low-rank smoothing splines for unusual spatial structures: smoothing estuaries and stream networks

M. Giovanna Ranalli
Department of Statistics, Colorado State University

Joint work with Jay Breidt and Haonan Wang, CSU
Thanks to Hal Walker, EPA and Erin Peterson, CSU
Problem # 1

New Hampshire Estuary – 97 sites where mercury in sediment concentrations has been surveyed in the years 2000/1 and 2003 (data from NHNCA and NHDES) – Relationship between mercury in sediment concentration and the grain size

observed log(HG/SILTCLAY)

- (2.445) - (1.520)
- (1.519) - (1.336)
- (1.335) - (1.218)
- (1.217) - (1.063)
- (1.062) - (0.851)
- (0.850) - (0.759)
- (0.758) - (0.512)
- (0.511) - (0.234)
- (0.233) - +0.019
+0.020 - +2.438

Estuary Boundary

0 1 2 3 4
Kilometers
Problem # 2

Maryland Stream Network – 955 sites where chemical, physical and biological variables were collected in the years 1995–1997 (data from MD NDR) – focus here on Acid Neutralizing Capacity.
Needs and Problems
Needs and Problems

NEEDS –

• Mapping quantities of interest, such as pollutants and water chemistry, by making predictions at non-observed locations

• Simple way to introduce covariates other than geographical coordinates
Needs and Problems

NEEDS –

• Mapping quantities of interest, such as pollutants and water chemistry, by making predictions at non-observed locations
• Simple way to introduce covariates other than geographical coordinates

Bivariate Smoothers like thin plate splines and kriging would obtain a map by employing covariance functions between locations that depend on their Euclidean distance.
Needs and Problems

NEEDS –

- Mapping quantities of interest, such as pollutants and water chemistry, by making predictions at non-observed locations
- Simple way to introduce covariates other than geographical coordinates

Bivariate Smoothers like thin plate splines and kriging would obtain a map by employing covariance functions between locations that depend on their Euclidean distance

PROBLEMS –

- Irregularly shaped non-convex domains
- Euclidean distance might not be a good way to measure similarity between data points
- Using non-euclidean distance metrics in kriging does not guarantee a positive definite covariance matrix (Rathbun, 1998; Gardner et al., 2003; Ver Hoef et al., 2004)
Low-rank thin plate splines – LTPS

Data from the example of the estuary (and of the stream network) are of the form \((x_i, y_i)\), for \(i = 1, \ldots, N\), with \(x_i\) geographical locations and \(y_i\) measurements of a variable of interest. Bivariate smoothing assumes that

\[
y = f(x) + \varepsilon, \tag{1}
\]

where \(y = (y_1, \ldots, y_N)^T\), \(f(\cdot)\) is some unspecified smooth function of \(x\) and the errors are such that \(E(\varepsilon) = 0\) and \(V(\varepsilon) = \sigma^2 I\).
Low-rank thin plate splines – LTPS

Data from the example of the estuary (and of the stream network) are of the form $(x_i, y_i)$, for $i = 1, \ldots, N$, with $x_i$ geographical locations and $y_i$ measurements of a variable of interest. Bivariate smoothing assumes that

$$y = f(x) + \varepsilon,$$

where $y = (y_1, \ldots, y_N)^T$, $f(\cdot)$ is some unspecified smooth function of $x$ and the errors are such that $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2 \varepsilon I$.

Ruppert et al. (2003) advocate the use of a mixed models–low rank representation of the problem to

1. speed computation
2. make computation easy through mixed models software
3. insert other covariates in the fixed part (parametric continuous or factors) or in the random part (nonparametric continuous and random effects)
The mixed model representation of model (1) is

\[ y = X\beta + Zu + \varepsilon, \]  

(2)
LTPS: the model and the predictions

The mixed model representation of model (1) is

\[ y = X\beta + Zu + \varepsilon, \]  

(2)

where

- \( X = \left[ 1 \ x_i \right]_{1 \leq i \leq N} \)
- \( Z \) contains \( T \leq N \) radial basis functions for the estimation of the non-linear structure of \( f(\cdot) \)
- \( u \) s.t. \( E(u) = 0, V(u) = \sigma_u^2 I \) are random effects independent of \( \varepsilon \)

