

Department of Statistics, UVA – March 4, 2005

STARMAP

Low-rank smoothing splines for complex domains and manifold recovery

M. Giovanna Ranalli

Department of Statistics, Colorado State University

Joint work with Haonan Wang, CSU

Outline

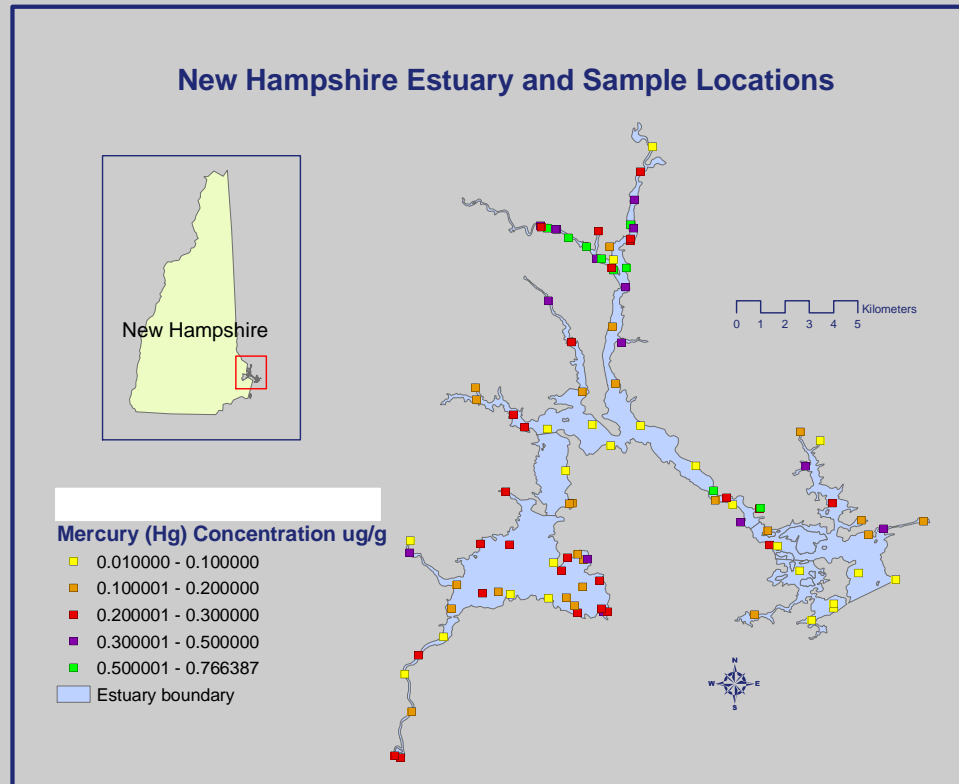
1. Low-rank smoothing splines for complex domains
 - Motivation
 - Low-rank thin plate splines
 - Geodesic Low-rank thin plate splines

Outline

1. Low-rank smoothing splines for complex domains
 - Motivation
 - Low-rank thin plate splines
 - Geodesic Low-rank thin plate splines
2. Low-rank smoothing splines for manifold recovery
 - Motivation
 - Examples of nonfunctional manifolds
 - How essentially the same idea employed in 1. can be used in these situations

Motivation

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)



Needs and Problems

Needs and Problems

NEEDS –

- Mapping quantities of interest, such as pollutants, 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, 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, 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 different 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 are of the form (\mathbf{x}_i, y_i) , for $i = 1, \dots, N$, with \mathbf{x}_i geographical locations and y_i measurements of a variable of interest. Bivariate smoothing assumes that

$$\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\varepsilon}, \quad (1)$$

where $\mathbf{y} = (y_1, \dots, y_N)^T$, $f(\cdot)$ is some unspecified smooth function of \mathbf{x} and the distribution of the errors is given by $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$.

Low-rank thin plate splines – LTPS

Data from the example of the estuary are of the form (\mathbf{x}_i, y_i) , for $i = 1, \dots, N$, with \mathbf{x}_i geographical locations and y_i measurements of a variable of interest. Bivariate smoothing assumes that

$$\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\varepsilon}, \quad (1)$$

where $\mathbf{y} = (y_1, \dots, y_N)^T$, $f(\cdot)$ is some unspecified smooth function of \mathbf{x} and the distribution of the errors is given by $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$.

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)

LTPS: the model and the predictions

The mixed model representation of model (1) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad (2)$$

LTPS: the model and the predictions

The mixed model representation of model (1) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad (2)$$

where

- $\mathbf{X} = [1 \ \mathbf{x}_i]_{1 \leq i \leq N}$
- \mathbf{Z} contains $T \leq N$ radial basis functions for the estimation of the non-linear structure of $f(\cdot)$
- $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{I})$ are random effects independent of $\boldsymbol{\varepsilon}$

This type of models can be fitted using PROC MIXED in SAS or the lme() function in Splus

The Z matrix

$$\mathbf{Z} = \left[C(\|\mathbf{x}_i, \boldsymbol{\kappa}_t\|_E) \right]_{\substack{1 \leq i \leq N \\ 1 \leq t \leq T}} \left[C(\|\boldsymbol{\kappa}_t, \boldsymbol{\kappa}_{t'}\|_E) \right]_{1 \leq t, t' \leq T}^{-1/2}, \quad (3)$$

The Z matrix

$$\mathbf{Z} = \left[C(\|\mathbf{x}_i, \boldsymbol{\kappa}_t\|_E) \right]_{\substack{1 \leq i \leq N \\ 1 \leq t \leq T}} \left[C(\|\boldsymbol{\kappa}_t, \boldsymbol{\kappa}_{t'}\|_E) \right]_{1 \leq t, t' \leq T}^{-1/2}, \quad (3)$$

where

- $\boldsymbol{\kappa}_1, \dots, \boldsymbol{\kappa}_T$ is a set of knot locations (next slide)
- $\|\cdot\|_E$ denotes Euclidean distance
- the function C is given by $C(r) = r^{2m-d} \log r$, if d is even and $C(r) = r^{2m-d}$, if d is odd, with d dimension of the predictors' space and $m > 1$ controlling the smoothness of f . If $m = 2$ we are penalizing the second derivative of f .

The Z matrix

$$\mathbf{Z} = \left[C(\|\mathbf{x}_i, \boldsymbol{\kappa}_t\|_E) \right]_{\substack{1 \leq i \leq N \\ 1 \leq t \leq T}} \left[C(\|\boldsymbol{\kappa}_t, \boldsymbol{\kappa}_{t'}\|_E) \right]_{1 \leq t, t' \leq T}^{-1/2}, \quad (3)$$

where

- $\boldsymbol{\kappa}_1, \dots, \boldsymbol{\kappa}_T$ is a set of knot locations (next slide)
- $\|\cdot\|_E$ denotes Euclidean distance
- the function C is given by $C(r) = r^{2m-d} \log r$, if d is even and $C(r) = r^{2m-d}$, if d is odd, with d dimension of the predictors' space and $m > 1$ controlling the smoothness of f . If $m = 2$ we are penalizing the second derivative of f .

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

The Z matrix

$$\mathbf{Z} = \left[C(\|\mathbf{x}_i, \boldsymbol{\kappa}_t\|_E) \right]_{\substack{1 \leq i \leq N \\ 1 \leq t \leq T}} \left[C(\|\boldsymbol{\kappa}_t, \boldsymbol{\kappa}_{t'}\|_E) \right]_{1 \leq t, t' \leq T}^{-1/2}, \quad (3)$$

where

- $\boldsymbol{\kappa}_1, \dots, \boldsymbol{\kappa}_T$ is a set of knot locations (next slide)
- $\|\cdot\|_E$ denotes Euclidean distance
- the function C is given by $C(r) = r^{2m-d} \log r$, if d is even and $C(r) = r^{2m-d}$, if d is odd, with d dimension of the predictors' space and $m > 1$ controlling the smoothness of f . If $m = 2$ we are penalizing the second derivative of f .

