Lecture 10: Multivariate Processes

SISMID 2025

Howard Chang
howard.chang@emory.edu

Multivariate Spatial Data

Examples:

- Multiple outcomes are collected at the same spatial locations that can be potentially correlated (e.g., rates of different cancers across counties).
- Spatially-varying coefficient models where regression coefficients (e.g., intercept and slope) are dependent.
- Multiple component models (e.g., mixture model, zero-inflated model) where the components are dependent.

Problem: Additional correlations within a spatial location needs to be accounted for.

Advantages: (1) Learn about correlation between outcomes and (2) improve statistical precision.

Approach 1: Multivariate CAR Model

This approach describes multivariate spatial dependence using multivariate conditional distributions.

Let $\mathbf{w}_s$ be a $K \times 1$ outcome vector at location s. We assume

$$\mathbf{w}_s \,|\, \mathbf{w}_{-s} \sim N_K\left(\frac{1}{n_s}\sum_{l \in \delta_s} \mathbf{w}_l,\ \frac{1}{n_s}\boldsymbol{\Sigma}\right)$$

where

- δ_s is the set of neighbors with unit s.
- n_s is the number of neighbors.
- Σ is a $K \times K$ covariance matrix. The off-diagonal elements captures conditional dependence between outcomes.

We can show that the above defines a unique joint Normal distribution of dimension $K \times S$.

Separable Model

Given the conditional distribution

$$\mathbf{w}_s \mid \mathbf{w}_{-s} \sim N_K\left(\frac{1}{n_s}\sum_{l\in\delta_s}\mathbf{w}_l,\ \frac{1}{n_s}\Sigma\right)$$

The joint distribution of $\mathbf{w} = (\mathbf{w}_1', \mathbf{w}_2', \ldots, \mathbf{w}_S')'$ is given by

$$\mathbf{w} \sim \mathrm{N}_{KN}\left(\mathbf{0}, (\mathbf{D} - \mathbf{W})^{-1} \otimes \boldsymbol{\Sigma}\ \right)$$

where $\mathbf{W}$ is the adjacency matrix and $\mathbf{D}$ is diagonal with element n_s.

We note that $(\mathbf{D} - \mathbf{W})^{-1}$ is the spatial correlation matrix induced by the iCAR model. We can similarly derive other CAR-based models (i.e., proper, Leroux).

Separable Models

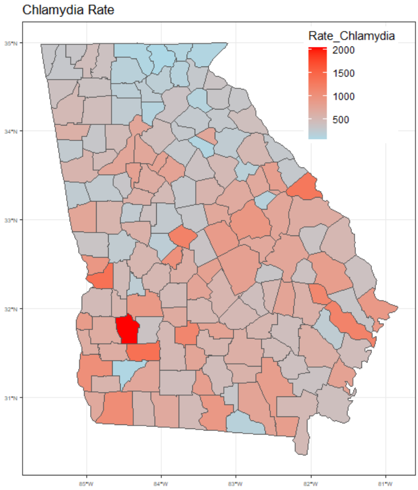
Let's consider the simple bivariate case. Assume two separable exponential covariance functions for bivariate $w_1(s)$ and $w_2(s')$ given by

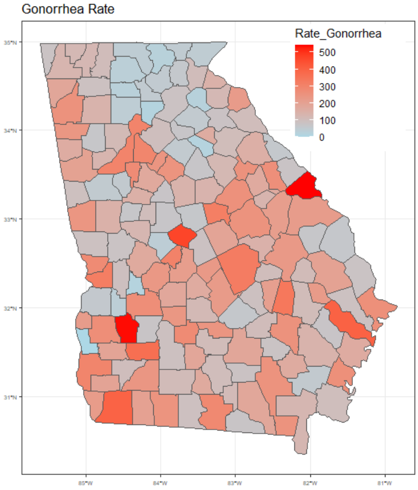
$$\sigma^2 \exp\left\{-\frac{1}{\rho_1}||s_i - s_j||\right\} \times \exp\left\{-\frac{1}{\rho_2}\mathbf{I}(k = k')\right\} .$$

- At the same location, the correlation between $w_1(s)$ and $w_2(s)$ is e^{-1/ρ_2}.
- For each outcome $k = 1$ or $k = 2$, the spatial dependence is identical.
- We often assume σ^2 is different across outcomes.

