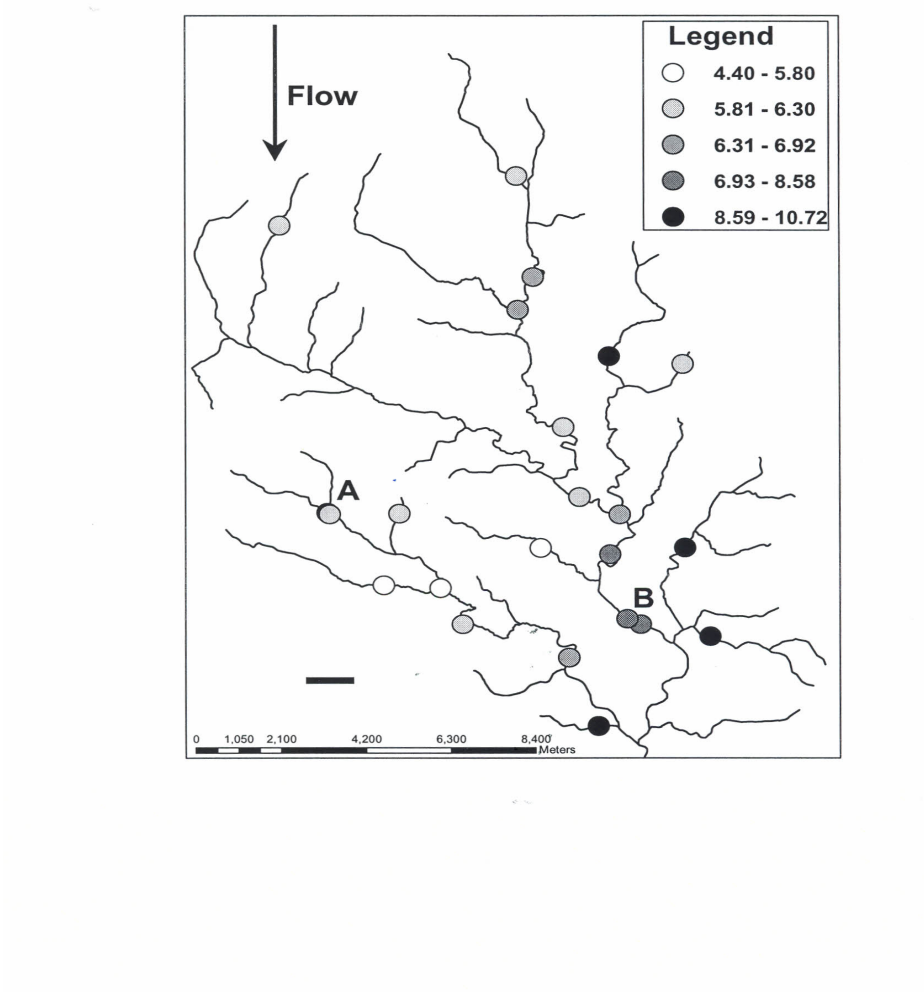
Fluvial Variography: Characterizing Spatial Dependence on Stream Networks

Dale Zimmerman
University of Iowa
(joint work with Jay Ver Hoef, NOAA)

March 5, 2015

Stream network data



Spatial statistics on streams?

- Premise: Values of a variable occurring in or along a stream network are as likely as Euclidean spatial data to obey Tobler's first law of spatial statistics (i.e., values from "nearby" sites tend to be similar).

- Consequently, stream ecologists want to apply methods of spatial statistics to address questions about the variable (e.g. to estimate an average or total over a stream or stream segment, make predictions at unsampled locations, estimate relationships between the primary variable and other variables, . . .).

- But a stream network is not a Euclidean space; does this matter? How much can we borrow or easily adapt from Euclidean geostatistics?

Euclidean geostatistics and classical variography

- Geostatistical approach: regard the observed data as a sample taken from one realization of $Y(\cdot) \equiv \{Y(\mathbf{s}) : \mathbf{s} \in D\}$, where D is a region in two-dimensional (usually) Euclidean space

- Common assumptions: $Y(\cdot)$ is intrinsically stationary and isotropic, i.e., $E[Y(\cdot)]$ is constant across space and its semivariogram $\gamma(\mathbf{s},\mathbf{t}) \equiv \frac{1}{2}E[Y(\mathbf{s}) - Y(\mathbf{t})]^2$ may be expressed as a function of $h = \|\mathbf{s} - \mathbf{t}\|$.