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

IF $T = N$ & $C(r)$ some Matérn, exponential, gaussian corr functions \rightsquigarrow FULL-RANK Kriging

Knots – 2 issues: how many & where

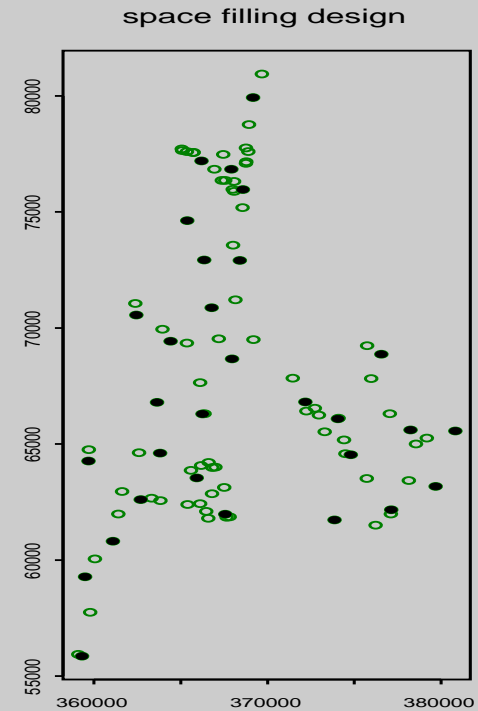
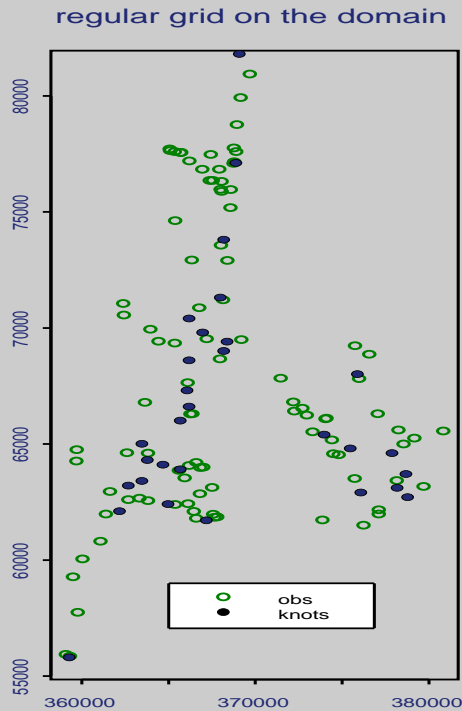
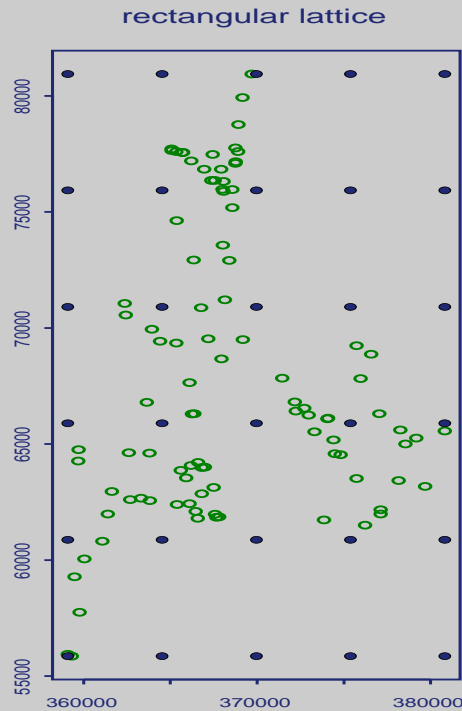
HOW MANY rule of thumb: 1 every 3-4 observations, never more than 100.

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

Knots – 2 issues: how many & where

HOW MANY rule of thumb: 1 every 3-4 observations, never more than 100.

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



Predictions

Once estimates from model (2) for β 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 β 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 `Spplus` commands for this would be simply

```
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 β and predictions for u are obtained through Maximum Likelihood or REstricted ML, predicted values at observed locations are given by

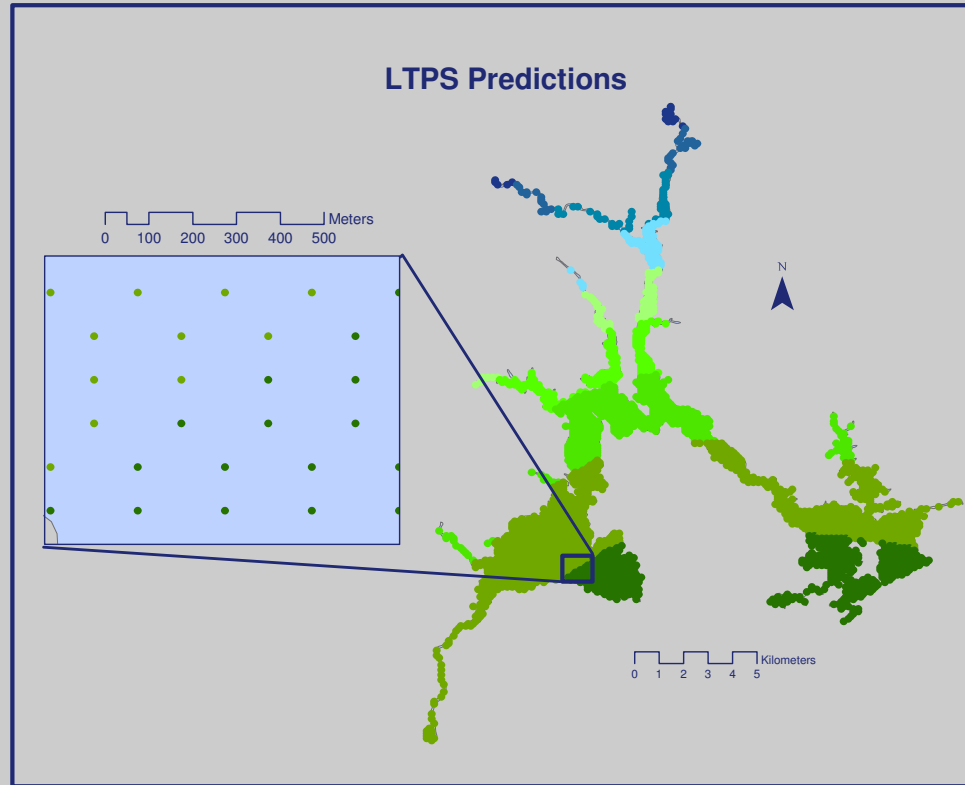
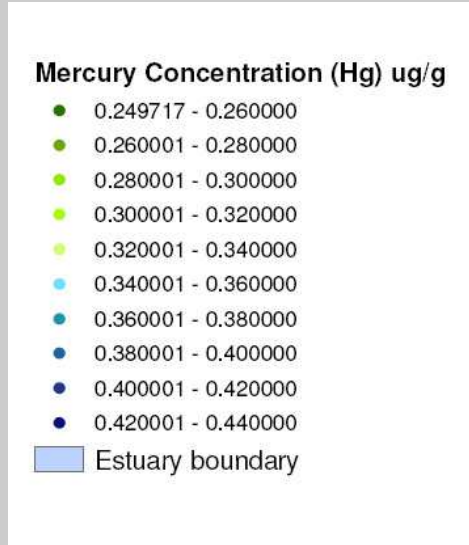
$$\hat{y} = X\hat{\beta} + Z\hat{u}$$

The `Spplus` commands for this would be simply

