

EPA / AED – March 1, 2005

STARMAP

## Low-rank smoothing splines on complex domains

M. Giovanna Ranalli

Department of Statistics, Colorado State University

Joint work with Jay Breidt and Haonan Wang, CSU

Slide number 0

STARMAP

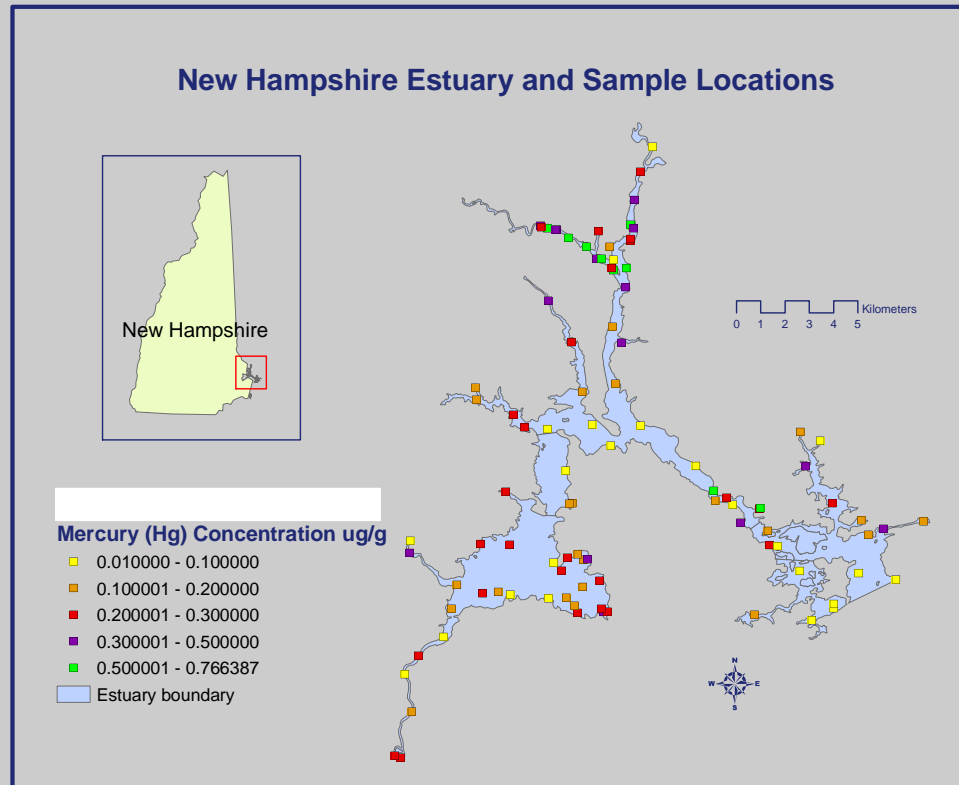
# Slide number 0

STARMAP

Given our very different backgrounds  
PLEASE feel free to stop me as soon as you think  
*What is she talking about????*  
being it statistics, biology, ecology, my accent, etc...

# Challenging problem

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  $i = 1, \dots, N$  ( $N=97$  for the estuary), with  $\mathbf{x}_i$  geographical locations and  $y_i$  measurements of a variable of interest (HG for the estuary). 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  $i = 1, \dots, N$  ( $N=97$  for the estuary), with  $\mathbf{x}_i$  geographical locations and  $y_i$  measurements of a variable of interest (HG for the estuary). 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  $C(r) = r^2 \log r$