This type of models can be fitted using PROC MIXED in SAS or the \texttt{lme()} function in R and Splus
The $Z$ matrix

$$Z = \left[ C(\|x_i, \kappa_t \|_E) \right]_{1 \leq i \leq N}^{1 \leq t \leq T} \left[ C(\|\kappa_t, \kappa_{t'} \|_E) \right]^{-1/2}_{1 \leq t, t' \leq T},$$

(3)
The Z matrix

\[
Z = \left[ C(\|x_i, \kappa_t\|_E) \right]_{1 \leq i \leq N} \left[ C(\|\kappa_t, \kappa_{t'}\|_E) \right]^{-1/2} \left\{ 1 \leq t, t' \leq T \right\},
\]

where

- \( \kappa_1, \ldots, \kappa_T \) is a set of knot locations (next slide)
- \( \| \cdot \|_E \) denotes Euclidean distance
- the function \( C \) is given by \( C(r) = r^2 \log r \)
The Z matrix

\[ Z = \left[ C(||x_i, \kappa_t||_E) \right]_{1 \leq i \leq N}^{1 \leq t \leq T} \left[ C(||\kappa_t, \kappa_{t'}||_E) \right]_{1 \leq t, t' \leq T}^{-1/2} \]

where

- \( \kappa_1, \ldots, \kappa_T \) is a set of knot locations (next slide)
- \( || \cdot ||_E \) denotes Euclidean distance
- the function \( C \) is given by \( C(r) = r^2 \log r \)

IF \( T = N \) \( \Rightarrow \) knots \( \equiv \) observations and FULL-RANK case (Thin plate splines)
The Z matrix

\[ Z = \left[ C(||x_i, \kappa_t||_E) \right]_{1 \leq i \leq N}^{1 \leq t \leq T} \left[ C(||\kappa_t, \kappa_{t'}||_E) \right]_{1 \leq t, t' \leq T}^{-1/2} \]  

where

- \( \kappa_1, \ldots, \kappa_T \) is a set of knot locations (next slide)
- \( || \cdot ||_E \) denotes Euclidean distance
- the function \( C \) is given by \( C(r) = r^2 \log r \)

IF \( T = N \) \( \Rightarrow \) knots \( \equiv \) observations and FULL-RANK case (Thin plate splines)

IF \( T = N \) \& \( C(r) \) some Matérn, exponential, gaussian corr functions \( \Rightarrow \) FULL-RANK Kriging
Knots – 2 issues: how many & where

**HOW MANY** rule of thumb: 1 every 3-4 observations

**WHERE** rectangular lattices, regular grids on the domain or space filling designs

*(FUNFITS in Splus and FIELDS in R do it)*
Knots – 2 issues: how many & where

**HOW MANY** rule of thumb: 1 every 3-4 observations

**WHERE** rectangular lattices, regular grids on the domain or space filling designs (FUNFITS in Splus and FIELDS in R do it)
Predictions

Once estimates from model (2) for $\beta$ and predictions for $u$ are obtained through Maximum Likelihood or REstricted ML, predicted values at observed locations are given by

$$\hat{y} = X\hat{\beta} + Z\hat{u}$$
Predictions

Once estimates from model (2) for $\beta$ and predictions for $u$ are obtained through Maximum Likelihood or REstricted ML, predicted values at observed locations are given by

$$\hat{y} = X\hat{\beta} + Z\hat{u}$$

The Splus commands for this would be simply

```r
fit<-lme(y~1+X, random=pdIdent(~1+Z))
beta<-fit$coef$fixed
u<-fit$coef$random
predictions<-X%*%beta+Z%*%u
```
Predictions

Once estimates from model (2) for $\beta$ and predictions for $u$ are obtained through Maximum Likelihood or REstricted ML, predicted values at observed locations are given by

$$\hat{y} = X\hat{\beta} + Z\hat{u}$$

The Splus commands for this would be simply

```r
fit<-lme(y~-1+X, random=pdIdent(~-1+Z))
beta<-fit$coef$fixed
u<-fit$coef$random
predictions<-X%*%beta+Z%*%u
```

Predictions at other locations can be done by adding new rows to $X$ and $Z$ to include the new prediction points.
LTPS fit of the estuary data