```
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

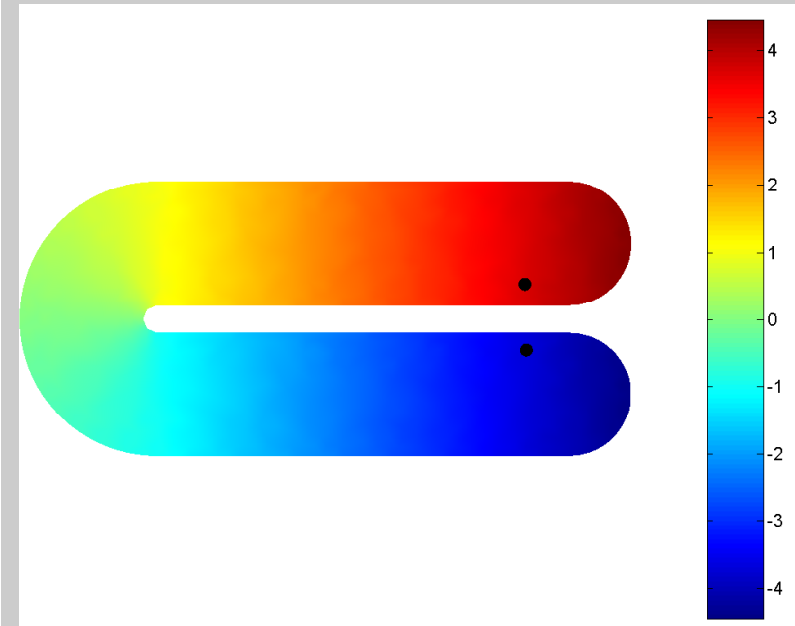
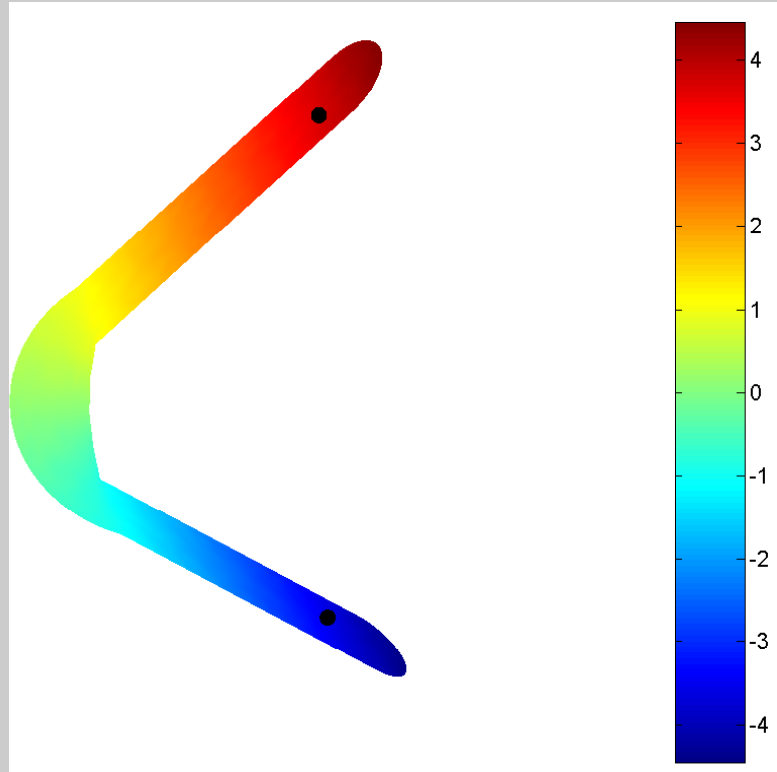
LTPS fit of the estuary data



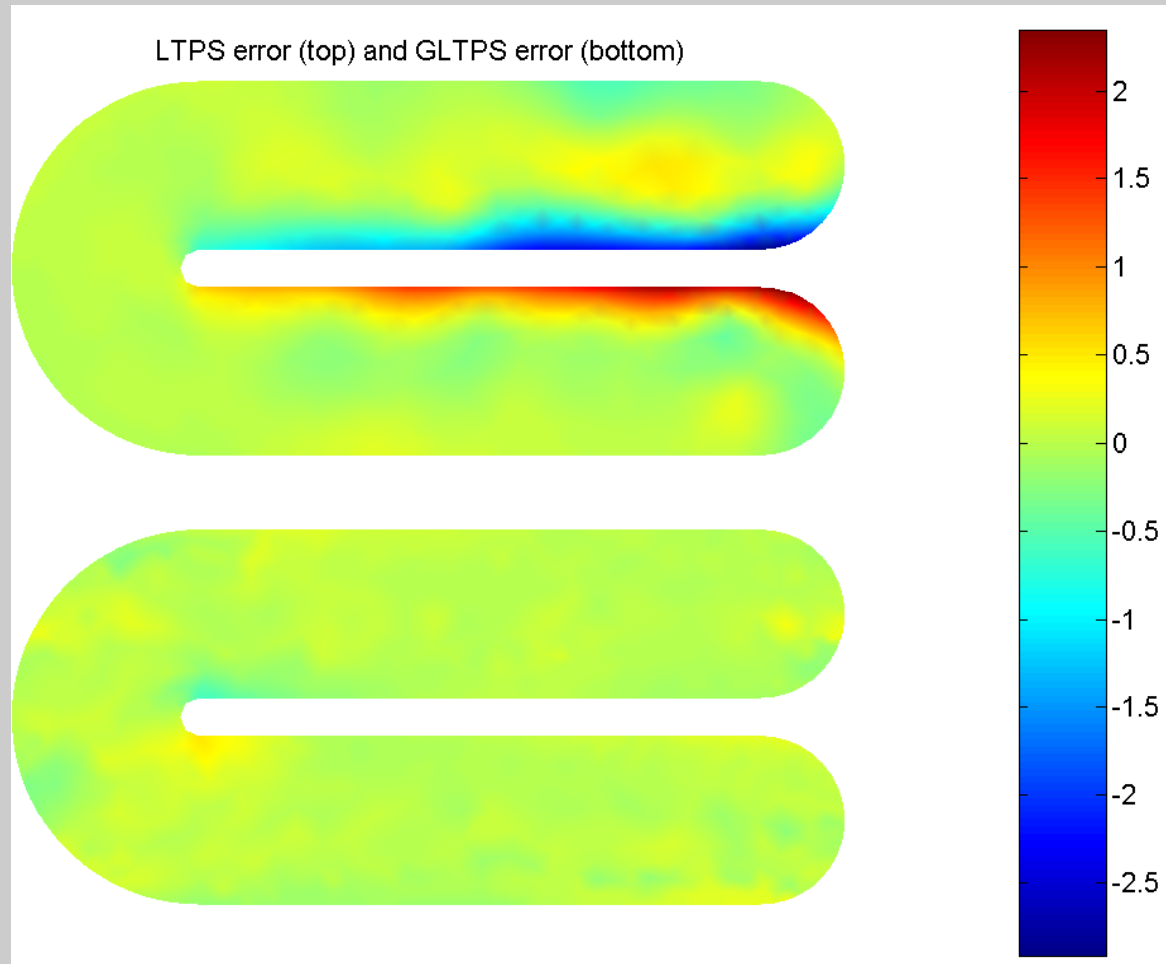
Recall Needs and Problems 3

Do we really need a different distance metric??