Case Study 1: Chlamydia and Gonorrhea Rates in 2019

Correlation between the two rates $= 0.85$





Bivariate CAR Models

- $Y_1(s)$ = county counts of Chlamydia
- $Y_2(s)$ = county counts of Gonorrhea

$$Y_1(s) \sim \text{Pois}\ \left(P_s \times \mu_1(s)\right) \qquad Y_2(s) \sim \text{Pois}\ \left(P_s \times \mu_2(s)\right)$$

$$\log \mu_1(s) = \beta_{0,1} + \phi_1(s)$$
$$\log \mu_2(s) = \beta_{0,1} + \phi_2(s)$$

We will consider modeling $\phi_1(s)$ and $\phi_2(s)$ jointly as

- Between-outcome = independent or correlated
- Between-county = instrinsic CAR (iCAR) or proper CAR (pCAR).

Model Comparison

Outcome	Spatial	WAIC	Eff. Param
Independent	iCAR	2,479	165
Independent	pCAR	2,493	171
Dependent	iCAR	2,462	152
Dependent	pCAR	2,452	148

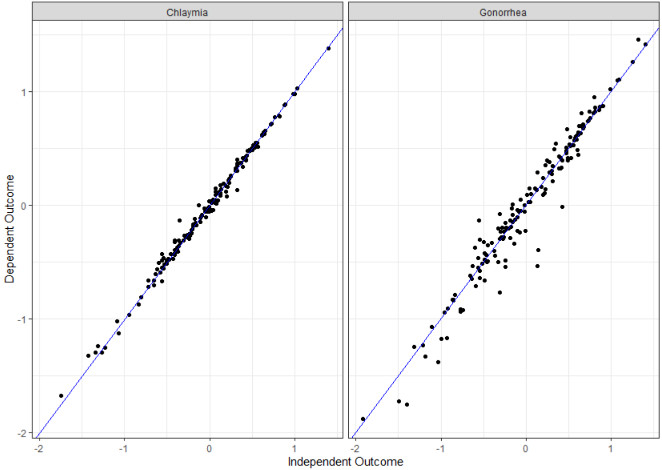
Model Parameter Estimates

Posterior Mean (SD) and 95% Interval			
Outcome	Spatial	$\log \tau_1^2$	$\log \tau_2^2$
Independent	iCAR	0.02 (0.12)	0.51 (0.13)
Independent	pCAR	0.11 (0.13)	0.59 (0.13)
Dependent	iCAR	0.00 (0.05)	0.52 (0.02)
Dependent	pCAR	0.06 (0.12)	0.61 (0.12)

Outcome	Spatial	pCAR (ρ)	Outcome Corr.
Independent	iCAR		
Independent	pCAR	0.859 (0.687, 0.858)	
Dependent	iCAR		0.933 (0.930, 0.936)
Dependent	pCAR	0.845 (0.653, 0.949)	0.922 (0.884, 0.949)

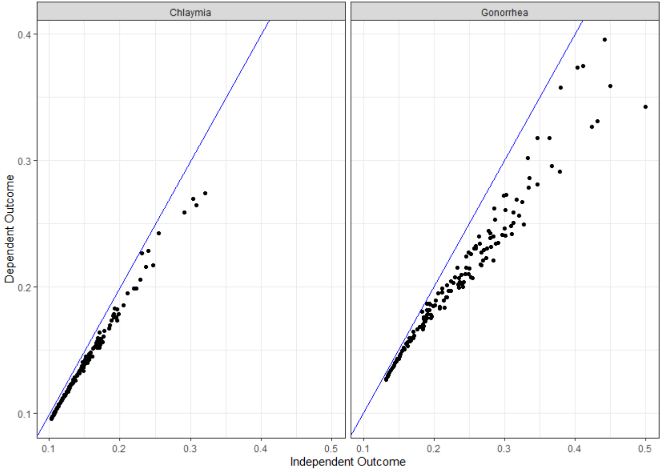
Random Effect Estimation

Posterior Mean of $\phi_k(s)$ from independent and dependent pCAR models.