Recall observed data 1
Geodesic LTPS

Change the Euclidean distance measure in the $Z$ matrix in (3) with the GEODESIC DISTANCE = THE SHORTEST PATH A FISH WOULD SWIM
Geodesic LTPS

Change the Euclidean distance measure in the $Z$ matrix in (3) with the GEODESIC DISTANCE = THE SHORTEST PATH A FISH WOULD SWIM

$$Z_g = \left[ C(|x_i, \kappa_t|_G) \right]_{1 \leq i \leq N}^{1 \leq t \leq T} \left[ C'(||\kappa_t, \kappa_{t'}||_G) \right]_{1 \leq t, t' \leq T}^{-1/2}.$$
Geodesic LTPS

Change the Euclidean distance measure in the \( Z \) matrix in (3) with the GEODESIC DISTANCE = THE SHORTEST PATH A FISH WOULD SWIM

\[
Z_g = \left[ C(\|x_i, \kappa_t\|_G) \right]_{1 \leq i \leq N}^{1 \leq t \leq T} \left[ C(\|\kappa_t, \kappa_{t'}\|_G) \right]_{1 \leq t, t' \leq T}^{-1/2},
\]

- The geodesic distance is estimated by means of the Floyd Algorithm:
  1. build a graph for which nodes are locations and each node is connected to its \( nn \) nearest neighbors;
  2. obtain the shortest path between two locations and get the geodesic distance as the length of such path.
- The density of the data influences the final estimate.
Floyd Algorithm for the Northern part of the estuary
My recipe – The steps to obtain GLTPS

1. Select knots through a space filling design;
My recipe – The steps to obtain GLTPS

1. Select knots through a space filling design;
2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances;
My recipe – The steps to obtain GLTPS

1. Select knots through a space filling design;

2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances;

3. Estimate it with the Floyd Algorithm starting with $nn = 3$ and then increasing $nn$ until all the points are connected;
My recipe – The steps to obtain GLTPS

1. Select knots through a space filling design;

2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances;

3. Estimate it with the Floyd Algorithm starting with $nn = 3$ and then increasing $nn$ until all the points are connected;

4. Calculate the $X_P$ and the $Z_P$ matrices for all the grid points
My recipe – The steps to obtain GLTPS

1. Select knots through a space filling design;

2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances;

3. Estimate it with the Floyd Algorithm starting with $nn = 3$ and then increasing $nn$ until all the points are connected;

4. Calculate the $X_P$ and the $Z_P$ matrices for all the grid points

5. Obtain the $X$ and the $Z_g$ matrices for the observed data locations as a subset of rows of $X_P$ and $Z_P$;
My recipe – The steps to obtain GLTPS

1. Select knots through a space filling design;
2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances;
3. Estimate it with the Floyd Algorithm starting with $nn = 3$ and then increasing $nn$ until all the points are connected;
4. Calculate the $X_P$ and the $Z_P$ matrices for all the grid points
5. Obtain the $X$ and the $Z_g$ matrices for the observed data locations as a subset of rows of $X_P$ and $Z_P$;
6. Fiddle with lme models. Issues:
   - other covariates: size of $X_P$;
   - tests for the significance of the covariates are carried in the usual way;
   - tests for the significance of the random components (i.e. the spatial component) if done within the mixed models framework can be really conservative (alternative: Crainiceanu & Ruppert, 2004, for one variance component).
GLTPS fit of the estuary data

GLTPS
log(HG/SILTCLAY)

- (1.410) - (1.182)
- (1.181) - (1.149)
- (1.148) - (1.098)
- (1.097) - (0.965)
- (0.964) - (0.805)
- (0.804) - (0.720)
- (0.719) - (0.638)
- (0.637) - (0.516)
- (0.515) - (0.358)
- (0.357) - +0.168

Estuary Boundary
GLTPS vs LTPS fit of the estuary data

- Great Bay and Cocheco River
- Recall observed data 1
As the crow flies or as the fish swims??