Monte Carlo simulation: real function



Simulation results: average prediction error



Recall estuary 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

Geodesic LTPS

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

$$\mathbf{Z} = \left[C(\|\mathbf{x}_i, \boldsymbol{\kappa}_t\|_G) \right]_{\substack{1 \leq i \leq N \\ 1 \leq t \leq T}} \left[C(\|\boldsymbol{\kappa}_t, \boldsymbol{\kappa}_{t'}\|_G) \right]_{1 \leq t, t' \leq T}^{-1/2},$$

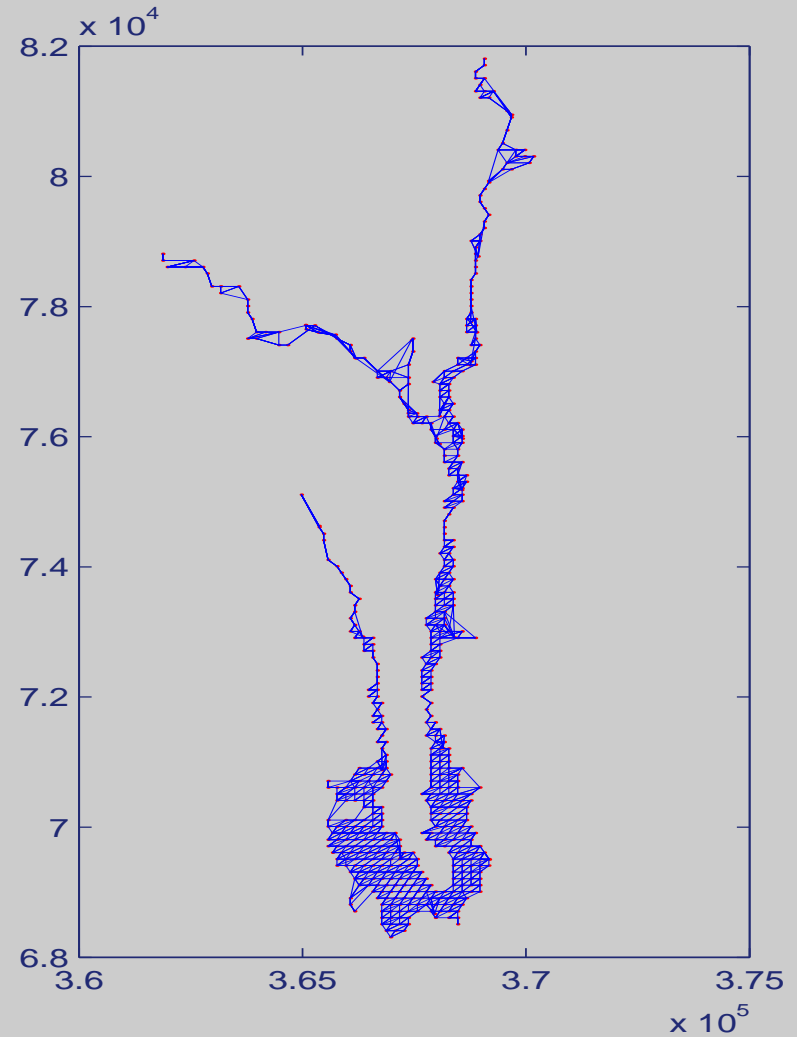
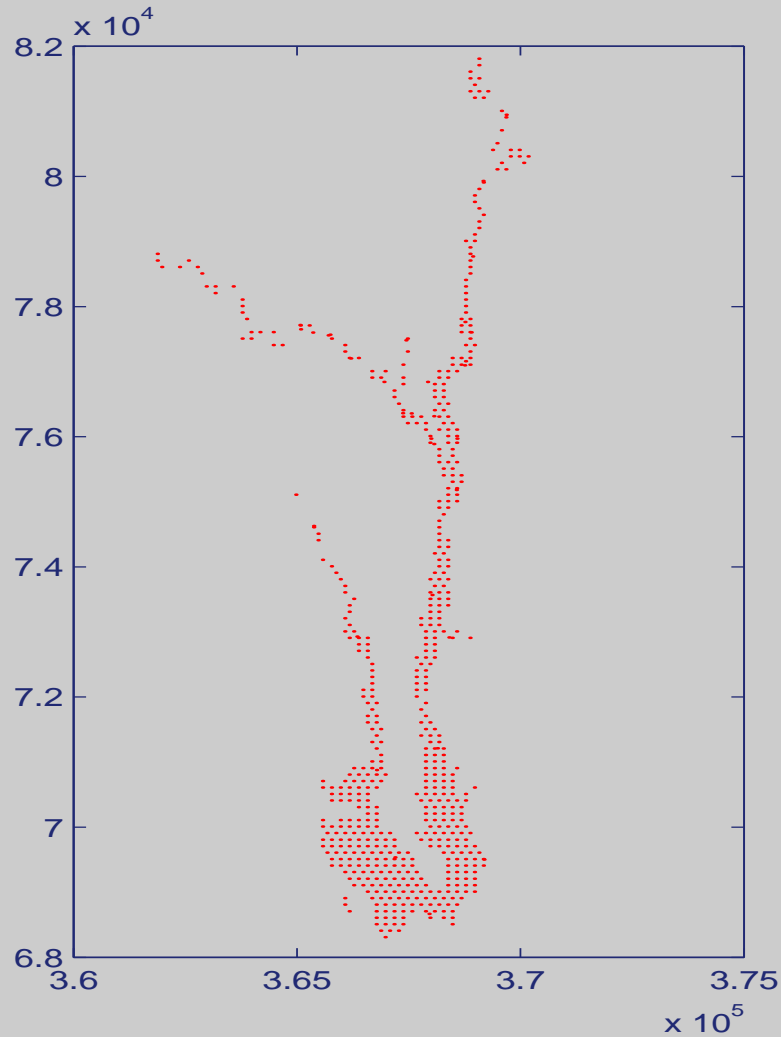
Geodesic LTPS

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

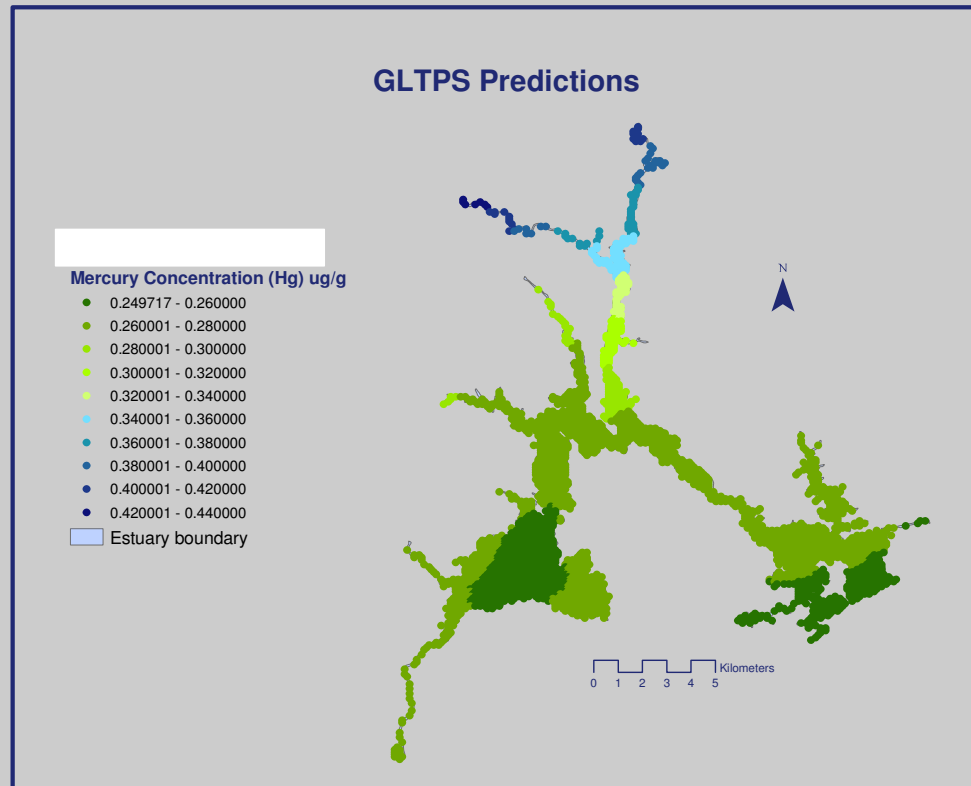
$$\mathbf{Z} = \left[C(\|\mathbf{x}_i, \boldsymbol{\kappa}_t\|_G) \right]_{\substack{1 \leq i \leq N \\ 1 \leq t \leq T}} \left[C(\|\boldsymbol{\kappa}_t, \boldsymbol{\kappa}_{t'}\|_G) \right]_{1 \leq t, t' \leq T}^{-1/2},$$

