Spatial Statistical Models on Stream Networks

Jay Ver Hoef\textsuperscript{1}, Erin Peterson\textsuperscript{2}, Dan Isaak\textsuperscript{3}

\textsuperscript{1}NOAA National Marine Mammal Lab, Seattle, USA
\textsuperscript{2}CSIRO Mathematics, Informatics, and Statistics, Brisbane, Australia
\textsuperscript{3}USDA Forest Service, Boise, USA
Concepts

NCEAS Working Group
## Concepts

### Dual Spatial Coordinate System

<table>
<thead>
<tr>
<th>Dual</th>
<th>Network</th>
<th>2-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Dual Diagram]</td>
<td>![Network Diagram]</td>
<td>![2-D Diagram]</td>
</tr>
</tbody>
</table>

Ver Hoef, Peterson, Isaak  
NOAA/NMML, CSIRO/CMIS, USDA/FS
Concepts

Classification of Analytical Methods

River Network

OFF

ABOUT

ON

OVER

ACROSS

Ver Hoef, Peterson, Isaak

NOAA/NMML, CSIRO/CMIS, USDA/FS
Concepts

Spatial Linear Model

\[
\begin{pmatrix}
Z_{\text{observed}} \\
Z_{\text{unobserved}}
\end{pmatrix}
= X\beta + \varepsilon, \quad \text{var}(\varepsilon) = \Sigma(\theta)
\]
Introduction

Concepts

Stream Network Covariance

\[
\begin{pmatrix}
Z_{\text{observed}} \\
Z_{\text{unobserved}}
\end{pmatrix} = X\beta + \varepsilon, \quad \text{var}(\varepsilon) = \Sigma(\theta)
\]

Empirical and Fitted Covariance

Ver Hoef, Peterson, Isaak

NOAA/NMML, CSIRO/CMIS, USDA/FS
Concepts

Distance and Flow

Euclidean distance

Hydrologic distance

Hydrologic distance
Why Bother?

Better Predictions

Stream temperature data

<table>
<thead>
<tr>
<th>Model</th>
<th>Estimate (°C)</th>
<th>95% Prediction Interval (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kriging</td>
<td>11.98</td>
<td>10.26 - 13.70</td>
</tr>
<tr>
<td>Tail-up</td>
<td>12.94</td>
<td>10.29 - 15.58</td>
</tr>
</tbody>
</table>

Covariance Parameter | Euclidean | Tail-up |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nugget</td>
<td>0.292</td>
<td>0.000</td>
</tr>
<tr>
<td>Partial Sill</td>
<td>7.72</td>
<td>11</td>
</tr>
<tr>
<td>Range</td>
<td>16200</td>
<td>203300</td>
</tr>
</tbody>
</table>

Ver Hoef, Peterson, Isaak

NOAA/NMML, CSIRO/CMIS, USDA/FS
Problem:

Valid Covariance Matrices Using Stream Distance

Not to scale. All lengths = 1
Models

Time Series Moving Average Construction
Models

Time Series Moving Average Construction

Ver Hoef, Peterson, Isaak

NOAA/NMML, CSIRO/CMIS, USDA/FS
Models

Stream Network Moving Average Construction
Models

Tail-up Models

\[ C_u(r_i, s_j|\theta_u) = \begin{cases} 
\pi_{i,j} C_t(h|\theta_u) & \text{if } r_i \text{ and } s_j \text{ are flow-connected}, \\
0 & \text{if } r_i \text{ and } s_j \text{ are flow-unconnected}, 
\end{cases} \]

- Tail-Up Linear-with-Sill Model

\[ C_t(h|\theta_u) = \sigma_u^2 \left( 1 - \frac{h}{\alpha_u} \right) I \left( \frac{h}{\alpha_u} \leq 1 \right), \]

- Tail-Up Exponential Model

\[ C_t(h|\theta_u) = \sigma_u^2 \exp\left( -\frac{3h}{\alpha_u} \right), \]
Models