# 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  $C(r) = r^2 \log r$

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  $C(r) = r^2 \log r$

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

---

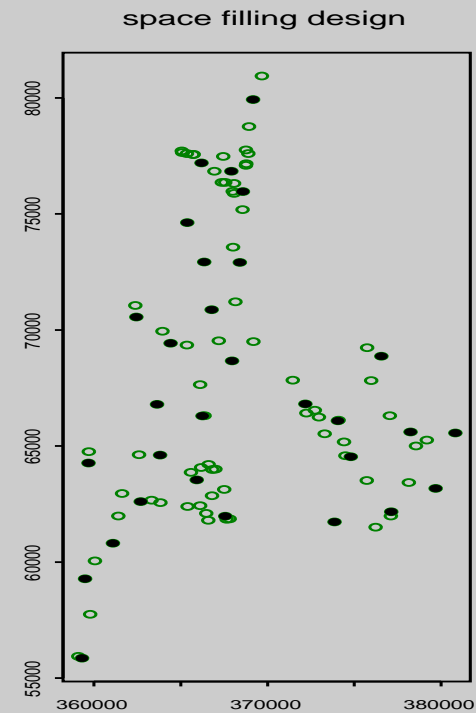
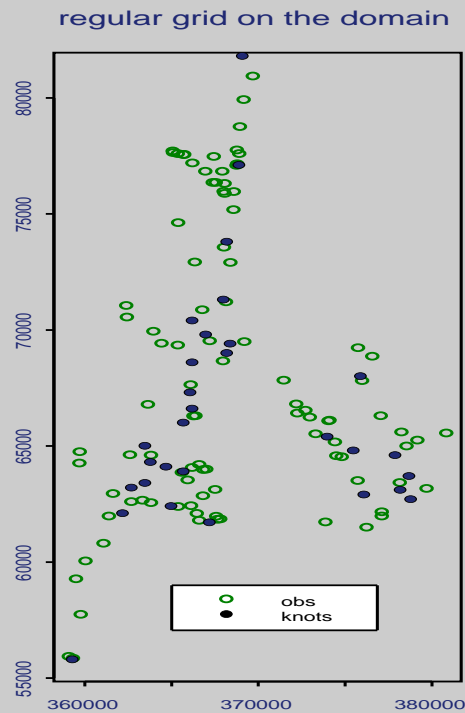
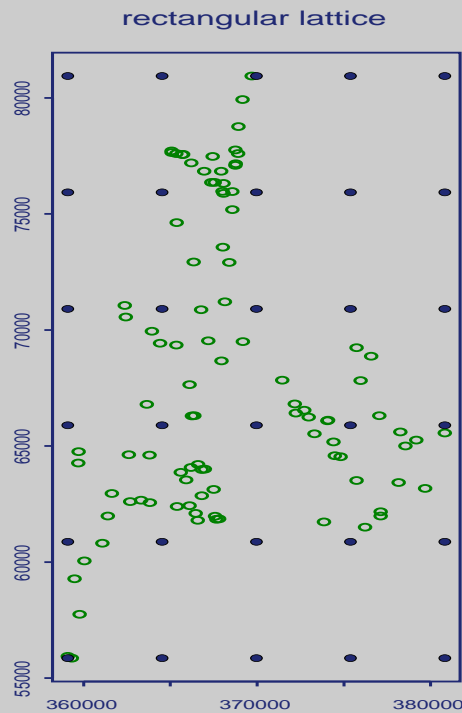
**HOW MANY** rule of thumb: 1 every 3-4 observations

**WHERE** rectangular lattices, regular grids on the domain or space filling designs  
(FUNFITS in Spplus 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 SpPlus 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 `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  $\beta$  and predictions for  $u$  are obtained through Maximum Likelihood or REstricted ML, predicted values at observed locations are given by

