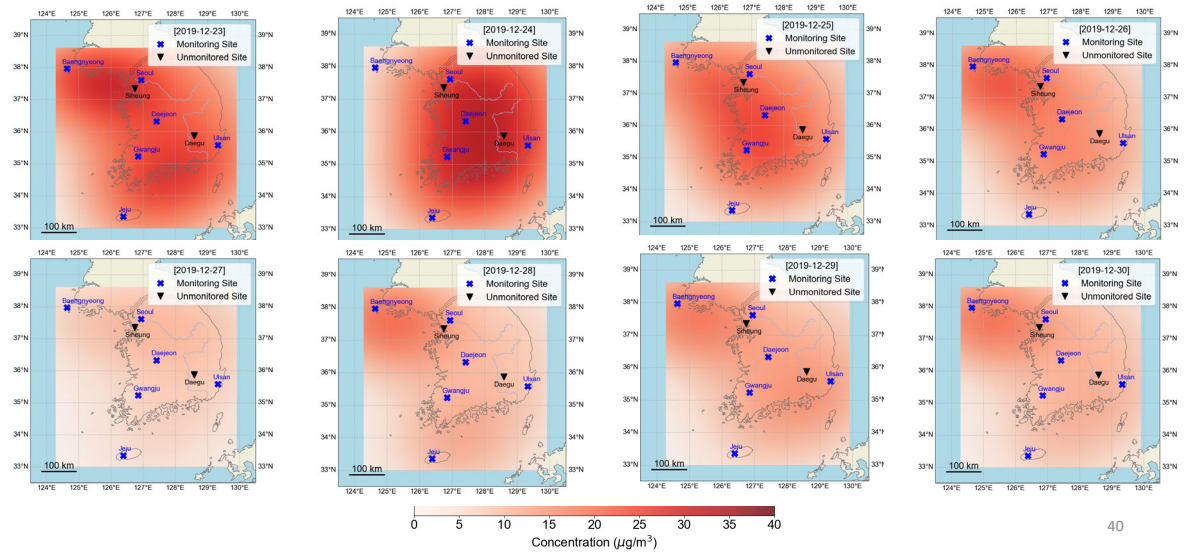
33

Predicted source contributions at unmonitored sites



34

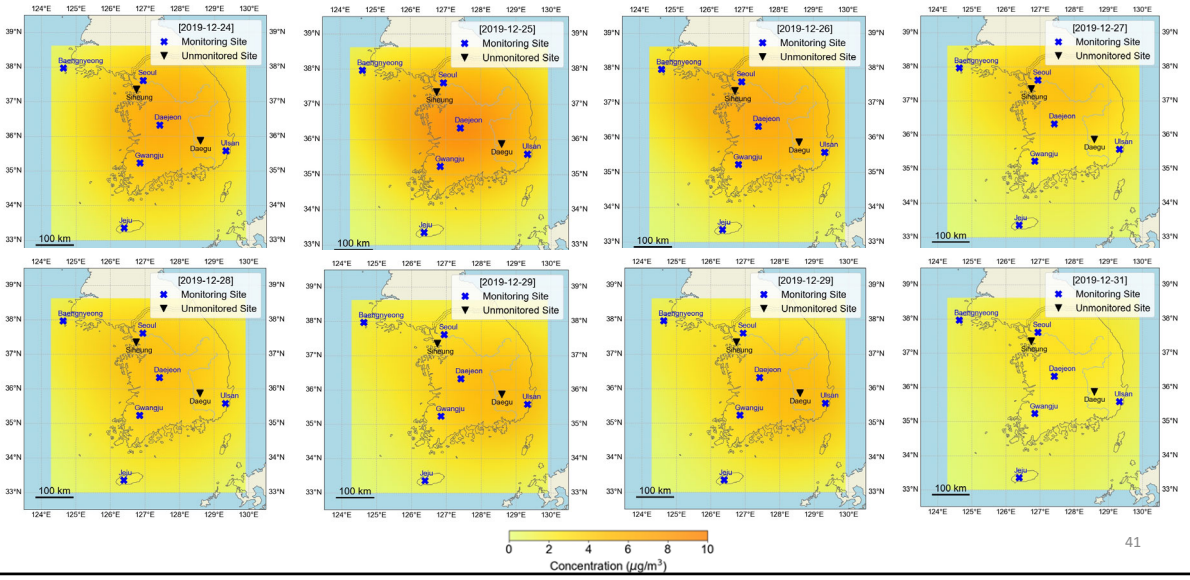
Secondary Sulfate and Nitrate contribution surface for eight consecutive days



40

35

Motor vehicle emission contribution surface for eight consecutive days



41

36

Summary and Discussion

- Multipollutant data obtained from ambient monitoring stations allow for identification of source profiles and contributions.
- Spatially extended Bayesian multivariate receptor modeling (**BSMRM**) can be used for improved source-specific exposure assessment.
  - BSMRM enables spatial prediction of source contributions at any location, which may significantly reduce exposure misclassification.
  - BSMRM can handle model uncertainty as well as providing uncertainty estimates of predicted source contributions simultaneously.
- It is important to select candidate models by good exploratory analysis and/or prior knowledge about the problem so that a reasonable model can be included in model comparison.

Work in Progress

- Extension of BSMRM
  - To incorporate traffic and land use covariates and total PM2.5 mass concentration data into modeling of source contributions to improve prediction of small-scale variation of source-specific exposures.
  - To explicitly incorporate meteorology variables into modeling
  - To include health outcome models
    - Joint estimation of source-specific exposures and health effect parameters based on the multipollutant data from a single monitoring site: Park et al. (Biostatistics 2014), Park and Oh (Environmetrics 2018).
    - Extension to the multipollutant data from multiple monitoring sites: ongoing

References

- Anderson (2003). 3rd ed., New York; Wiley.
- Buzcu and Fraser (2006). Atmospheric Environment, 40: 2385-2400.
- Calder (2007). Environ Ecol Stat 14: 229–247.
- Oh (1999), Computational Statistics and Data Analysis, 29, 411–427.
- Paatero and Tapper (1994). Environmetrics, 5, 111–126.
- Park, Oh, and Guttorp (2002). Chemometrics and Intelligent Laboratory Systems, 60: 49-67.
- Park, Spiegelman, and Henry (2002), Environmetrics, 13, 775-798.
- Park et al. (2014). Biostatistics, 15: 484-497.
- Park and Oh (2018). Environmetrics, 29(1), 2484.
- Park et al. (2018). Technometrics, 60: 306-318.
- Park, Lee, and Oh (2021). Chemometrics and Intelligent Laboratory Systems, 211, 104280.

Thank you

Email: e-park@tti.tamu.edu