Tail-down Models

- Tail-Down Linear-with-Sill Model, $b \geq a \geq 0$,

$$C_d(a, b, h|\theta_d) = \begin{cases} 
\sigma_d^2 \left(1 - \frac{h}{\alpha_d}\right) I \left(\frac{h}{\alpha_d} \leq 1\right) & \text{if flow-connected,} \\
\sigma_d^2 \left(1 - \frac{b}{\alpha_d}\right) I \left(\frac{b}{\alpha_d} \leq 1\right) & \text{if flow-unconnected,}
\end{cases}$$

- Tail-down Exponential Model,

$$C_d(a, b, h|\theta_d) = \begin{cases} 
\sigma_d^2 \exp\left(-\frac{3h}{\alpha_d}\right) & \text{if flow-connected,} \\
\sigma_d^2 \exp\left(-\frac{3(a + b)}{\alpha_d}\right) & \text{if flow-unconnected,}
\end{cases}$$

Ver Hoef, Peterson, Isaak

NOAA/NMML, CSIRO/CMIS, USDA/FS
Models

Mixed Models

- Mixed Model (Variance Component Model),

\[ Y = X\beta + z_u + z_d + z_e + W_1\gamma_1 + \ldots + W_p\gamma_p + \epsilon, \]

- Tail-down Exponential Model,

\[ \text{var}(Y) = \Sigma = \sigma_u^2 R(\alpha_u) + \sigma_d^2 R(\alpha_d) + \sigma_e^2 R(\alpha_e) + \sigma_1^2 W_1 W_1' + \ldots + \sigma_p^2 W_p W_p' + \sigma_0^2 I. \]
Models

Mixed Models

\[
\begin{pmatrix}
    Z_{\text{observed}} \\
    Z_{\text{unobserved}}
\end{pmatrix} = X\beta + \epsilon, \quad \text{var}(\epsilon) = \Sigma(\theta)
\]

Variance Components:

SO$_4$ Data

Estimation

Ver Hoef, Peterson, Isaak
NOAA/NMML, CSIRO/CMIS, USDA/FS
Models

Estimation

**SO4, Tail-up Model Only**

<table>
<thead>
<tr>
<th>effect</th>
<th>estimate</th>
<th>std.err</th>
<th>df</th>
<th>t.value</th>
<th>prob.t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9.07</td>
<td>0.984</td>
<td>21</td>
<td>9.22</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Distance Upstream</td>
<td>-0.1337</td>
<td>0.0537</td>
<td>21</td>
<td>-2.488</td>
<td>0.0213</td>
</tr>
</tbody>
</table>

**SO4, Components Model**

<table>
<thead>
<tr>
<th>effect</th>
<th>estimate</th>
<th>std.err</th>
<th>df</th>
<th>t.value</th>
<th>prob.t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9.20</td>
<td>1.399</td>
<td>21</td>
<td>6.58</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Distance Upstream</td>
<td>-0.1259</td>
<td>0.0664</td>
<td>21</td>
<td>-1.893</td>
<td>0.0722</td>
</tr>
</tbody>
</table>

\[
\begin{pmatrix}
Z_{\text{observed}} \\
Z_{\text{unobserved}}
\end{pmatrix} = X\beta + \varepsilon, \quad \text{var}(\varepsilon) = \Sigma(\theta)
\]
Prediction

\[
\begin{pmatrix}
Z_{\text{observed}} \\
Z_{\text{unobserved}}
\end{pmatrix} = X \beta + \epsilon, \quad \text{var}(\epsilon) = \Sigma(\theta)
\]

Block Kriging

\[
E = 5.82 \quad \text{se} = 1.13
\]

\[
F = 9.67 \quad \text{se} = 0.67
\]
Software / Tools - Collaborators

Jay Ver Hoef, Statistician
NOAA National Marine Mammal Laboratory

David Clifford, Statistician
CSIRO Mathematics, Informatics & Statistics

Rohan Shah, Mathematician & Programmer
CSIRO Mathematics, Informatics & Statistics
Software / Tools

• Multidisciplinary skills are required
  • Knowledge of aquatic ecology
  • Specialized skills using geographic information systems (GIS)
  • Spatial statistics