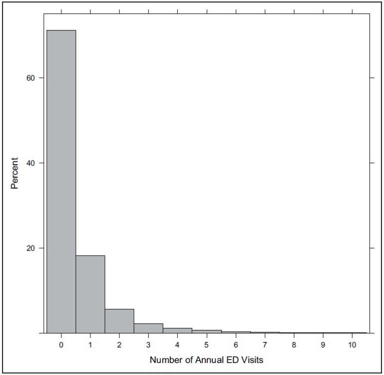
Random Effect Estimation Uncertainty

Posterior Standard Deviation of $\phi_k(s)$ from independent and dependent pCAR models.



Case Study 2: Spatial-Temporal ED Visits

Counts of emergency department visit are often right-skewed with an excess of zeros.



Duke (2008-2011). Neelon et al. (2014). Spatiotemporal hurdle models for zero-inflated count data: Exploring trends in emergency department visits. Stat Methods Med Res. PMID: 24682266.

Zero-Inflated versus Hurdle Model

Let π be the probability that a count outcome Y is not-zero.

Zero-Inflated Model

$$P(Y = y) = (1 - \pi)1_{y=0} + \pi f(y)1_{y>0}$$

Hurdle Model

$$P(Y = y) = (1 - \pi)1_{y=0} + \pi \frac{f(y)}{1 - f(y = 0)}1_{y>0}$$

The 0's in a zero-inflation model is contaminated by the count distribution.

Spatial Hurdle Model

A hurdle model is one approach to model zero-inflated count data.

Let Y_{ijk} denote the number of ED visits for the kth patient in blockgroup i and year j. Let π_{ijk} denote the probability that $Y_{ij} > 0$.

The hurdle model is given by

$$P(Y_{ijk} = y_{ijk}) = (1 - \pi_{ijk})1_{y_{ijk}=0} + \pi_{ijk}\frac{f(y_{ijk})}{1 - f(0)}1_{y_{ijk}>0}$$

where $f(y_{ijk})$ is the density function of a distribution for count (e.g. Poisson, negative binomial) with mean μ_{ijk}.

Covariates enter in both zero and the non-zero component:

$$\text{logit}(\pi_{ijk}) = \mathbf{X}'_{ijk}\boldsymbol{\beta}_1 + \phi_{1i} + \nu_{1j} + \delta_{1ij}$$

$$\log(\mu_{ijk}) = \mathbf{X}'_{ijk}\boldsymbol{\beta}_2 + \phi_{2i} + \nu_{2j} + \delta_{2ij}$$

Spatial Hurdle Model

The above model assumes the space-time residuals is decomposed into:

- A purely spatial term:

$$\boldsymbol{\phi}_i = [\phi_{1i}, \phi_{2i}]' \sim iCAR(\boldsymbol{\Sigma}_\phi)$$

- A purely temporal term:

$\mathbf{v}_i = [\nu_{1j}, \nu_{2j}]'$ estimated as fixed effects or bi-variate iCAR.

- Dynamic spatial residual:

$$\boldsymbol{\delta}_j = [\delta_{1j}, \delta_{2j}, \ldots, \delta_{Sj}]'$$

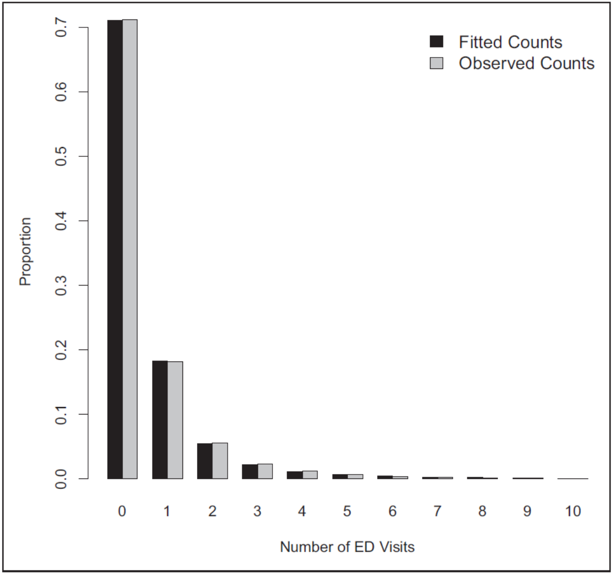
$$\boldsymbol{\delta}_j = \alpha \boldsymbol{\delta}_{j-1} + \boldsymbol{\theta}_j \qquad \boldsymbol{\theta}_j \sim iCAR(\boldsymbol{\Sigma}_\theta)$$