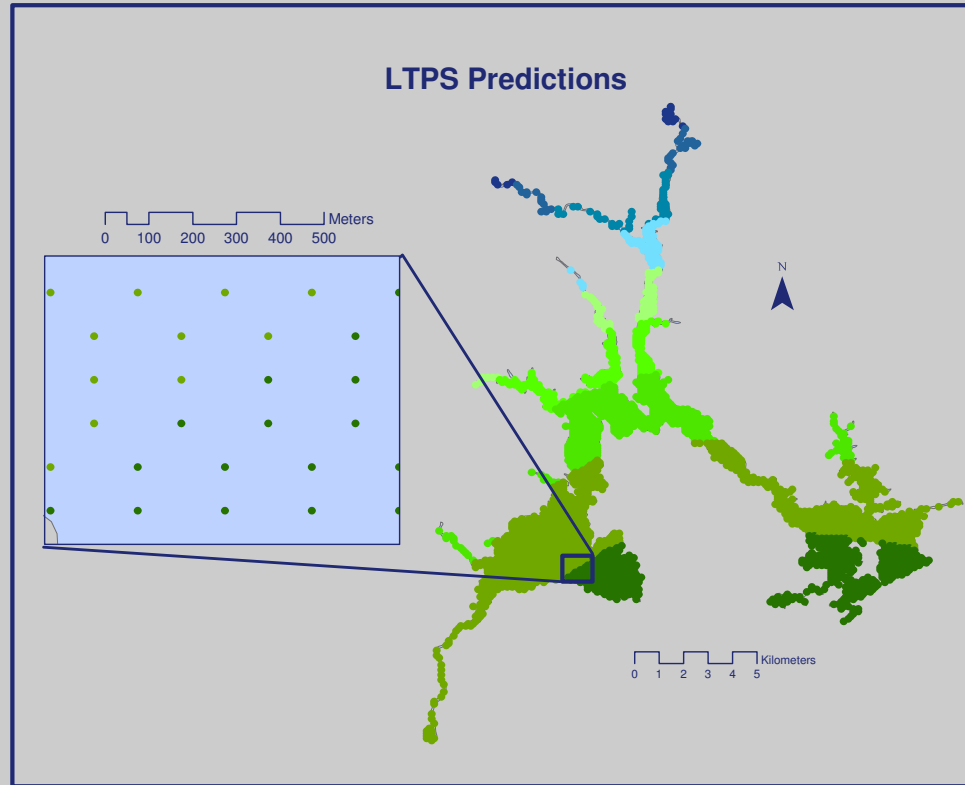
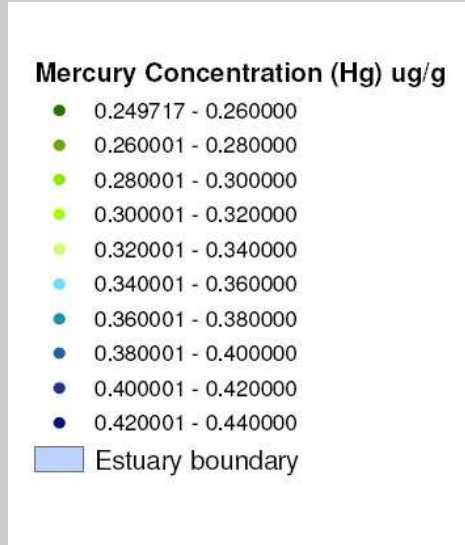
$$\hat{y} = X\hat{\beta} + Z\hat{u}$$

The `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

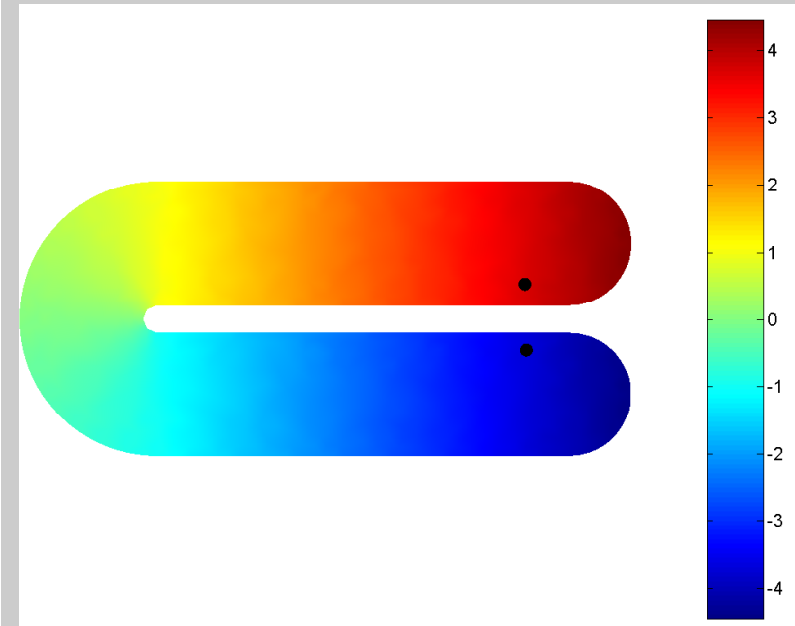
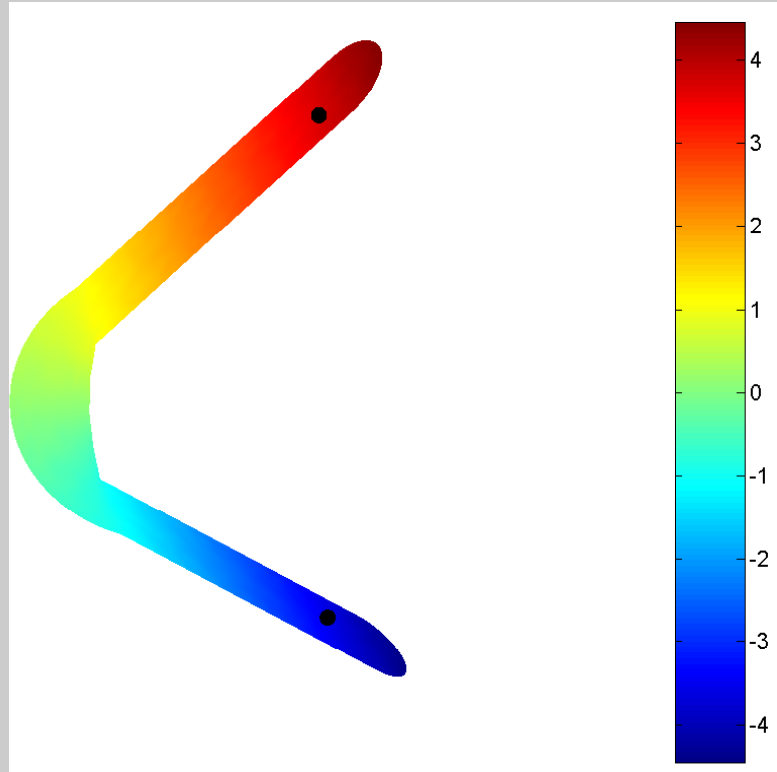
# LTPS fit of the estuary data