• Knowledge transfer & methodological uptake requires software/tools
  • Journal articles are not enough
  \[
  \begin{pmatrix}
    z_{\text{observed}} \\
    z_{\text{unobserved}}
  \end{pmatrix} = X\beta + \varepsilon, \quad \text{var}(\varepsilon) = \Sigma(\theta)
  \]
Suite of GIS and Statistical Tools

ArcGIS

- FLoWS Custom Toolset
- STARS Custom Toolset

R

- SSN Package
- .SSN
Software / Tools – FLoWS

Functional Linkage of Watersheds and Streams (FLoWS)

- ArcGIS Geoprocessing Toolbox written in Python for ArcGIS v9.3
- Developed by Dave Theobald and John Norman at Colorado State University
- Website: http://www.nrel.colostate.edu/projects/starmap/flows_index.htm

What are the FLoWS v9.3 tools?

- Graph theoretic-based analysis tools
- Functionally link aquatic and terrestrial components of a landscape based on hydrologic processes
- Hydrologic modelling framework
Spatial Tools for the Analysis of Streams (STARS)

- ArcGIS Geoprocessing Toolbox written in Python for ArcGIS v9.3.1

Toolsets

- **Pre-processing:** Identify unique topological relationships that are prohibited in the SpatialStreamNetwork class

- **Calculate:** Derive spatial data needed to fit a spatial statistical model to streams data:
  - Hydrologic distances
  - Spatial weights
  - Covariates for both observed & unobserved locations

- **Export:** Export the spatial features, topological relationships, and attribute information to a format that can be easily accessed using R
Export – Create .ssn Directory
Stores the spatial features, topological relationships, and attribute information to a format that can be easily accessed using R

- .SSN directory components:
  - Sites & edges shapefiles
  - Text file containing topological information for each network
  - Prediction site shapefile(s): optional

- Easily transferable
  - If it’s in the .ssn directory, you need it

Note, shapefiles will always have at least three files associated with them (.shp, .shx, .dbf) and may have additional files (.sbn, .sbx, .prj, etc.)
Software / Tools

ArcGIS

FLoWS
Custom Toolset

STARS
Custom Toolset

R

SSN
Package

.ssnn
Software / Tools – SSN Package

Spatial Stream Networks (SSN) Package

Purpose
• Read, store, visualize, analyse & export spatial streams data

Our design goals
• Make the package user friendly
  • Easy to install on multiple operating systems
  • Fit models to multiple response types
  • Include helper functions so that you don’t have to be an R developer or a statistician to fit a model
  • Reduce opportunities for user error
SpatialStreamNetwork Class: S4 object

- Formally defined class structure
  - Ensures data is in the “right” format

- Conventions for spatial classes set out by Bivand et al. 2008
  - Co-authored sp package to extend R classes and methods for spatial data
  - 49 other packages depend on sp

- SpatialStreamNetwork class
  - Contains both point and line features: edges, sites, prediction sites
  - Unique coordinate system used to navigate network
Software / Tools – SSN Package

**.ssn directory**
- Explicitly linked to SpatialStreamNetwork object
- R reads/writes data to this location

**binaryID.db: SQLite database**
- Stores some topological information for line segments
- Automatically created when data are imported into R

**Distance matrices**
- Stored as .Rdata files
- External storage means they only have to be calculated once
SSN Package Functionality…

• Fit generalised linear models with spatially-autocorrelated errors
  • Normal likelihood methods (including REML)
  • Quasi-likelihood for Poisson and Binomial families

• Supported distributions: Gaussian, binomial, Poisson
  • Repeated measurements are permitted
  • Random effects may be included

• Block-kriging estimates

• Multiple covariance models supported:
  • Tail-up & tail-down: Exponential, Linear-with-Sill, Spherical, Mariah
  • Euclidean: Exponential, Spherical, Linear-with-Sill, Cauchy
Software / Tools – SSN Package

SSN Package Functionality…

• Influence & cross-validation diagnostics
  • Raw residuals, studentized residuals, cross-validation residuals, Leverage values, Cooks D, cross-validation predictions, and standard errors for the cross validation predictions