Fit a model with both effects and then test for their significance

\[ y = X\beta + Zu + Z_g u_g + \varepsilon \]

where

\[
C = \begin{pmatrix}
\varepsilon \\
u \\
u_g
\end{pmatrix} = \begin{pmatrix}
\sigma^2_{\varepsilon} I & 0 & 0 \\
0 & \sigma^2_u I & 0 \\
0 & 0 & \sigma^2_g I
\end{pmatrix}
\]

Estimates of the variance components are again obtained via REML, testing their significance can be carried via nonparametric bootstrap (Opsomer et al., 2005).
## Bootstrap results for the estuary data

<table>
<thead>
<tr>
<th>Model</th>
<th>testing for</th>
<th>logReLik</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0 + Zu + Z_g u_g$</td>
<td>(full model)</td>
<td>-109.82</td>
<td>—</td>
</tr>
<tr>
<td>$\beta_0 + Zu$</td>
<td>$\sigma^2_g = 0$ (no within water spatial structure)</td>
<td>-114.17</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$\beta_0 + Z_g u_g$</td>
<td>$\sigma^2_g = 0$ (no Euclidean spatial structure)</td>
<td>-109.82</td>
<td>0.962</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>$\sigma^2_g = \sigma^2_u = 0$ (no spatial structure at all)</td>
<td>-113.23</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

⇝ The spatial structure suggested by the data is the within-water one
Application to stream networks – ANC

ANC is defined as the capacity of water to buffer acid \( \sim \) SMALL is bad!

<table>
<thead>
<tr>
<th>Observed ANC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(319.70) - 100.20</td>
<td></td>
</tr>
<tr>
<td>100.21 - 177.81</td>
<td></td>
</tr>
<tr>
<td>177.82 - 246.70</td>
<td></td>
</tr>
<tr>
<td>246.71 - 303.92</td>
<td></td>
</tr>
<tr>
<td>303.93 - 382.80</td>
<td></td>
</tr>
<tr>
<td>382.81 - 466.20</td>
<td></td>
</tr>
<tr>
<td>466.21 - 589.90</td>
<td></td>
</tr>
<tr>
<td>589.91 - 805.40</td>
<td></td>
</tr>
<tr>
<td>805.41 - 1252.25</td>
<td></td>
</tr>
<tr>
<td>1252.26 - 5715.83</td>
<td></td>
</tr>
</tbody>
</table>

Streams
New Issues

1. Intrinsic dimension of a stream network $\leadsto$ different $\mathbb{Z}_n$ matrix
New Issues

1. Intrinsic dimension of a stream network $\sim$ different $\mathbb{Z}_n$ matrix

2. Hydrologic distances cannot be efficiently calculated via the Floyd Alg $\sim$ GIS tools
   (thanks to D. Theobald and E. Peterson); now waiting for covariates and stream distances
   for locations different from the observed locations to make predictions
New Issues

1. Intrinsic dimension of a stream network $\leadsto$ different $Z_n$ matrix

2. Hydrologic distances cannot be efficiently calculated via the Floyd Alg $\leadsto$ GIS tools (thanks to D. Theobald and E. Peterson); now waiting for covariates and stream distances for locations different from the observed locations to make predictions

3. We will model ANC, account for both Euclidean and Asymmetric Hydrologic distance and test for which one is more supported by the data with the same bootstrap procedure employed earlier.
Meaningful distance measures for water chemistry along a stream network
P-splines for a stream network – loosely speaking!

1. Our approach is to define a 1-dimensional penalized spline (see Rupper, Wand & Carroll, 2003) along the stream network.
1. Our approach is to define a 1-dimensional penalized spline (see Rupper, Wand & Carroll, 2003) along the stream network.

2. If truncated linear basis functions are used, this is equivalent to use the matrix of Asymmetric Hydrologic distances to a set of knots’ locations as our $Z_n$ matrix.
P-splines for a stream network – loosely speaking!

1. Our approach is to define a 1-dimensional penalized spline (see Rupper, Wand & Carroll, 2003) along the stream network.

2. If truncated linear basis functions are used, this is equivalent to use the matrix of Asymmetric Hydrologic distances to a set of knots’ locations as our $Z_n$ matrix.

