Modeling and mapping non-stationary multivariate processes for large spatial datasets

Andrew O. Finley¹ and Sudipto Banerjee²

¹ Department of Forestry & Department of Geography, Michigan State University, East Lansing, Michigan, U.S.A.
² Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

March 22, 2011
Point-referenced spatial data often come as *multivariate measurements* at each location.
Point-referenced spatial data often come as multivariate measurements at each location.

Examples:

- **Environmental monitoring**: stations yield measurements on ozone, NO$_2$, CO, SO$_2$ and PM
Point-referenced spatial data often come as multivariate measurements at each location. Examples:

- **environmental monitoring**: stations yield measurements on ozone, NO$_2$, CO, SO$_2$ and PM
- **community ecology**: assemblage of plant species due to water, nutrients, temperature, and light requirements
Point-referenced spatial data often come as multivariate measurements at each location. Examples:

- **environmental monitoring**: stations yield measurements on ozone, NO$_2$, CO, SO$_2$ and PM
- **community ecology**: assemblage of plant species due to water, nutrients, temperature, and light requirements
- **forestry**: measurements of stand characteristics age, total biomass, and average tree diameter.
Point-referenced spatial data often come as multivariate measurements at each location. Examples:

- **environmental monitoring**: stations yield measurements on ozone, NO$_2$, CO, SO$_2$ and PM
- **community ecology**: assemblage of plant species due to water, nutrients, temperature, and light requirements
- **forestry**: measurements of stand characteristics age, total biomass, and average tree diameter.
- **atmospheric modeling**: at a given location we observe surface temperature, precipitation and wind speed
Point-referenced spatial data often come as multivariate measurements at each location.

Examples:

- **environmental monitoring**: stations yield measurements on ozone, NO$_2$, CO, SO$_2$ and PM
- **community ecology**: assemblage of plant species due to water, nutrients, temperature, and light requirements
- **forestry**: measurements of stand characteristics age, total biomass, and average tree diameter.
- **atmospheric modeling**: at a given location we observe surface temperature, precipitation and wind speed

We anticipate dependence between outcomes within a given location and across proximate locations (e.g., regional difference in the dependence within locations).
La Selva Biological Station soil data

Soil sample locations

Phosphorus (P)

Sum of base cations (SBC)

Sum of nitrogen (SN)
Ultimate goal is fit tree competition/survival models. Covariates include individuals’ access to environmental resources e.g., water, soil nutrients, light.
Overview and motivating data set

Motivating example from community ecology

Ultimate goal is fit tree competition/survival models. Covariates include individuals’ access to environmental resources e.g., water, soil nutrients, light.

Our objectives:

- predict soil nutrients for each tree’s location (i.e., to serve as competition model covariates)
- document how nutrients co-vary in these tropical soils
Ultimate goal is fit tree competition/survival models. Covariates include individuals’ access to environmental resources e.g., water, soil nutrients, light.

Our objectives:

- predict soil nutrients for each tree’s location (i.e., to serve as competition model covariates)
- document how nutrients co-vary in these tropical soils

Data from La Selva Biological Station in Costa Rica:

- soil samples $n = 251$
- three soil nutrients measured at each location
Each location contains \( m \) spatial outcomes

\[
y_k(s) = \mu_k(s) + w_k(s) + \epsilon_k(s), \quad k = 1, \ldots, m.
\]

- **Mean:** \( \mu(s) = \{\mu_k(s)\}_{k=1}^m = \{x_k(s)'\beta_k\}_{k=1}^m \)
- **Cov:** \( w(s) = \{w_k(s)\}_{k=1}^m \sim MVGP(\mathbf{0}, \Gamma(\cdot, \cdot)) \)

\[
\Gamma(s_1, s_2) = \{\text{cov}(w_i(s_1), w_j(s_2))\}_{i,j=1}^m
\]

- **Error:** \( \epsilon(s) = [\epsilon_k(s)]_{k=1}^m \sim MVN(\mathbf{0}, \Psi) \)

\( \Psi \) is an \( m \times m \) p.d. matrix, e.g. usually \( \text{diag}\{\tau_k^2\}_{k=1}^m \).
Multivariate Gaussian process are characterized by a cross-covariance function, $\Gamma(s_1, s_2; \theta) = \text{cov}\{w(s_1), w(s_2)\}$.

We use a constructive approach following factor analysis ideas and the “Linear Model of Coregionalization”:

$$w_k(s) = a_{k1}(s)v_1(s) + a_{k2}(s)v_2(s) + \cdots + a_{kk}(s)v_k(s)$$

$$\implies w(s) = A(s)v(s).$$

- $v_k(s)$’s follow univariate Gaussian processes, independent for $k = 1, \ldots, m$
**Multivariate** Gaussian process are characterized by a cross-covariance function, \( \Gamma(s_1, s_2; \theta) = \text{cov}\{w(s_1), w(s_2)\} \).