• Predictions & std errors at unobserved locations
  • Predicted response variable & standard errors for the predictions

• Create SSN objects
  • Generate network(s) with observations & prediction locations
  • Point distribution based on different sample designs: Poisson, hard core, systematic

• Simulate data
  • Gaussian, binomial, Poisson distribution
  • Different covariance structures & fixed effects
Software / Tools – SSN Package

Helper Functions for 4 Classes

Allow users to examine data, produce model diagnostics, visualize data, and export results easily

Model Diagnostics

• AIC, generalized R² for covariates, cross-validation statistics, variance components

Visualisation

• Plot: map of observed values, residuals, or predictions and/or standard errors
• Torgegram: flow-connected & flow-unconnected
• QQplots, histograms, boxplot

Others:

Access, alter, export, subset, summary…
Model Fitting

```r
> mf04.glmssn0c <- glmssn(Summer_mn ~ ELEV_DEM + SLOPE, mf04c,
+ CorModels = c("Exponential.tailup", "Exponential.taildown",
+ "Exponential.Euclid"),
+ addfunccol = "afvArea")
>
> summary(mf04.glmssn0c)

$fixed.effects.estimates

<table>
<thead>
<tr>
<th>FactorLevel</th>
<th>Estimate</th>
<th>std.err</th>
<th>t.value</th>
<th>prob.t</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Intercept)</td>
<td>29.348881916</td>
<td>5.540699171</td>
<td>5.296964</td>
<td>0.00000</td>
</tr>
<tr>
<td>2 ELEV_DEM</td>
<td>-0.009970748</td>
<td>0.002758303</td>
<td>-3.614812</td>
<td>0.00051</td>
</tr>
<tr>
<td>3 SLOPE</td>
<td>-12.812193742</td>
<td>12.340454558</td>
<td>-1.038227</td>
<td>0.30207</td>
</tr>
</tbody>
</table>

$covariance.parameter.estimates

<table>
<thead>
<tr>
<th>Covariance.Model Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Exponential.tailup parsill</td>
<td>1.782467e+00</td>
</tr>
<tr>
<td>2 Exponential.tailup range</td>
<td>7.845687e+05</td>
</tr>
<tr>
<td>3 Exponential.taildown parsill</td>
<td>2.370570e+00</td>
</tr>
<tr>
<td>4 Exponential.taildown range</td>
<td>1.951213e+05</td>
</tr>
<tr>
<td>5 Exponential.Euclid parsill</td>
<td>3.625637e-01</td>
</tr>
<tr>
<td>6 Exponential.Euclid range</td>
<td>1.631959e+05</td>
</tr>
<tr>
<td>7 Nugget parsill</td>
<td>1.536211e-02</td>
</tr>
</tbody>
</table>

> varcomp(mf04.glmssn0c)

<table>
<thead>
<tr>
<th>VarComp</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariates (R-sq)</td>
<td>0.139588592</td>
</tr>
<tr>
<td>Exponential.tailup</td>
<td>0.336483306</td>
</tr>
<tr>
<td>Exponential.taildown</td>
<td>0.450161531</td>
</tr>
<tr>
<td>Exponential.Euclid</td>
<td>0.066849368</td>
</tr>
<tr>
<td>Nugget</td>
<td>0.002917203</td>
</tr>
</tbody>
</table>
```
Torgegrams

Temperature

Abundance Residuals

Semivariance

Stream Distance (m)

Flow-connected
Flow-unconnected

Semivariance

Stream Distance (m)

Flow-connected
Flow-unconnected

(a)

(b)
Future Tool Development

Statistical tools
• Continue to improve the SSN package…

Update GIS tools
• Personal geodatabase is not supported in ArcGIS v10
• FLoWS toolset is only available up to ArcGIS v9.3

Tools must be more computationally efficient
• Sensor networks = massive datasets
• Model-fitting algorithms can’t handle observations

CSIRO Accelerated Computing Projects
• Speed up matrix operations
  • Intel’s MKL library (commercial software)
  • Multiple GPU-based calculations
  • Implement MAGMA multi-GPU qr() decomposition function in R
Implicit goal
Make the tools and dataset available so that others can extend/apply these methods

http://www.nceas.ucsb.edu/featured/peterson
US Forest Service website (currently under construction):