- 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.
- Small $nn \rightarrow$ the Graph might not be connected; too big $nn \rightarrow$ Euclidean distance.
 \Rightarrow take the smallest nn for which the Graph is connected.
- The density of the data influences the final estimate.

Floyd Algorithm for the Northern part of the estuary



GLTPS fit of the estuary



	GLTPS	
	estimate	p-value
Intercept	0.164	< 0.001
F(Year ₀₃)	0.090	0.011
σ_u	0.0016	0.011

r

AIC of models and standard deviations of predictions

X=1, lat, lon, year AIC=-55.258

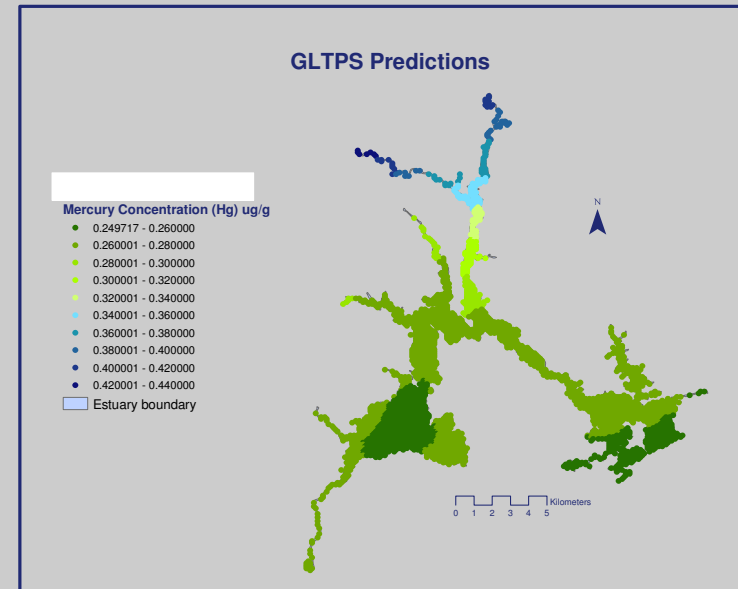
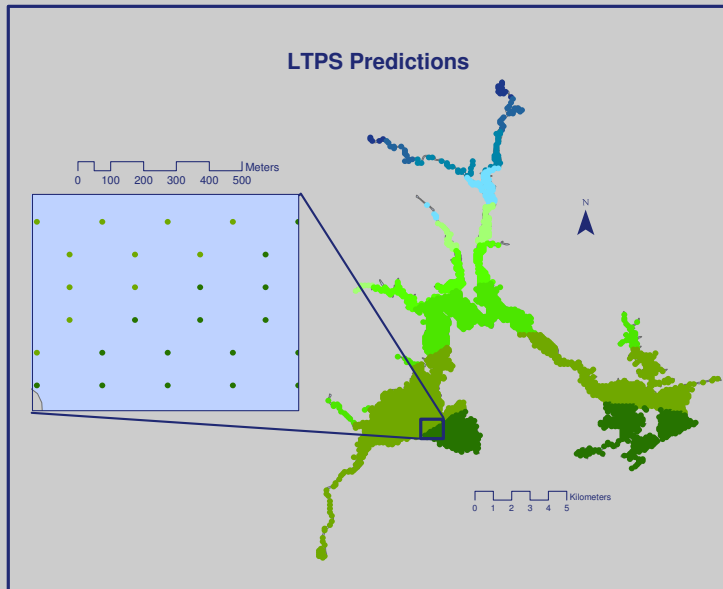
X=1, lat, lon AIC=-55.122

X=1, year AIC=-72.477

AIC: smaller is better

	Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.
observed HG	0.010	0.110	0.210	0.232	0.300	0.766
pred at obs locations	0.163	0.172	0.186	0.232	0.267	0.399
stdev at obs locations	0.006	0.011	0.020	0.018	0.021	0.031
pred at nonobs locations	0.250	0.260	0.262	0.271	0.268	0.436
stdev at nonobs locations	0.148	0.148	0.148	0.148	0.148	0.152

LTPS vs GLTPS fit of the estuary data



- Airborne deposition vs different patterns
- Great Bay and Cocheco River
- Recall estuary 2

The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;

The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;
2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances (and eventually predictions)

The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;
2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances (and eventually predictions)
3. Estimate the matrix of geodesic distances from this grid with the Floyd Algorithm starting with $nn = 3$ and then increasing nn until the graph is connected

The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;
2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances (and eventually predictions)
3. Estimate the matrix of geodesic distances from this grid with the Floyd Algorithm starting with $nn = 3$ and then increasing nn until the graph is connected
4. Calculate the \mathbf{X}_P and the \mathbf{Z}_P matrices for all the grid points

The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;
2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances (and eventually predictions)
3. Estimate the matrix of geodesic distances from this grid with the Floyd Algorithm starting with $nn = 3$ and then increasing nn until the graph is connected
4. Calculate the \mathbf{X}_P and the \mathbf{Z}_P matrices for all the grid points
5. Obtain the \mathbf{X} and the \mathbf{Z} matrices for the observed data locations as a subset of rows of \mathbf{X}_P and \mathbf{Z}_P (time saver and more accurate)

The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;
2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances (and eventually predictions)
3. Estimate the matrix of geodesic distances from this grid with the Floyd Algorithm starting with $nn = 3$ and then increasing nn until the graph is connected
4. Calculate the \mathbf{X}_P and the \mathbf{Z}_P matrices for all the grid points
5. Obtain the \mathbf{X} and the \mathbf{Z} matrices for the observed data locations as a subset of rows of \mathbf{X}_P and \mathbf{Z}_P (time saver and more accurate)
6. Fiddle with `lme` models. Issues:
 - other covariates: size of \mathbf{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 or bootstrap for more than one).