We use a constructive approach following factor analysis ideas and the “Linear Model of Coregionalization”:

\[
w_k(s) = a_{k1}(s)v_1(s) + a_{k2}(s)v_2(s) + \cdots + a_{kk}(s)v_k(s)
\]

\[\implies w(s) = A(s)v(s).\]

- \( v_k(s) \)'s follow univariate Gaussian processes, independent for \( k = 1, \ldots, m \)
- \( A(s) = \{a_{ij}(s)\} \) is an \( m \times m \) non-singular, *lower-triangular* matrix
Multivariate Gaussian process are characterized by a cross-covariance function, \( \Gamma(s_1, s_2; \theta) = \text{cov}\{w(s_1), w(s_2)\} \).

We use a constructive approach following factor analysis ideas and the “Linear Model of Coregionalization”:

\[
\begin{align*}
  w_k(s) &= a_{k1}(s)v_1(s) + a_{k2}(s)v_2(s) + \cdots + a_{kk}(s)v_k(s) \\
  \implies w(s) &= A(s)v(s).
\end{align*}
\]

- \( v_k(s) \)'s follow univariate Gaussian processes, independent for \( k = 1, \ldots, m \)
- \( A(s) = \{a_{ij}(s)\} \) is an \( m \times m \) non-singular, \textit{lower-triangular} matrix
- \( a_{ij}(s) \)'s follow independent univariate (log) Gaussian processes for \( k = 1, \ldots, m(m + 1)/2 \)
**Multivariate** Gaussian process are characterized by a cross-covariance function, \( \Gamma(\mathbf{s}_1, \mathbf{s}_2; \theta) = \text{cov}\{\mathbf{w}(\mathbf{s}_1), \mathbf{w}(\mathbf{s}_2)\} \).

We use a constructive approach following factor analysis ideas and the “Linear Model of Coregionalization”:

\[
\begin{align*}
  w_k(\mathbf{s}) = a_{k1}(\mathbf{s})v_1(\mathbf{s}) + a_{k2}(\mathbf{s})v_2(\mathbf{s}) + \cdots + a_{kk}(\mathbf{s})v_k(\mathbf{s}) \\
  \implies \mathbf{w}(\mathbf{s}) = \mathbf{A}(\mathbf{s})\mathbf{v}(\mathbf{s}).
\end{align*}
\]

- \( v_k(\mathbf{s})'s \) follow univariate Gaussian processes, independent for \( k = 1, \ldots, m \)
- \( \mathbf{A}(\mathbf{s}) = \{a_{ij}(\mathbf{s})\} \) is an \( m \times m \) non-singular, lower-triangular matrix
- \( a_{ij}(\mathbf{s})'s \) follow independent univariate (log) Gaussian processes for \( k = 1, \ldots, m(m + 1)/2 \)
- \( \Gamma(\mathbf{s}_1, \mathbf{s}_2; \phi) = \mathbf{A}(\mathbf{s}_1)\text{diag}\{\rho_k(\mathbf{s}_1, \mathbf{s}_2; \phi_{vk})\}_{k=1}^{m}\mathbf{A}(\mathbf{s}_2)' \)
For a collection of locations $\mathcal{I} = \{s_1, \ldots, s_n\}$, we have $w = (w(s_1)', \ldots, w(s_n)')'$. With $w \sim N(0, \Sigma_w)$, where $\Sigma_w$ is an $nm \times nm$ matrix partitioned into $m \times m$ blocks with $\Gamma$ forming the $(i, j)$-th block.
For a collection of locations $\mathcal{S} = \{s_1, \ldots, s_n\}$, we have $w = (w(s_1)', \ldots, w(s_n)')'$. With $w \sim N(0, \Sigma_w)$, where $\Sigma_w$ is an $nm \times nm$ matrix partitioned into $m \times m$ blocks with $\Gamma$ forming the $(i, j)$-th block.

**Stumbling block:** For each MCMC iteration we must compute:

1. $\Sigma_w^{-1}$ and $|\Sigma_w|$
2. inverse and determinant of the $n \times n$ cov matrices for each of the $a_{ij}(s)$’s ($\frac{m(m+1)}{2}$ of them).