Tools:
• FLoWS, STARS, & SSN package

Documentation:
• Tutorials, journal articles, examples

Datasets:
• Full Lower Snake dataset
• Smaller subsets of the Lower Snake
• Space-time, block kriging, macroinvertebrates
Regional Infrastructure for VHP Analysis of Stream Data

Stream datasets can be more accurately analyzed & scaled consistently with GIS infrastructure complete. 

\[
\begin{array}{c}
\text{Symmetric Distance Classes} \\
\begin{array}{c}
\text{(a)} \quad \text{(b)} \\
\end{array} \\
\text{Asymmetric Distance Classes} \\
\begin{array}{c}
\text{(a)} \quad \text{(b)} \\
\end{array}
\end{array}
\]

\[\sim 350,000 \text{ stream kilometers}\]
A Regional Stream Temperature Model for Mapping Thermal Habitats & Species Climate Vulnerability Across the GNLCC

Dan Isaak¹, Erin Peterson², Seth Wenger³, Jay Verhoef⁴, Charlie Luce¹, Jason Dunham⁵, Dave Nagel¹, Jeff Kershner⁵, Brett Roper¹, Steve Hostetler⁵, Dona Horan¹, Gwynne Chandler¹, Sherry Wollrab¹, Sharon Parkes¹, Dave Hockman⁵
GNLCC Temperature Project

Landscape Conservation Cooperatives
GNLCC plus Temperature Project

Great Northern Landscape Conservation Cooperatives

Isaak et al. 2011. GNLCC proposal

& south Idaho & the coast

Landscape Conservation Cooperatives

Range-wide climate vulnerability assessment for bull trout in the conterminous United States

"Judging by one criterion, it is Extinct!"

"But judging by another, it is alive and healthy in places!"
Interagency Database

15,000+ unique stream sites
45,000+ summers measured
T3 Temperature Team

T3 Temperature Team
Interagency Database

Database Status (4/2/12)
15,000+ unique stream sites
45,000+ summers measured
Regional Temperature Model

Cross-jurisdictional “maps” of stream temperatures

Consistent datum for strategic assessments

VHP models

NorWeST

Stream Temp

Replace with validation
Data from 2007
Computation Challenges: Model Fitting

NCEAS - Lower Snake Hydrologic Region

- 42,000 stream km
- 5,498 summers
- 1,667 temperature sites
Non-spatial Stream Temp =
- 0.0041*Ele (m)
- 13.9*Slope (%)
+ 0.016*Wat_size (100km²)
-0.0022*Ave_Precip
- 0.041*Flow (m³/s)
+ 0.42*AirMean (°C)

Spatial Stream Temp =
- 0.0045*Ele (m)
- 9.8*Slope (%)
+ 0.012*Wat_size (100km²)
- 0.00061*Ave_Precip
- 0.037*Flow (m³/s)
+ 0.46*AirMean (°C)

Mean Summer Temperature

\[ r^2 = 0.63; \text{RMSE} = 1.88°C \]

\[ r^2 = 0.93; \text{RMSE} = 0.82°C \]
Kriged Temperature “Maps”

2006 - Summer Mean - Boise River

Prediction Precision “Maps”

Payette National Forest
Spatial Uncertainty Map
Data Model for Prediction Points

1 km default, but customizable
NHD+ Hydrolayer Conditioning

1. **Converging streams** - stream segments upstream that are not connected to the main channel were deleted.

2. **Downstream divergences** - side-channels were deleted.

3. **Multi-segment confluences** (i.e., > 2 stream converging) were adjusted.
GIS Layers Online

Shapefiles of prediction points

Reconditioned NHD+ Hydrolayers

Websites for Distribution
Regional Infrastructure for VHP Analysis of Stream Data

Just add your data…
Temperature Data, but also...

- Genetic Attributes
- Water Quality Parameters
- Response Metrics
  - Gaussian
  - Poisson
  - Binomial
- Distribution & abundance