Outline

1. Low-rank smoothing splines for complex domains

- Motivation
- Low-rank thin plate splines
- Geodesic Low-rank thin plate splines

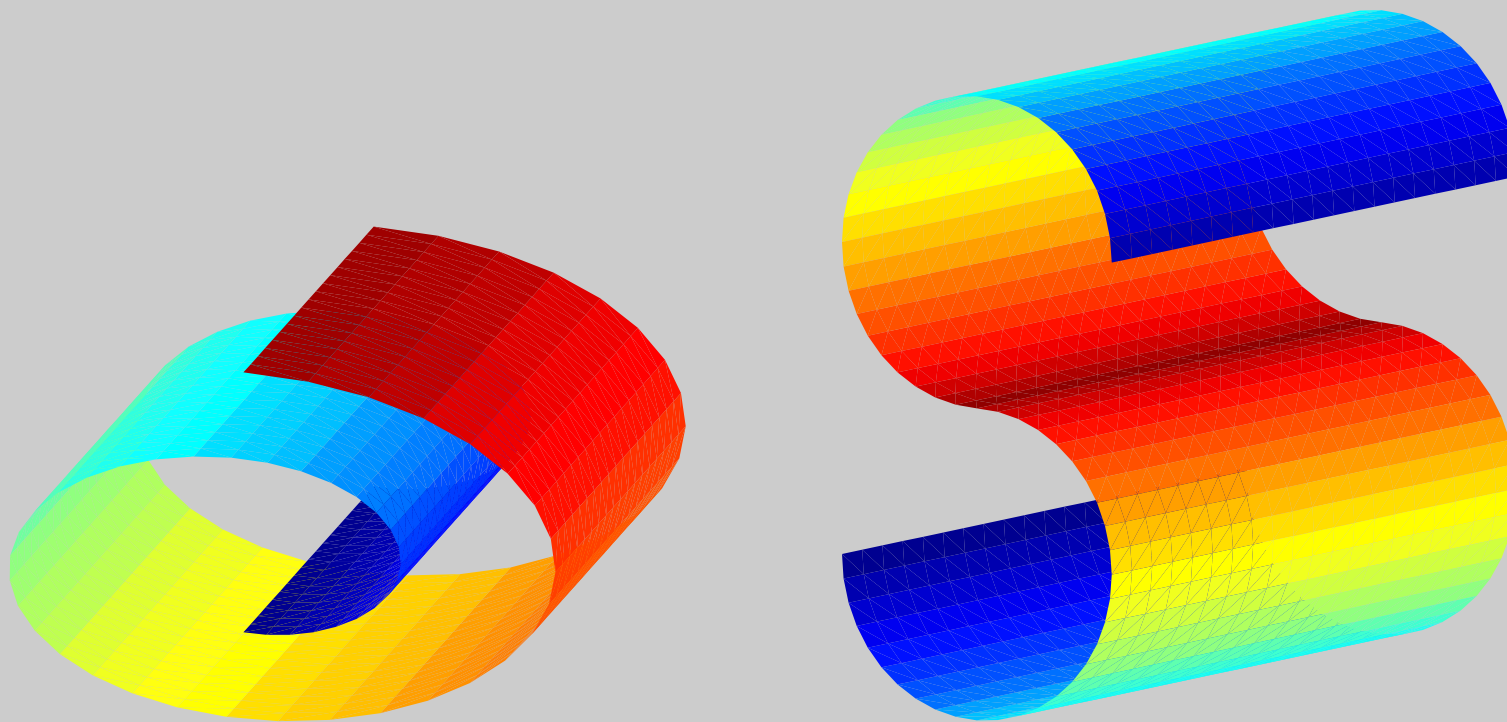
2. Low-rank smoothing splines for manifold recovery

- Motivation
- Examples of nonfunctional manifolds
- How essentially the same idea employed in 1. can be used in these situations

Motivation: data mining framework

- Understand the structure of large high dimensional data
- Handle nonlinear structures (nonlinear manifolds)
- Visualization in lower dimensional spaces
- Classification and clustering – Image recognition

Manifold recovery: Swiss-roll and S examples



NEW Needs and Problems

NEW Needs and Problems

NEEDS –

- Recover the manifold shape from noisy data

NEW Needs and Problems

NEEDS –

- Recover the manifold shape from noisy data

PROBLEMS –

- Bivariate smoothers run into difficulties when the manifold is not a functional relationship of the locations, see model (1).
- Both Euclidean and geodesic distance on the domain are not of use for shape recovery

Proposed approach: 3LTPS and 3GLTPS

Recall the key modification employed in GLTPS: 12.

Proposed approach: 3LTPS and 3GLTPS

Recall the key modification employed in GLTPS: 12.

Use the Euclidean or the geodesic distance in the manifold space, instead of the domain space:

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

where $z_i = (x_i, y_i)$ and knots are chosen on the 3d space. Assume for the moment that we can calculate both the Euclidean and the geodesic distance on the manifold, $n=3$

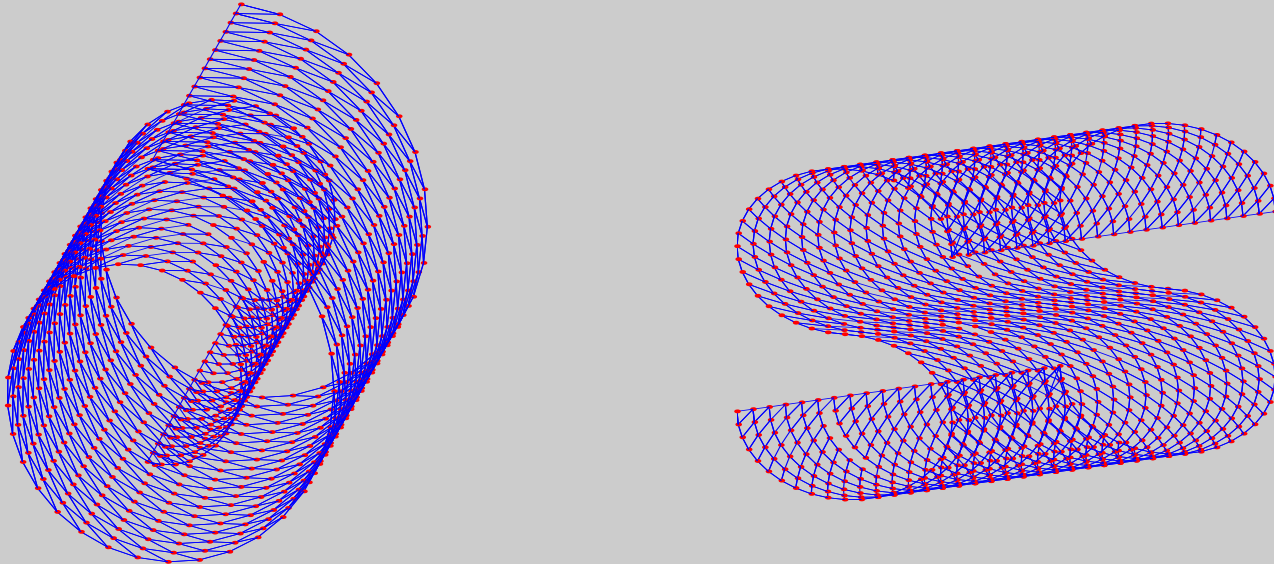
Proposed approach: 3LTPS and 3GLTPS

Recall the key modification employed in GLTPS: 12.