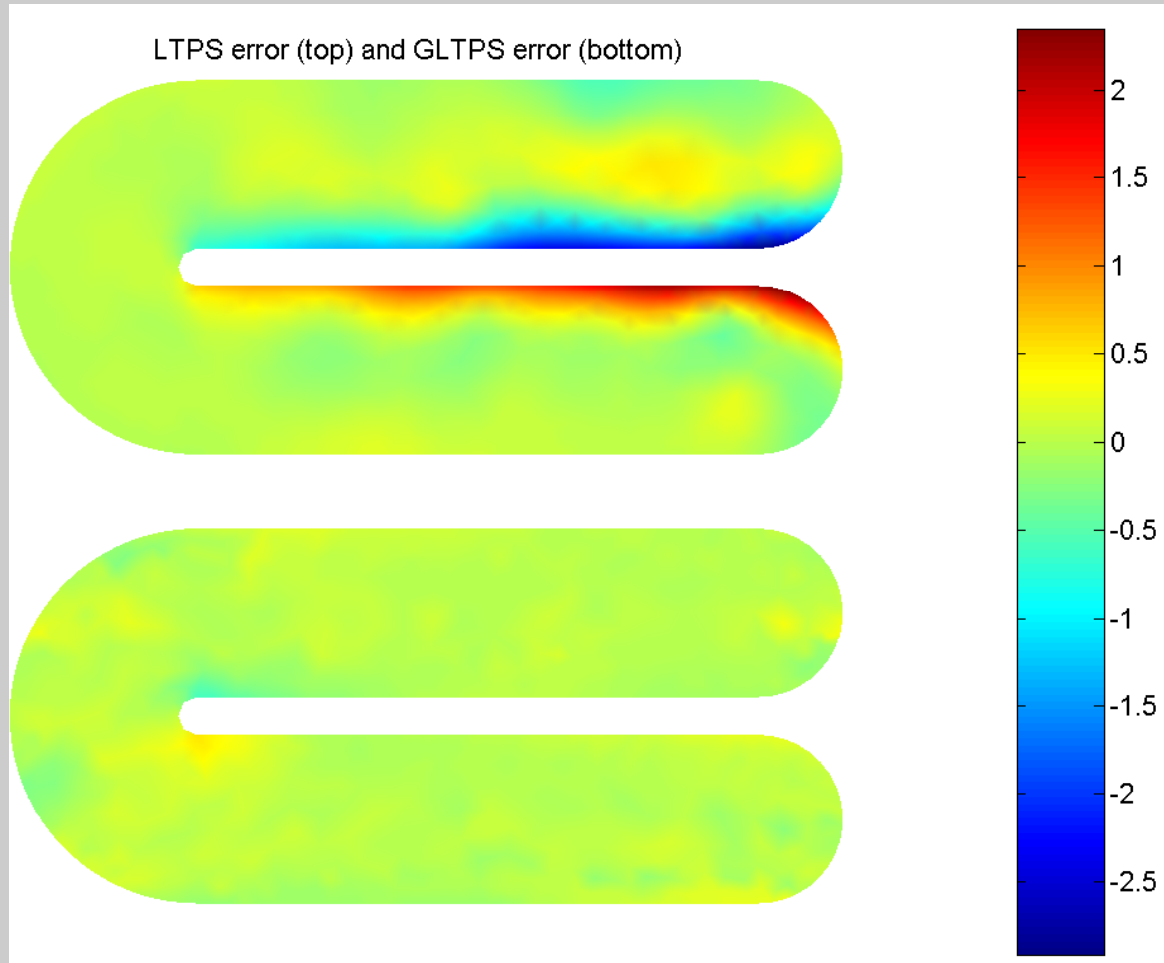
Recall Needs and Problems 3

# 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