By assuming $\boldsymbol{\Sigma}_\phi$ and $\boldsymbol{\Sigma}_\theta$ to be non-diagonal, we allow dependence between the zero and the non-zero component. This dependence is separable from the spatial and temporal dependence.

Model Comparison

Base Distribution	Temporal Effects	DIC	p_D
Poisson	Fixed	232,158	566
Poisson	Bivariate iCAR	232,171	574
Negative Binomial	Fixed	211,198	367
Negative Binomial	Bivariate iCAR	211,209	377
Generalized Poisson	Fixed	211,035	367
Generalized Poisson	Bivariate iCAR	211,046	374

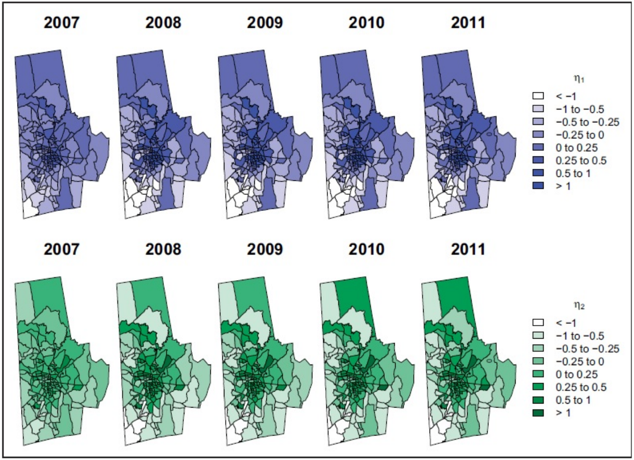
Observed vs Fitted Values



ED Visits in Durham County, NC

Estimated space-time residual random effects.

$$\eta_1 = \phi_{1i} + \nu_{1j} + \delta_{1ij} \qquad \eta_2 = \phi_{2i} + \nu_{2j} + \delta_{2ij}$$



Parameter Estimates

Covariate	logit(π_{ijk}) model OR	log (μ_{ijk}) model RR
Baseline prob or mean	0.23 (0.20, 0.25)	0.26 (0.19, 0.32)
Self-pay vs Private	4.66 (4.43, 4.90)	1.66 (1.58, 1.77)
Male vs Female	1.21 (1.17, 1.25)	0.92 (0.88, 0.96)
Hisp vs Non-Hisp	1.63 (1.54, 1.73)	1.68 (1.84, 1.55)

Spatial Parameter Estimates

Covariate	Spatial Random Effect Estimate (95% P.I.)	Dynamic Spat. Residual Estimate (95% P.I.)
$\Sigma[1,1]$ (logit(π) Model)	0.22 (0.16, 0.29)	0.03 (0.02, 0.04)
$\Sigma[2,2]$ (log(μ) Model)	0.14 (0.09, 0.21)	0.06 (0.04, 0.09)
Correlation	0.56 (0.37, 0.72)	-0.03 (-0.28, 0.25)
Autoregressive param		0.59 (0.25, 0.84)

Approach 2: Linear Model of Coregionalization

For multivariate continuous spatial process, we often wish to have different spatial dependence across outcomes. One constructive approach is known as linear model of coregionalization (LMC).

Assume

$$\begin{bmatrix} W_1(s) \\ W_2(s) \end{bmatrix} = \mathbf{A} \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix}$$

- A_{11}, A_{21}, A_{22} are unknown constants.
- $U_1(s)$ and $U_2(s)$ are two *independent* zero-mean Gaussian spatial processes. The covariance functions of U_1 and U_2 may be different!

