**PRACTICUM: Assessing Potential Nonstationarity in Spatial Data**

*Pan-American Advanced Study Institute on Spatio-Temporal Statistics*

Búzios, RJ, Brazil
June 16-26, 2014

Catherine Calder (calder@stat.osu.edu) and Mark Risser (risser.13@osu.edu)

## PART I: Simulating Nonstationary Spatial Data

What do nonstationary spatial processes look like? How can we determine whether we need to fit a nonstationary model to our data? To answer these questions, it can be helpful to examine synthetic data generated from a nonstationary spatial statistical model, such as the one given below.

---

Let $Z(\cdot)$ be a mean-zero spatial process defined for all $\boldsymbol{s} \in G \subset \mathbb{R}^2$, where

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s}),$$

$Y(\cdot) \sim \text{GP}(\mathbf{0}, \boldsymbol{\Omega}(\boldsymbol{\Xi}))$, $\epsilon(\boldsymbol{s}) \overset{iid}{\sim} \text{N}(0, \tau^2)$ for all $\boldsymbol{s} \in G$, and $Y(\cdot)$ and $\epsilon(\cdot)$ are independent. The elements of $\boldsymbol{\Omega}(\boldsymbol{\Xi})$ are $\Omega_{ij} \equiv \text{Cov}(Y(\mathbf{s_i}), Y(\mathbf{s_j}))$, which are given by a parametric covariance function

$$\text{Cov}(Y(\mathbf{s_i}), Y(\mathbf{s_j})) = C(\mathbf{s_i}, \mathbf{s_j} | \boldsymbol{\Xi}),$$

which depends on the generic parameter vector $\boldsymbol{\Xi}$. Following Paciorek and Schervish (2006) and Stein (2005), we take

$$C(\mathbf{s_i}, \mathbf{s_j}) = \sigma^2 |\boldsymbol{\Sigma_i}|^{1/4} |\boldsymbol{\Sigma_j}|^{1/4} \left| \frac{\boldsymbol{\Sigma_i} + \boldsymbol{\Sigma_j}}{2} \right|^{-1/2} \mathcal{M}_\nu \left( \sqrt{Q_{ij}} \right),$$

where $\mathcal{M}_\nu(\cdot)$ is the Matérn correlation function with smoothness parameter $\nu$, $\boldsymbol{\Sigma_i} \equiv \boldsymbol{\Sigma}(\mathbf{s_i})$ is the $(d \times d)$ covariance matrix for the Gaussian kernel function centered at location $\mathbf{s_i}$ (henceforth called the **kernel matrix**), and

$$Q_{ij} = (\mathbf{s_i} - \mathbf{s_j})' \left( \frac{\boldsymbol{\Sigma_i} + \boldsymbol{\Sigma_j}}{2} \right)^{-1} (\mathbf{s_i} - \mathbf{s_j}).$$

The kernel matrices are parameterized as follows. Each $2 \times 2$ kernel matrix is

$$\boldsymbol{\Sigma_i} = \begin{bmatrix} \sigma^2_{1,i} & \sigma_{3,i} \\ \sigma_{3,i} & \sigma^2_{2,i} \end{bmatrix},$$

and we define

$$\boldsymbol{\Psi_i} = \begin{bmatrix} \log(\sigma^2_{1,i}) \\ \log(\sigma^2_{2,i}) \\ \tan(\frac{\pi}{2}\rho_i) \end{bmatrix} \equiv \begin{bmatrix} \psi^i_1 \\ \psi^i_2 \\ \psi^i_3 \end{bmatrix}$$

(where $\rho_i = \sigma_{3,i}/(\sigma_{1,i} \cdot \sigma_{2,i}) \in (-1, 1)$), so that each element of $\mathbf{\Psi_i}$ has support on $\mathbb{R}$ and can be modeled using a Gaussian process with an exponential covariance structure. For example, $\mathrm{E}[\psi_1^i] = m_1$ and

$$\mathrm{cov}(\psi_1^i, \psi_1^j) = v_1^2 \exp\left\{-\frac{||\mathbf{s_i} - \mathbf{s_j}||}{r_1}\right\}$$

for all $\boldsymbol{s}_i, \boldsymbol{s}_j \in G$. Thus, the model for the kernel matrices has three mean parameters $\mathbf{m} = (m_1, m_2, m_3)'$, three spatial variance parameters $\mathbf{v} = (v_1, v_2, v_3)'$, and three range parameters $\mathbf{r} = (r_1, r_2, r_3)'$.

---

The file `part1.R` contains R code to generate realizations from this model for a set values of $\tau^2, \sigma^2, \nu, \boldsymbol{m}, \boldsymbol{v},$ and $\boldsymbol{r}$. The kernel matrix parameters $\mathbf{\Psi}_i$ vary smoothly across space.

**Some ideas:**

1. Try generating data using different parameter values for the kernel matrix model parameters (i.e., $\boldsymbol{m}, \boldsymbol{v},$ and $\boldsymbol{r}$). How do the kernel ellipses change? Do the realizations of $Z(\cdot)$ look noticeably different?

2. Using functions in the `geoR` library, use classical geostatistical tools to assess whether the simulated data come from a stationarity process (e.g., empirical variograms of subsets of the data) pretending that you do not know the true data generating model. Can you find values of $\boldsymbol{m}, \boldsymbol{v},$ and $\boldsymbol{r}$ such that exploratory techniques such as variogram analyses clearly pick up evidence of nonstationarity?

**PART II: Data Challenge**

The amount of annual precipitation across the state of Colorado (USA) is known to vary considerably due in part to the state's diverse topology. In particular, western Colorado is highly mountainous (part of the Rocky Mountains) whereas eastern Colorado is fairly flat. See the elevation map below.

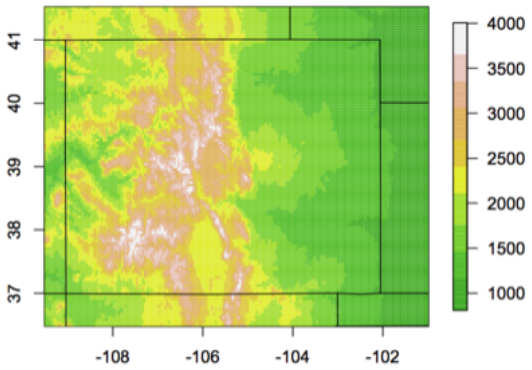


Figure 1: Colorado elevation (in meters) map.

The R data file `precip.Rdata` contains the `geodata` object `precip.geo`, which includes seven years of annual precipitation anomalies[1] at 202 locations (given by the longitude and latitude coordinates) across Colorado. (The meteorological data are available online from the National Center for Atmospheric Research at `http://www.image.ucar.edu/GSP/Data/US.monthly.met/CO.html`) Also included are the elevation (in meters) of the meteorological station locations.

**Some ideas:**

1. Perform exploratory analyses of the data to determine whether the spatial dependence structure of precipitation varies across the state. Do mountainous areas exhibit shorter or longer range spatial dependence than flatter regions?

2. Write down a spatial statistical model for these data. How can you include the elevation information in the model?

## References

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modeling using a new class of nonstationary covariance functions. *Environmetrics*, 17:483–506.

Stein, M. L. (2005). Nonstationary spatial covariance functions. *Unpublished technical report.*

[1] in site-specific standard deviation units based on historical records