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

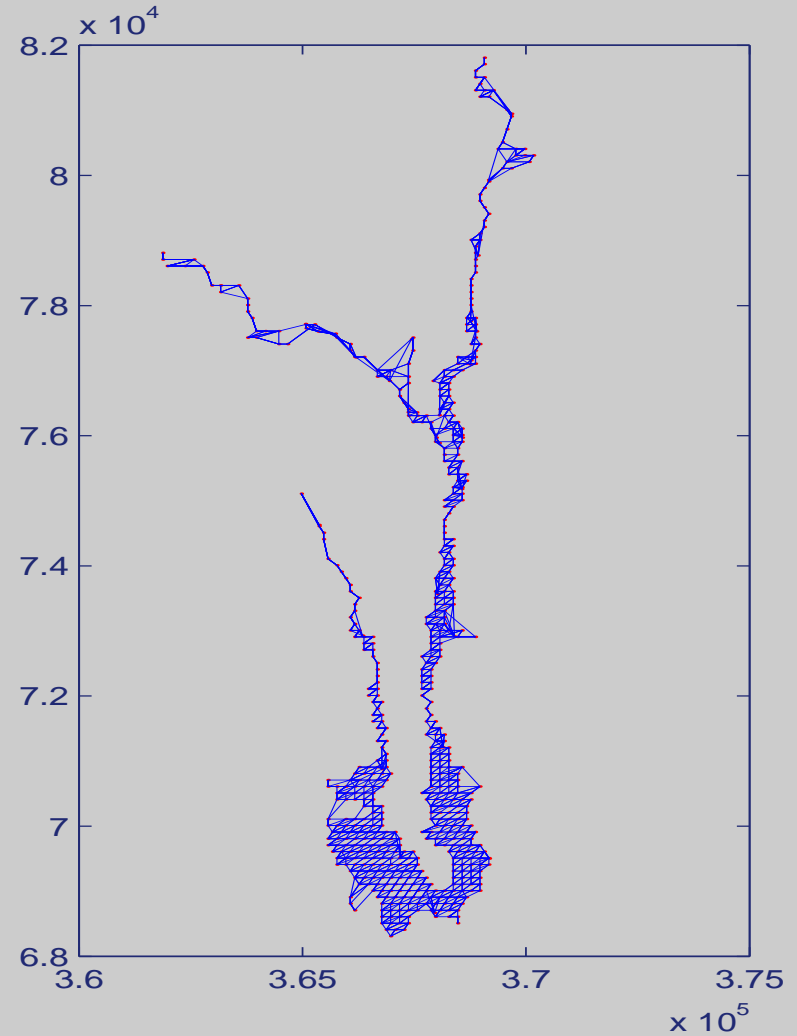
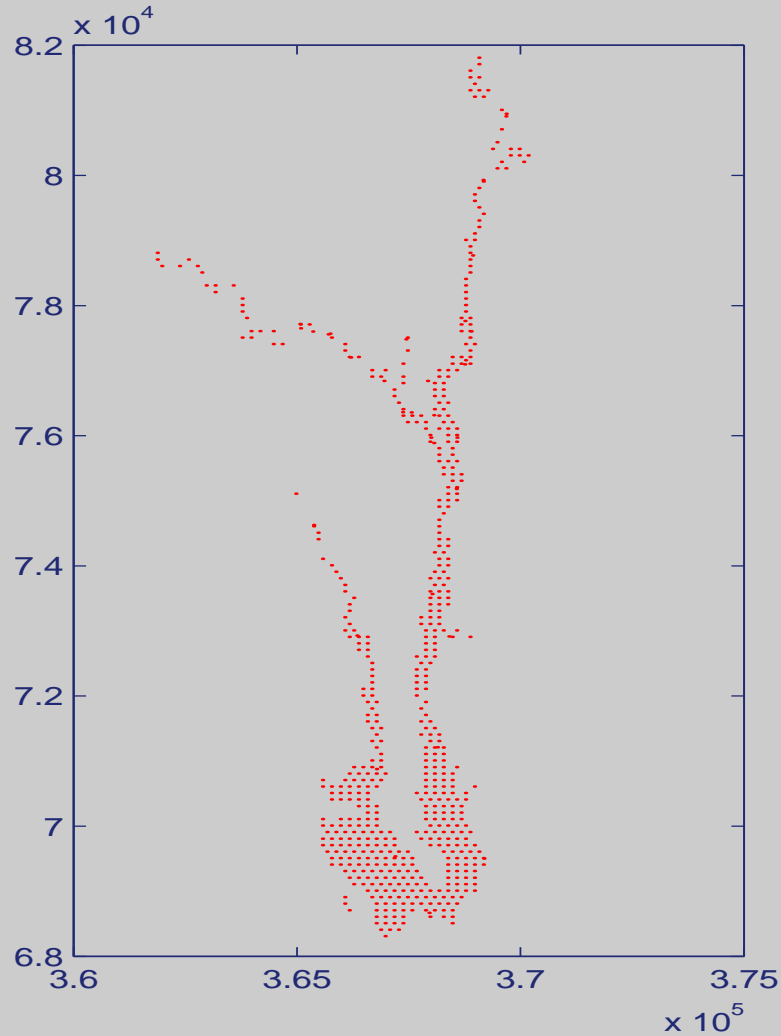
# Geodesic LTPS