For the bivariate random variable at the same location s:

$$Cov[W_1(s), W_2(s)] = \begin{bmatrix} A_{11}^2 & A_{11}A_{21} \\ A_{11}A_{21} & A_{21}^2 + A_{22}^2 \end{bmatrix} .$$

Linear Model of Coregionalization

The covariance for $W_1(2)$ at different locations is

$$Cov[W_1(s), W_1(s')] = A_{11}^2 \, Corr[\, U_1(s), U_1(s') \,] \, .$$

The covariance for $W_2(2)$ at different locations is

$$Cov[W_2(s), W_2(s')] = A_{21}^2 \, Corr[\, U_1(s), U_1(s') \,] + A_{22}^2 \, Corr[\, U_2(s), U_2(s') \,] \, .$$

Let $\mathbf{H}_1$ and $\mathbf{H}_2$ denote the covariance matrix of the latent processes $U_1(s)$ and $U_2(s)$. Also let $\mathbf{T}_j = \mathbf{a}_j' \mathbf{a}_j$, where $\mathbf{a}_j$ is the jth column of $\mathbf{A}$. The joint distribution of $\mathbf{W}$ has a non-separable covariance matrix

$$\mathbf{H}_1 \otimes \mathbf{T}_1 + \mathbf{H}_2 \otimes \mathbf{T}_2 \, .$$

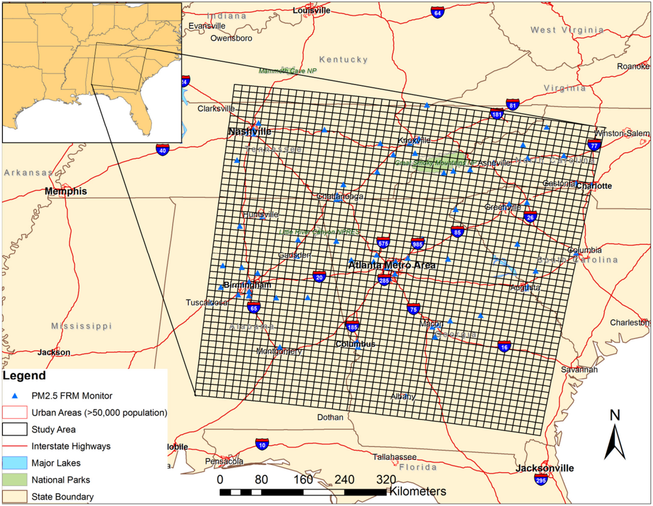
Case Study 3: Calibrating Satellite AOD for Fine Particulate Air Pollution

Aerosol optical depth (AOD)

- Remotely-sensed satellite images $\rightarrow$ large spatial coverage.
- Measures the degree to which aerosols prevent light from penetrating the atmosphere.
- Contain missing data (e.g. due to cloud cover).
- Positive empirical associations between AOD and ambient concentrations.

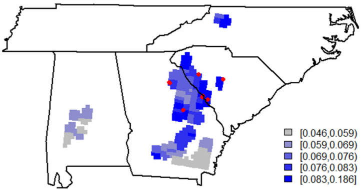
AOD Southeastern US Study Area, 2003-2005

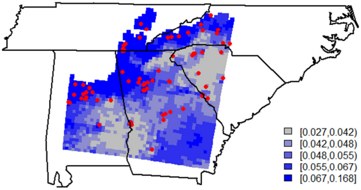
2,400 10km×10km AOD grid cells; 85 monitors

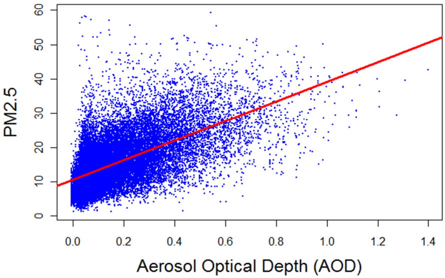


AOD Data

Example Days







Statistical Model

At monitoring location s on day t,

$$\text{PM}(\mathbf{s},t) = \alpha_0(\mathbf{s},t) + \alpha_1(\mathbf{s},t)\text{AOD}(s,t) + \epsilon(\mathbf{s},t)$$