3. Given the sparsity of the dataset, knots are chosen to be all those locations which have at least 4 upstream neighbors.
ANC models and bootstrap testing

Significant covariates that enter the model linearly (the $X$ matrix): % pasture, % woody wetlands, % low density urban area in the watershed above a site and survey year
ANC models and bootstrap testing

Significant covariates that enter the model linearly (the $X$ matrix): % pasture, % woody wetlands, % low density urban area in the watershed above a site and survey year

<table>
<thead>
<tr>
<th>Model</th>
<th>testing for</th>
<th>logReLik</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X\beta + Zu + Z_nu_n$</td>
<td>(full model)</td>
<td>91.99</td>
<td>—</td>
</tr>
<tr>
<td>$X\beta + Zu$</td>
<td>$\sigma_n^2 = 0$ (no asym.hydro. spatial structure)</td>
<td>91.99</td>
<td>0.958</td>
</tr>
<tr>
<td>$X\beta + Z_nu_n$</td>
<td>$\sigma_u^2 = 0$ (no Euclidean spatial structure)</td>
<td>46.15</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$X\beta$</td>
<td>$\sigma_n^2 = \sigma_u^2 = 0$ (no spatial structure at all)</td>
<td>74.76</td>
<td>0.012</td>
</tr>
</tbody>
</table>
ANC models and bootstrap testing

Significant covariates that enter the model linearly (the $X$ matrix): % pasture, % woody wetlands, % low density urban area in the watershed above a site and survey year

<table>
<thead>
<tr>
<th>Model</th>
<th>testing for</th>
<th>logReLik</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X\beta + Zu + Z_n u_n$</td>
<td>(full model)</td>
<td>91.99</td>
<td>—</td>
</tr>
<tr>
<td>$X\beta + Zu$</td>
<td>$\sigma_n^2 = 0$ (no asym.hydro. spatial structure)</td>
<td>91.99</td>
<td>0.958</td>
</tr>
<tr>
<td>$X\beta + Z_n u_n$</td>
<td>$\sigma_u^2 = 0$ (no Euclidean spatial structure)</td>
<td>46.15</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$X\beta$</td>
<td>$\sigma_n^2 = \sigma_u^2 = 0$ (no spatial structure at all)</td>
<td>74.76</td>
<td>0.012</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Model</th>
<th>testing for</th>
<th>logReLik</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X\beta + Z_n u_n$</td>
<td>(full model)</td>
<td>91.99</td>
<td>—</td>
</tr>
<tr>
<td>$X\beta$</td>
<td>$\sigma_n^2 = 0$ (no asym.hydro. spatial structure)</td>
<td>74.76</td>
<td>0.425</td>
</tr>
</tbody>
</table>
Wrap up

1. It is possible to account for non regular domains when estimating quantities of interest.
Wrap up

1. It is possible to account for non regular domains when estimating quantities of interest.

2. The LTPS and p-splines framework allows inserting other covariates available for all prediction locations in a parametric or nonparametric way easily.
Wrap up

1. It is possible to account for non regular domains when estimating quantities of interest.

2. The LTPS and p-splines framework allows inserting other covariates available for all prediction locations in a parametric or nonparametric way easily.

3. This framework also allows to use bootstrap techniques to test for the distance metric more suitable for the problem at hand.
**Wrap up**

1. It is possible to account for non regular domains when estimating quantities of interest.

2. The LTPS and p-splines framework allows inserting other covariates available for all prediction locations in a parametric or nonparametric way easily.

3. This framework also allows to use bootstrap techniques to test for the distance metric more suitable for the problem at hand.

4. Applications other than estuaries and stream networks include domains with holes and irregular boundaries (lakes with islands/land with lakes), response over a nonflat domain (measurements on mountains) ...
Current and future work

1. Inserting other covariates to better map the NH estuary
Current and future work

1. Inserting other covariates to better map the NH estuary

2. Obtain hydrologic distances and covariates at prediction locations for the Maryland stream network
Current and future work

1. Inserting other covariates to better map the NH estuary
2. Obtain hydrologic distances and covariates at prediction locations for the Maryland stream network
3. How to model symmetric hydrologic distances if modeling something else than water chemistry?
Essential bibliography ... and Thank you!


Ver Hoef, J.M., Peterson, E. and Theobald D. (2004), Spatial statistical models that use flow and stream distance *Manuscript*.

The work reported here was developed under the STAR Research Assistance Agreement CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This presentation has not been formally reviewed by EPA. The views expressed here are solely those of the presenter and STARMAP. EPA does not endorse any products or commercial services mentioned in this presentation.