These are $O(n^3)$ operations! We must reduce the dimension of the problem.
Approaches to dimension reduction:

- covariance tapering (Furrer et al. 2006; Zhang and Du, 2007; Du et al. 2009; Kaufman et al., 2009)
- spectral domain: (Fuentes 2007; Paciorek, 2007)
- approximate using MRFs: INLA (Rue et al. 2009)
- approximations using cond. indep. (Vecchia 1988; Stein et al. 2004)
- low-rank approaches (Wahba, 1990; Lin et al., 2000; Kamman & Wand, 2003; Paciorek, 2007; Rasmussen & Williams, 2006; Stein 2007, 2008; Cressie & Johannesson, 2008; Banerjee et al., 2008)
Univariate predictive process case:

- consider $n^*$ “knots” $\mathcal{S}^* = \{s_1^*, \ldots, s_n^*\}$ with $n^* << n$

- process realization over knots: $w^* \sim MVN_{n^*}\{0, \sigma^2 R^*(\phi)\}$

$\mathcal{S}^*$ at $s_0$:

$$\tilde{w}(s) = E[w(s) \mid w^*] = \text{cov}\{w(s_1), w^*\} \text{var}(w^*)^{-1} w^*$$

- $\tilde{w}(s) \sim GP\{0, \sigma^2 \tilde{\rho}(\cdot)\}$, where

$$\sigma^2 \tilde{\rho}(s_1, s_2; \phi) = \text{cov}\{w(s_1), w^*\} \text{var}(w^*)^{-1} \text{cov}\{w^*, w(s_2)\}$$
This effectively reduces the dimension of the problem but introduces a bias in the spatial variance parameters.

Use a rectified predictive process (Finley et al., 2009; Finley et al., 2010; Banerjee et al., 2010):

\[
\tilde{w}_\varepsilon(s) \sim N \left[ \tilde{w}(s), \sigma^2 \{ 1 - r(s, \phi)' R^*(\phi)^{-1} r(s, \phi) \} \right].
\]

Here, we simply replace \( w(s) \) with \( \tilde{w}_\varepsilon(s) \) and \( a_{ij}(s) \)'s with corresponding \( \tilde{a}_{ij,\varepsilon}(s) \)'s.
Synthetic data generated from the non-stationary model using \( n = 500 \) and \( m = 2 \).

\[ y_1(s) \]

\[ y_2(s) \]

\[ a_{1,1}(s) \]

\[ a_{2,1}(s) \]

\[ a_{2,2}(s) \]
consider three candidate models

1. stationary (i.e., $A(s) = A$)
2. full non-stationary
3. predictive process non-stationary (100 and 225 knot intensity)

assess models’ ability to estimate parameters

estimate posterior of $\rho_{i,j}(s) = \frac{\Gamma_{i,j}(s)}{\sqrt{\Gamma_{i,i}(s)\Gamma_{j,j}(s)}}$