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

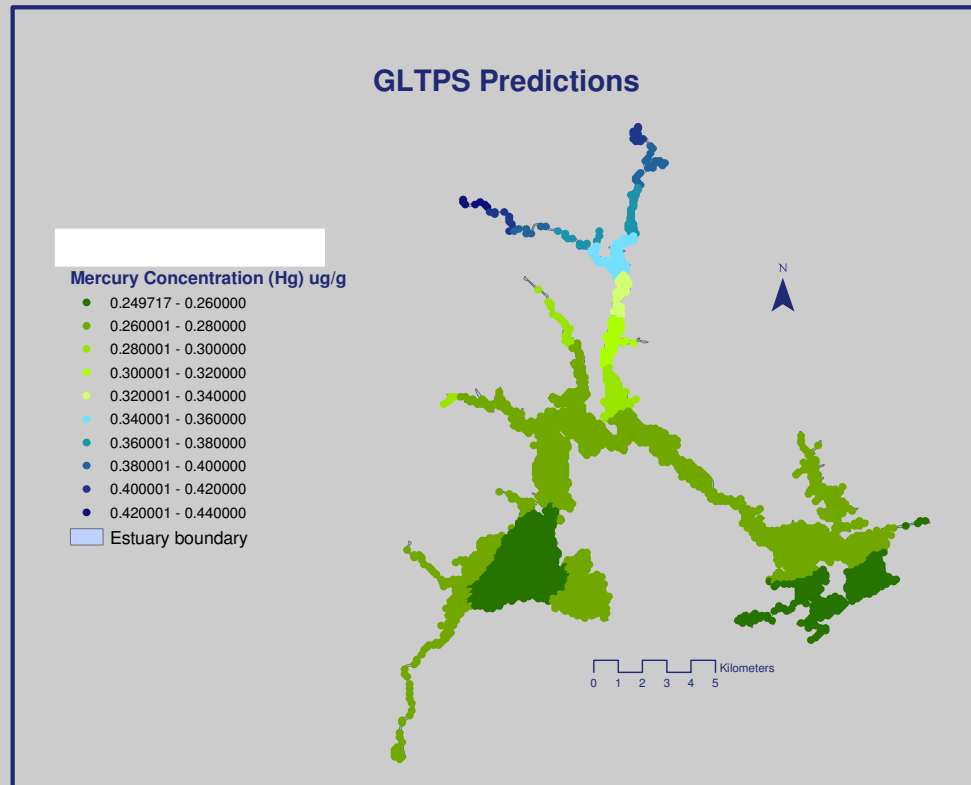
$$\mathbf{Z} = \left[ C(\|\mathbf{x}_i, \boldsymbol{\kappa}_t\|_G) \right]_{\substack{1 \leq i \leq N \\ 1 \leq t \leq T}} \left[ C(\|\boldsymbol{\kappa}_t, \boldsymbol{\kappa}_{t'}\|_G) \right]_{1 \leq t, t' \leq T}^{-1/2},$$