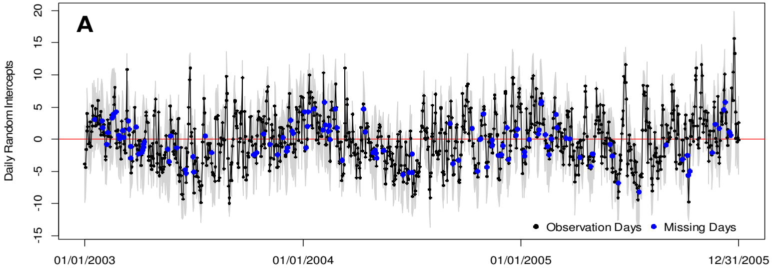
$$\begin{bmatrix}\alpha_0(\mathbf{s},t)\\ \alpha_1(\mathbf{s},t)\end{bmatrix} = \begin{bmatrix}\mathbf{Z}_0(\mathbf{s},t)\boldsymbol{\beta}_0\\ \mathbf{Z}_1(\mathbf{s},t)\boldsymbol{\beta}_1\end{bmatrix} + \begin{bmatrix}\theta_0(\mathbf{s})\\ \theta_1(\mathbf{s})\end{bmatrix} + \begin{bmatrix}\gamma_0(t)\\ \gamma_1(t)\end{bmatrix}$$

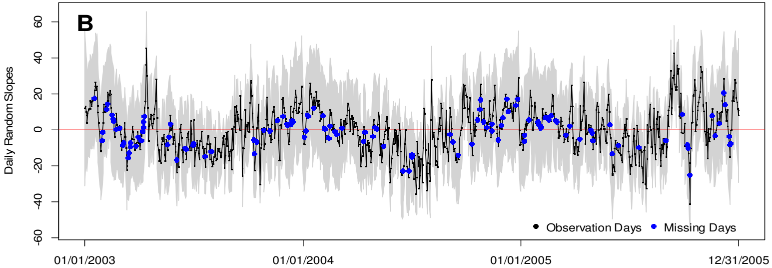
$$\epsilon(\mathbf{s},t) \sim \text{N}\ (0,\sigma^2)$$

- $\mathbf{Z}$: daily meteorological variables and land use variables*.
- Spatial effects $[\theta_0(s), \theta_1(s)]$ follows LMC with latent GP and exponential covariance functions.
- Temporal effects $[\gamma_0(t), \gamma_1(t)]$ are independent first-order pCAR.

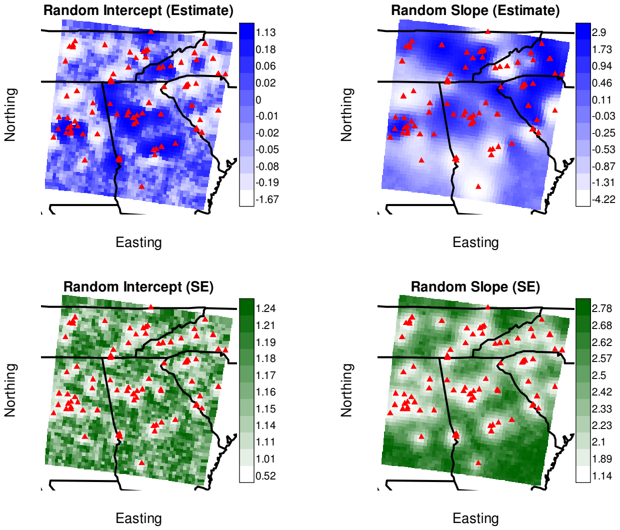
*elevation, wind speed, average daily temperature, major road way length, percent forest cover, and the presence of source emissions

Residual Temporal Biases





Residual Spatial Biases



Prediction Performance

	RMSE	MAE	90% PI Length	90% PI Coverage	R^2
On days without AOD-observation pairs					
Temporal	4.43	3.34	16.2	0.94	0.66
Independent	5.45	4.18	19.9	0.95	0.48
At locations without monitors					
Spatial	3.75	2.69	12.2	0.91	0.79
Independent	3.81	2.72	12.3	0.91	0.79