additional model selection

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Stationary</th>
<th>Full</th>
<th>Non-stationary</th>
<th>Predictive process</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{1,1}$</td>
<td>1</td>
<td>1.42 (0.89, 2.00)</td>
<td>1.04 (0.87, 1.21)</td>
<td>1.04 (0.86, 1.22)</td>
<td>1.10 (0.91, 1.28)</td>
</tr>
<tr>
<td>$\beta_{2,2}$</td>
<td>1</td>
<td>1.02 (0.94, 1.11)</td>
<td>1.01 (0.94, 1.08)</td>
<td>1.00 (0.94, 1.07)</td>
<td>1.01 (0.94, 1.07)</td>
</tr>
<tr>
<td>$\beta_{2,1}$</td>
<td>1</td>
<td>-0.50 (-1.43, 1.11)</td>
<td>0.66 (0.16, 1.21)</td>
<td>0.60 (0.15, 1.16)</td>
<td>0.45 (-0.17, 0.94)</td>
</tr>
<tr>
<td>$\beta_{2,2}$</td>
<td>5</td>
<td>5.03 (4.89, 5.17)</td>
<td>5.03 (4.92, 5.14)</td>
<td>5.04 (4.93, 5.14)</td>
<td>5.04 (4.92, 5.15)</td>
</tr>
<tr>
<td>$\sigma^2_{a_{1,1}}$</td>
<td>1</td>
<td>0.39 (0.24, 0.57)</td>
<td>0.88 (0.53, 1.52)</td>
<td>0.90 (0.52, 1.49)</td>
<td>0.84 (0.52, 1.38)</td>
</tr>
<tr>
<td>$\sigma^2_{a_{2,1}}$</td>
<td>5</td>
<td>-1.29 (-1.88, -0.91)</td>
<td>6.90 (4.11, 11.92)</td>
<td>5.42 (3.08, 9.54)</td>
<td>4.82 (2.29, 9.04)</td>
</tr>
<tr>
<td>$\sigma^2_{a_{2,2}}$</td>
<td>1</td>
<td>1.12 (0.98, 1.28)</td>
<td>0.42 (0.17, 1.27)</td>
<td>0.41 (0.15, 1.09)</td>
<td>0.45 (0.21, 0.98)</td>
</tr>
<tr>
<td>$\phi_a$</td>
<td>4</td>
<td>–</td>
<td>4.66 (3.15, 6.92)</td>
<td>4.13 (3.07, 5.95)</td>
<td>3.66 (3.03, 5.35)</td>
</tr>
<tr>
<td>$\phi_v$</td>
<td>6</td>
<td>9.20 (6.27, 12.53)</td>
<td>5.60 (3.54, 8.70)</td>
<td>6.59 (3.86, 9.66)</td>
<td>6.86 (4.00, 10.64)</td>
</tr>
<tr>
<td>$\tau^2_1$</td>
<td>0.5</td>
<td>0.53 (0.37, 0.72)</td>
<td>0.44 (0.35, 0.53)</td>
<td>0.42 (0.33, 0.52)</td>
<td>0.42 (0.33, 0.52)</td>
</tr>
<tr>
<td>$\tau^2_2$</td>
<td>0.5</td>
<td>0.37 (0.16, 0.75)</td>
<td>0.36 (0.19, 0.63)</td>
<td>0.29 (0.14, 0.53)</td>
<td>0.24 (0.12, 0.46)</td>
</tr>
<tr>
<td>G</td>
<td>156.64</td>
<td>174.76</td>
<td>149.03</td>
<td>140.02</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>765.43</td>
<td>636.63</td>
<td>569.88</td>
<td>533.37</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>922.07</td>
<td>811.4</td>
<td>718.9</td>
<td>673.39</td>
<td></td>
</tr>
</tbody>
</table>
Synthetic data

Full and predictive process model results

True $a_{1,1}(\mathbf{s})$

True $a_{2,1}(\mathbf{s})$

True $a_{2,2}(\mathbf{s})$

Full $a_{1,1}(\mathbf{s})$

Full $a_{2,1}(\mathbf{s})$

Full $a_{2,2}(\mathbf{s})$

Pred. proc.

100 knot $a_{1,1}(\mathbf{s})$

Pred. proc.

100 knot $a_{2,1}(\mathbf{s})$

Pred. proc.

225 knot $a_{2,2}(\mathbf{s})$
Median residual within location spatial correlation, $\rho_{2,1}(s) = \frac{\Gamma_{2,1}(s)}{\sqrt{\Gamma_{2,2}(s)\Gamma_{1,1}(s)}}$
Residual within location spatial correlation, significant at the 0.05 level (positive ●, negative●)
La Selva Biological Station soil data

Soil sample locations

Phosphorus (P)

Sum of base cations (SBC)

Sum of nitrogen (SN)

Wageningen University
consider three candidate models

1. stationary (i.e., $A(s) = A$)
2. full non-stationary
3. predictive process non-stationary (100 and 225 knot intensity)

estimate posterior of $\rho_{i,j}(s) = \frac{\Gamma_{i,j}(s)}{\sqrt{\Gamma_{i,i}(s)\Gamma_{j,j}(s)}}$