The geodesic distance is estimated by means of the Floyd Algorithm. It estimates it by summing the Euclidean distances between locations along a constructed path. This path is obtained by determining for each location the set of  $nn$  nearest neighbors and linking the locations in an increasing fashion. For small values of  $nn$  this kind of *net* might not be connected, for  $nn$  too big it could provide the Euclidean distance. Our advice would be to take the shortest  $nn$  for which the net 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 <sub>03</sub> ) | 0.090    | 0.011   |
| $\sigma_u$             | 0.0016   | 0.011   |

r

# AIC of models and standard deviations of predictions

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

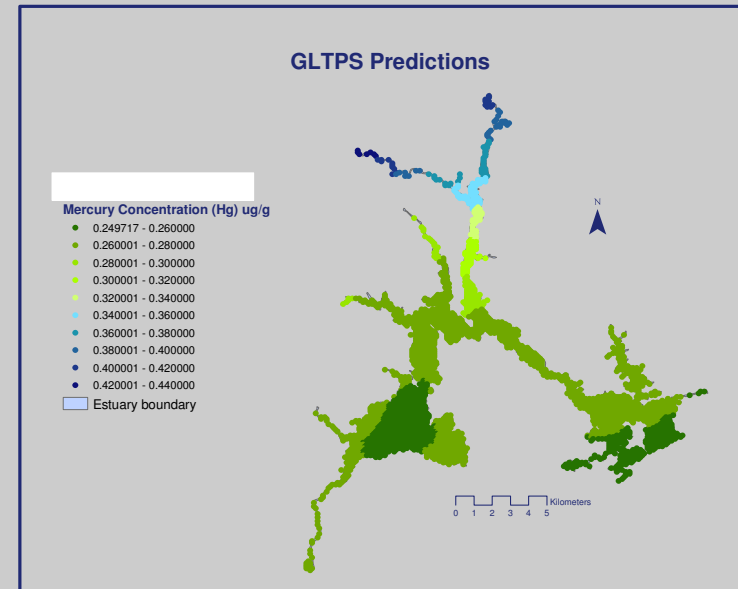
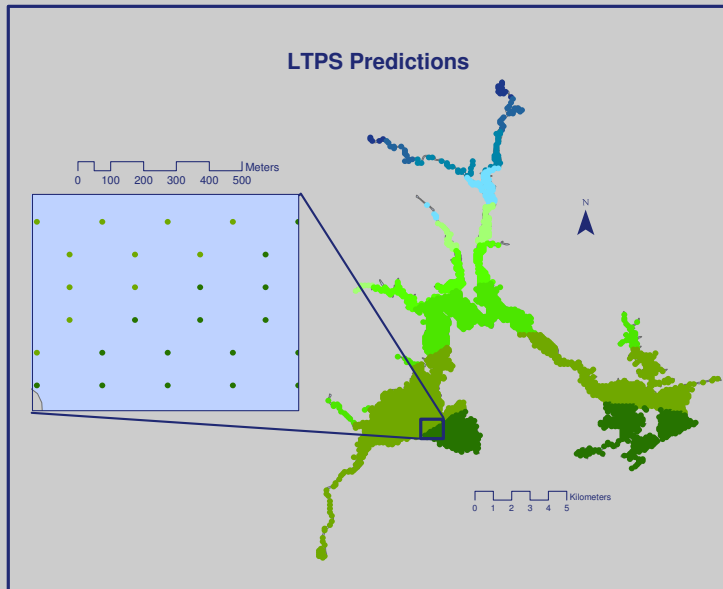
X=1,lat,lon AIC=-55.122