Use the Euclidean or the geodesic distance in the manifold space, instead of the domain space:

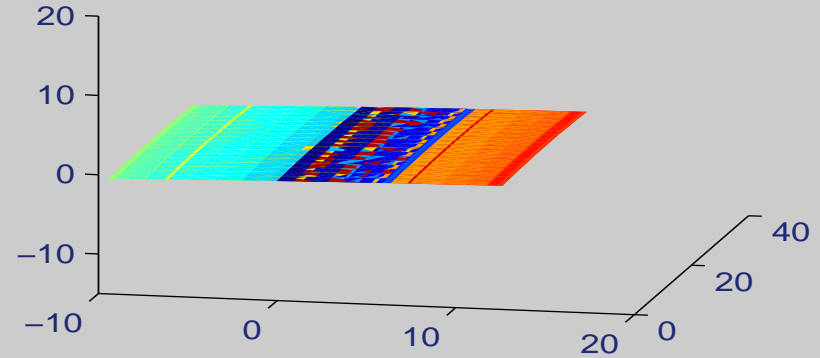
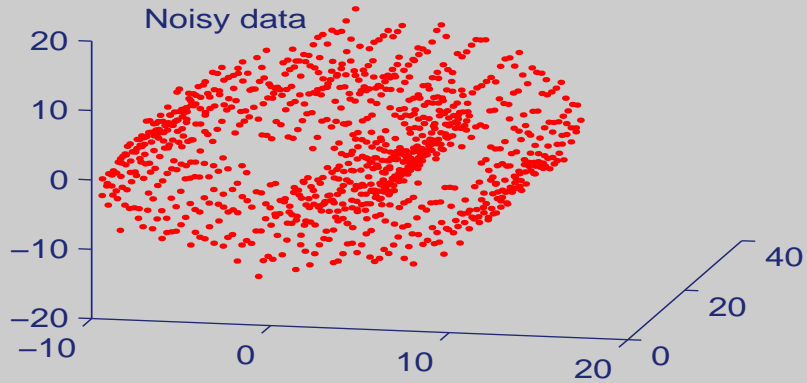
$$\mathbf{Z} = \left[C(\|z_i, \kappa_t\|_{E,G}) \right]_{\substack{1 \leq i \leq N \\ 1 \leq t \leq T}} \left[C(\|\kappa_t, \kappa_{t'}\|_{E,G}) \right]_{1 \leq t, t' \leq T}^{-1/2},$$

where $z_i = (x_i, y_i)$ and knots are chosen on the 3d space. Assume for the moment that we can calculate both the Euclidean and the geodesic distance on the manifold, $n=3$