again using $G + P = D$
La Selva Biological Station soil data

Parameter credible intervals, 50 (2.5 97.5) percentiles, for soil nutrient data analysis candidate models.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Stationary</th>
<th>Full</th>
<th>Predictive process</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{0,P}$</td>
<td>0.71 (0.26, 1.35)</td>
<td>0.66 (0.22, 1.05)</td>
<td>0.64 (0.33, 1.20)</td>
</tr>
<tr>
<td>$\beta_{0,SBC}$</td>
<td>5.38 (5.03, 6.08)</td>
<td>5.18 (4.86, 5.49)</td>
<td>5.16 (4.83, 5.40)</td>
</tr>
<tr>
<td>$\beta_{0,SN}$</td>
<td>5.42 (4.97, 5.86)</td>
<td>5.53 (5.30, 5.78)</td>
<td>5.53 (5.31, 5.73)</td>
</tr>
<tr>
<td>$\sigma^2_{P,P}$</td>
<td>0.92 (0.52, 2.29)</td>
<td>0.20 (0.09, 0.53)</td>
<td>0.22 (0.08, 0.57)</td>
</tr>
<tr>
<td>$\sigma^2_{SBC,P}$</td>
<td>0.47 (0.25, 1.23)</td>
<td>0.24 (0.10, 0.63)</td>
<td>0.21 (0.10, 0.54)</td>
</tr>
<tr>
<td>$\sigma^2_{SN,P}$</td>
<td>0.49 (0.26, 1.25)</td>
<td>0.20 (0.09, 0.50)</td>
<td>0.23 (0.11, 0.75)</td>
</tr>
<tr>
<td>$\sigma^2_{SBC,SBC}$</td>
<td>0.44 (0.27, 1.08)</td>
<td>0.54 (0.18, 1.64)</td>
<td>0.36 (0.13, 1.01)</td>
</tr>
<tr>
<td>$\sigma^2_{SN,SBC}$</td>
<td>0.19 (0.06, 0.51)</td>
<td>0.14 (0.06, 0.36)</td>
<td>0.15 (0.07, 0.38)</td>
</tr>
<tr>
<td>$\sigma^2_{SN,SN}$</td>
<td>0.39 (0.19, 1.08)</td>
<td>1.85 (0.62, 6.11)</td>
<td>1.77 (0.41, 10.38)</td>
</tr>
<tr>
<td>$\phi_a$</td>
<td>–</td>
<td>0.0135 (0.0125, 0.0173)</td>
<td>0.0134 (0.0125, 0.0170)</td>
</tr>
<tr>
<td>$\phi_w$</td>
<td>0.0499 (0.0165, 0.0873)</td>
<td>0.0371 (0.0180, 0.0737)</td>
<td>0.0284 (0.0133, 0.0603)</td>
</tr>
<tr>
<td>Eff. range$_{a}$</td>
<td>–</td>
<td>222.13 (173.32, 238.97)</td>
<td>224.36 (176.57, 239.17)</td>
</tr>
<tr>
<td>Eff. range$_{w}$</td>
<td>60.04 (34.31, 181.08)</td>
<td>80.68 (40.66, 166.33)</td>
<td>105.64 (49.65, 225.05)</td>
</tr>
<tr>
<td>$\tau^2_P$</td>
<td>0.21 (0.14, 0.30)</td>
<td>0.19 (0.13, 0.28)</td>
<td>0.19 (0.13, 0.28)</td>
</tr>
<tr>
<td>$\tau^2_{SBC}$</td>
<td>0.07 (0.05, 0.11)</td>
<td>0.06 (0.04, 0.09)</td>
<td>0.06 (0.04, 0.09)</td>
</tr>
<tr>
<td>$\tau^2_{SN}$</td>
<td>0.15 (0.11, 0.21)</td>
<td>0.11 (0.07, 0.16)</td>
<td>0.09 (0.06, 0.14)</td>
</tr>
</tbody>
</table>
Parameter credible intervals, 50 (2.5 97.5) percentiles, for soil nutrient data analysis candidate models.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Stationary</th>
<th>Full</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>39.45</td>
<td>28.02</td>
<td>24.4</td>
</tr>
<tr>
<td>P</td>
<td>92.62</td>
<td>79.9</td>
<td>77.28</td>
</tr>
<tr>
<td>D</td>
<td>132.07</td>
<td>107.92</td>
<td>101.68</td>
</tr>
</tbody>
</table>
Non-stationary – full versus predictive process

Full $\rho(s)_{P,SBC}$

Full $\rho(s)_{P,SN}$

Full $\rho(s)_{SN,SBC}$

Pred. proc. $\rho(s)_{P,SBC}$

Pred. proc. $\rho(s)_{P,SN}$

Pred. proc. $\rho(s)_{SN,SBC}$
Non-stationary – full versus predictive process, 
\( \rho(\mathbf{s}) \) sig. at 0.05 level (●) positive, (●) negative
Non-stationary – full versus predictive process, 
\( \rho(s) \) range between 0.025-0.975 CI

Full \( \rho(s)_{P,SBC} \)  
Full \( \rho(s)_{P,SN} \)  
Full \( \rho(s)_{SN,SBC} \)  

Pred. proc. \( \rho(s)_{P,SBC} \)  
Pred. proc. \( \rho(s)_{P,SN} \)  
Pred. proc. \( \rho(s)_{SN,SBC} \)
Non-stationary – observed (interpolated) versus predictive process (predicted)

Obs. Phosphorus (P)  
Sum of base cations (SBC)  
Sum of nitrogen (SN)  
Pred. proc. P  
Pred. proc. SBC  
Pred. proc. SN
In summary:

- proposed framework to quantify occurrence of non-stationary among outcomes within and across locations

- computational burden is reduced via the predictive process

- samples from the posterior predictive distribution can now be fed into tree competition models (provides for a way to acknowledge uncertainty in the soil nutrient covariates)
Thank you!