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

# My recipe – The steps to obtain GLTPS

---

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

# My recipe – 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)

# My recipe – 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 all the points are connected

# My recipe – 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 all the points are connected
4. Calculate the  $\mathbf{X}_P$  and the  $\mathbf{Z}_P$  matrices for all the grid points

# My recipe – 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 all the points are 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)

# My recipe – The steps to obtain GLTPS

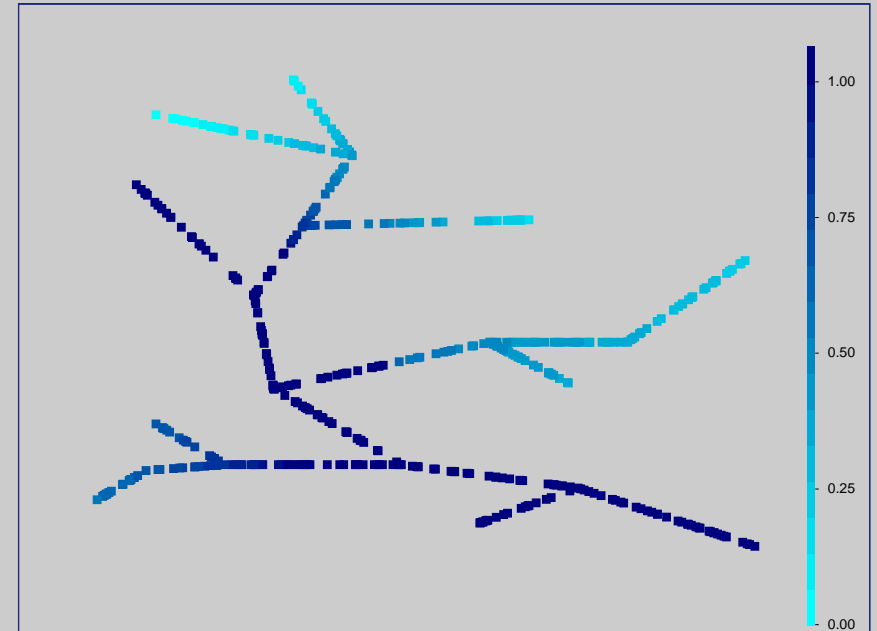
1. Determine number and location of the knots from the observed data locations through a space filling design;
2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances (and eventually predictions)
3. Estimate the matrix of geodesic distances from this grid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

# Another possible application – stream networks

The extension is straightforward but there are some issues:

**DISTANCES:** The Floyd Algorithm requires the data to be dense enough to work nicely. Dave Theobald at NREL-CSU is working on alternative algorithms to obtain stream network distances that can also account for flow to be included in ARCGIS.

**REAL DATA:** We are still in the *simulation* phase and looking for a real dataset with more than one observation for each stream segment.



# 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 framework allows inserting other covariates available for all prediction locations in a parametric or nonparametric way easily and guarantees positive definite \*covariance\* functions

# Wrap up

---

1. It is possible to account for non regular domains when estimating 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 and stream networks include domains with holes and irregular boundaries (lakes with islands/land with lakes), response over a nonflat domain (measurements on mountains) ...

# Wrap up

---

1. It is possible to account for non regular domains when estimating 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 and stream networks include 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

# Current and future work

---

1. Inserting other covariates to better model the NH estuary

# Current and future work

---

1. Inserting other covariates to better model the NH estuary
2. Find some real data on stream networks

# Current and future work

---

1. Inserting other covariates to better model the NH estuary
2. Find some real data on stream networks
3. Application of GLTPS to shape recovery of functional and nonfunctional manifolds

# Current and future work

---

1. Inserting other covariates to better model the NH estuary
2. Find some real data on stream networks
3. Application of GLTPS to shape recovery of functional and nonfunctional manifolds
4. Account for complex designs

# Current and future work

---

1. Inserting other covariates to better model the NH estuary
2. Find some real data on stream networks
3. Application of GLTPS to shape recovery of functional and nonfunctional manifolds
4. Account for complex designs
5. Fix the Floyd Algorithm to allow only for on-water paths and for flow directions

# 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*.

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.