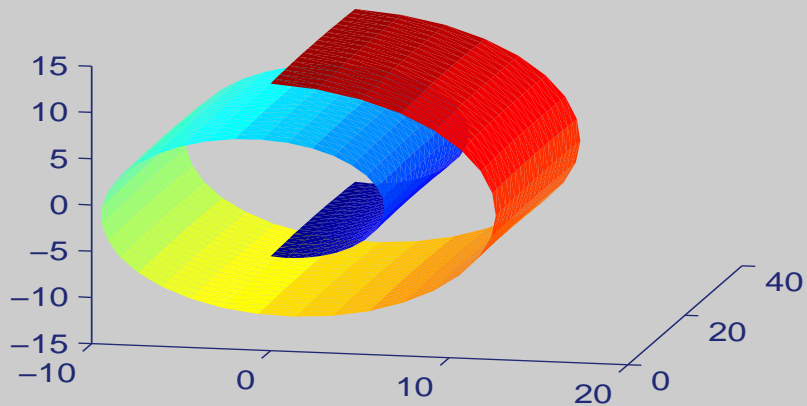


3LTPS and 3GLTPS fit of the Swiss roll

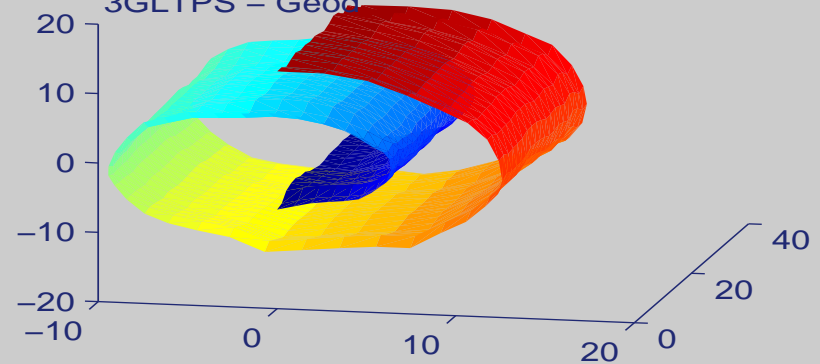
LTPS – Eucl



3LTPS – Eucl

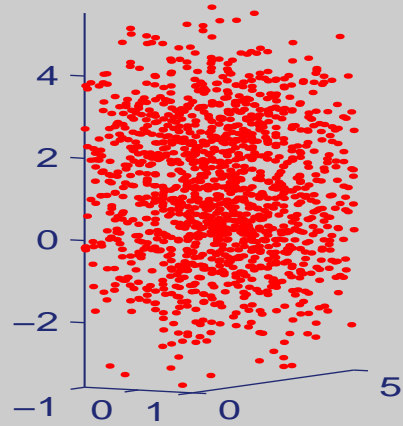


3GLTPS – Geod

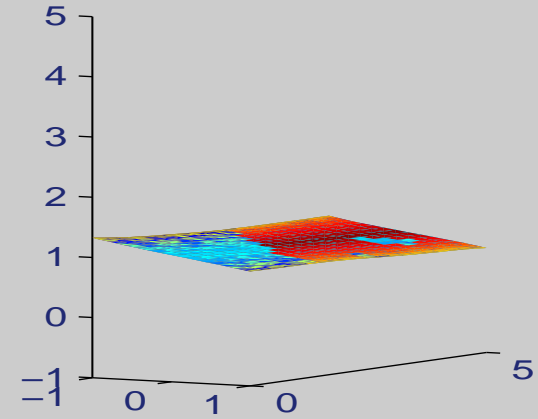


3GLTPS fit of the S manifold

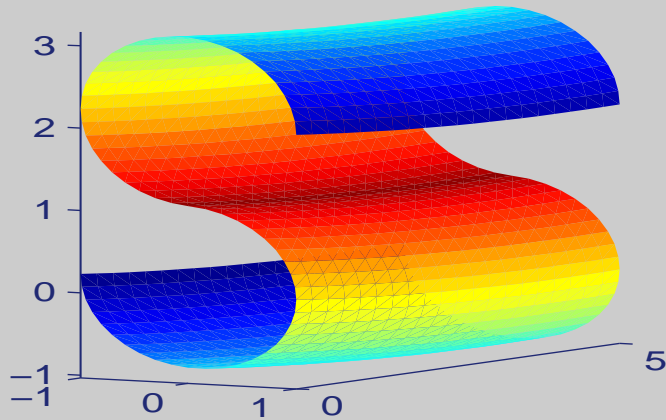
Noisy data



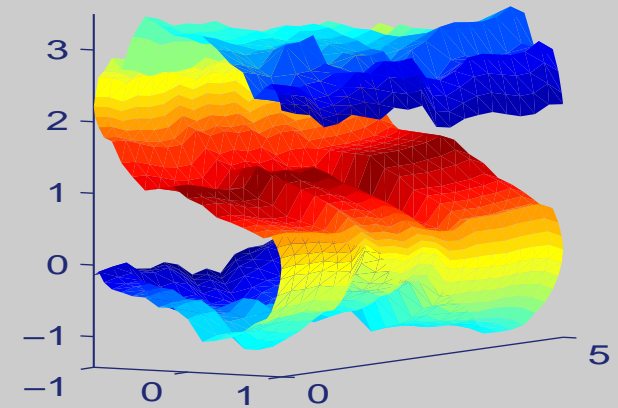
LTPS



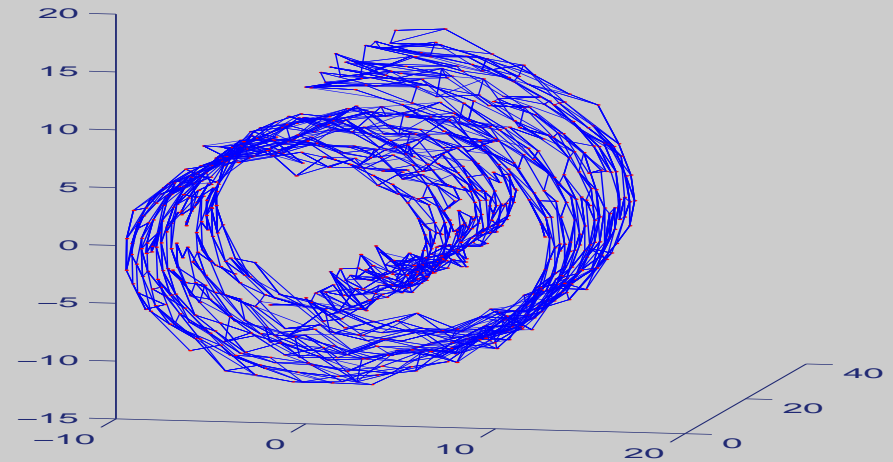
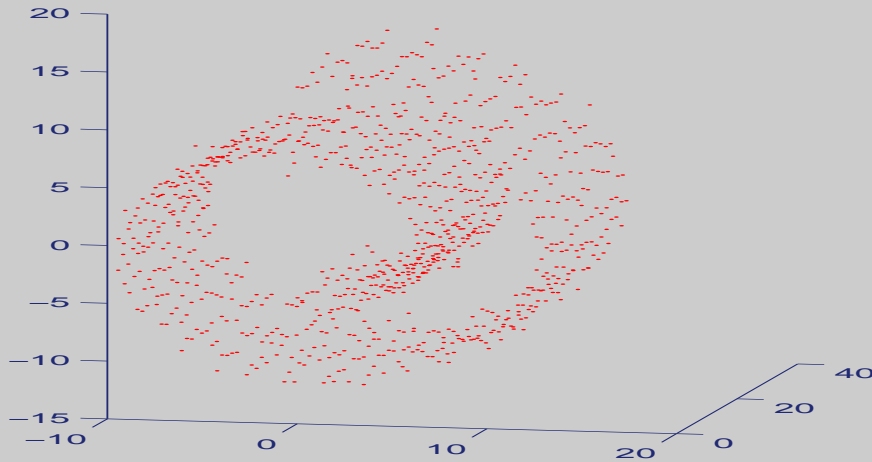
3LTPS – Eucl



3GLTPS – Geo



A key issue: estimating 3d distances from noisy data



HIGH QUALITY DATA We are still ok, the estimates are not as smooth as before, but still reasonable

HIGH NOISE Unreliable and wiggly estimates → we are now trying to work out estimates of these distances by smoothing the output of the Floyd algorithm.

Wrap up

GLTPS

1. It is possible to account for non regular domains when mapping quantities of interest
2. The LTPS framework allows inserting other covariates available for all prediction locations in a parametric or nonparametric way easily and guarantees positive definite *covariance* functions
3. Applications other than estuaries include stream networks, domains with holes and irregular boundaries (lakes with islands/land with lakes), response over a nonflat domain (measurements on mountains) ...
4. Other distance measures can be employed

Wrap up

GLTPS

1. It is possible to account for non regular domains when mapping quantities of interest
2. The LTPS framework allows inserting other covariates available for all prediction locations in a parametric or nonparametric way easily and guarantees positive definite *covariance* functions
3. Applications other than estuaries include stream networks, domains with holes and irregular boundaries (lakes with islands/land with lakes), response over a nonflat domain (measurements on mountains) ...
4. Other distance measures can be employed

3LTPS and 3GLTPS: Nonfunctional manifolds can be recovered if observational data is high quality or, if not, if we have a priori information about the manifold.

Current and future work

GLTPS Add remote sensing covariates to the modeling of mercury in NH estuary (land cover) —
altern: multiphase sampling for less expensive covariates to be collected on site (grain size)

GLTPS Application on stream networks

GLTPS Fix the Floyd Algorithm to allow only for on-water paths and for flow direction

Current and future work

GLTPS Add remote sensing covariates to the modeling of mercury in NH estuary (land cover) —
altern: multiphase sampling for less expensive covariates to be collected on site (grain size)

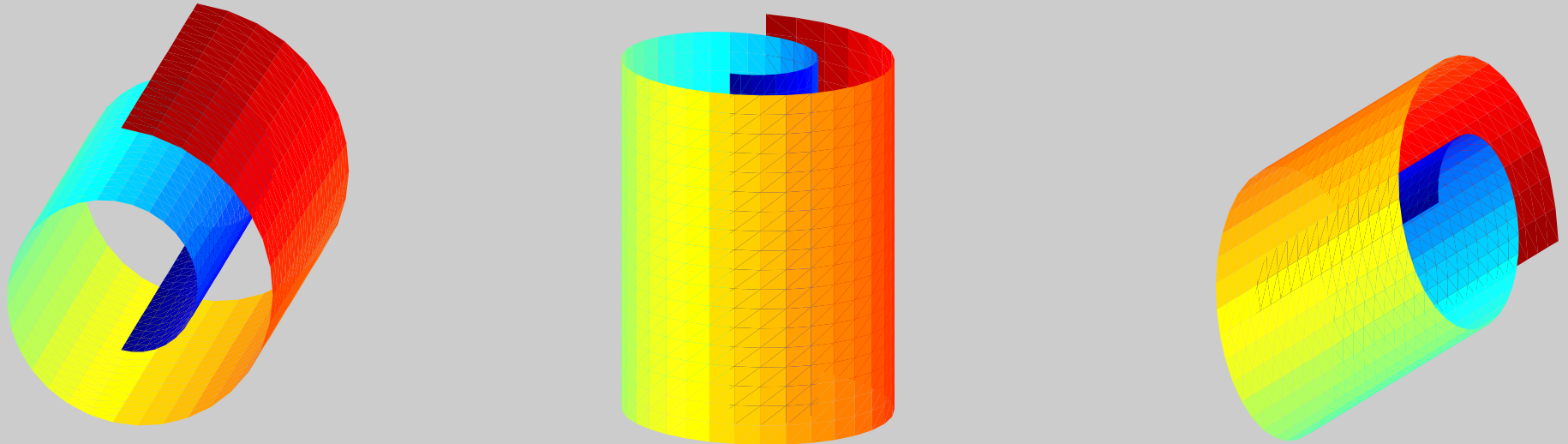
GLTPS Application on stream networks

GLTPS Fix the Floyd Algorithm to allow only for on-water paths and for flow direction

3GLTPS Obtain reliable estimates of distances from high noise level data

3GLTPS Application to higher dimensional good quality datasets

3GLTPS Manifold recovery under rotation



Essential bibliography

- Crainiceanu, C. and Ruppert, D. (2004), Likelihood ratio tests in linear mixed models with one variance component, *J.R.S.S.– B*, **66**, 165–185.
- Gardner, B., Sullivan, P.J. and Lembo, A.J.Jr (2003), Predicting stream temperatures: geostatistical model comparison using alternative distance metrics, *Can. J. Fish. Aquat. Sci.*, **60**, 344–351.
- Rathbun, S.L. (1998), Spatial modelling in irregularly shaped regions: kriging estuaries, *Environmetrics*, **9**, 109–129.
- Ruppert, D., Wand, M. P. and Carroll, R. (2003), *Semiparametric Regression*. Cambridge University Press, Cambridge, New York.
- Ver Hoef, J.M., Peterson, E. and Theobald D. (2004), Spatial statistical models that use flow and stream distance *Manuscript*.

ACKNOWLEDGEMENTS: Hal Walker (EPA) and Phil Townbridge (NHDES) for providing the NH data, Jay Breidt for helpful discussion. 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.