- Under second-order stationarity, $\gamma(h) = C(0) - C(h)$, where $C(\cdot)$ is the covariance function.

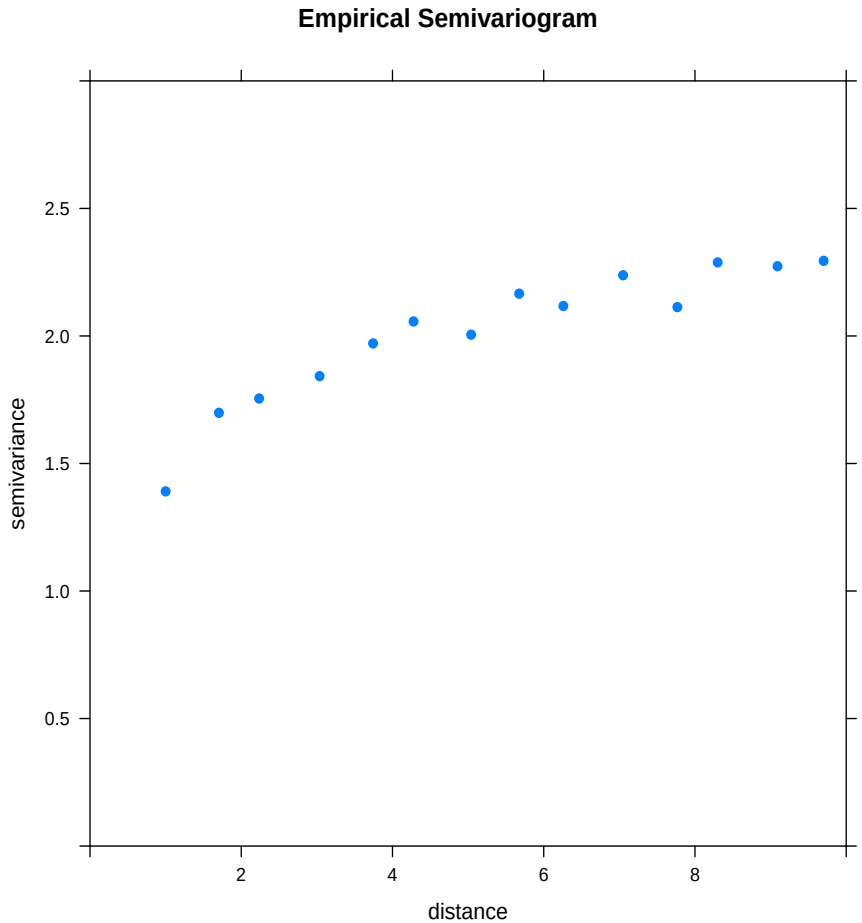
- Modeling/estimation/characterization of the semivariogram is known as *variography*, for which the main tool is the empirical semivariogram.

Empirical semivariogram

$$\hat{\gamma}(h_k) = \frac{1}{2N(H_k)} \sum_{\|\mathbf{s}_i - \mathbf{s}_j\| \in H_k} \Big(Y(\mathbf{s}_i) - Y(\mathbf{s}_j) \Big)^2, \qquad k = 1, \ldots, K,$$

where $Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)$ are the observed data, h_k is a representative distance (often the average or midrange) for a distance bin H_k, and $N(H_k)$ is the number of distinct site-pairs in H_k.

A “typical” empirical semivariogram



Diagnostic value of empirical semivariogram

- The estimator is (approximately) unbiased
- Can often discern a *range*, a *sill*, a *nugget*, and a shape that informs semivariogram model selection
- A chosen model may be fit by any of several methods (WLS, REML)
- Absence of a sill (i.e. unbounded increase) may be evidence of trend contamination
- Can compute and plot in several directions/subregions and compare to assess isotropy/stationarity

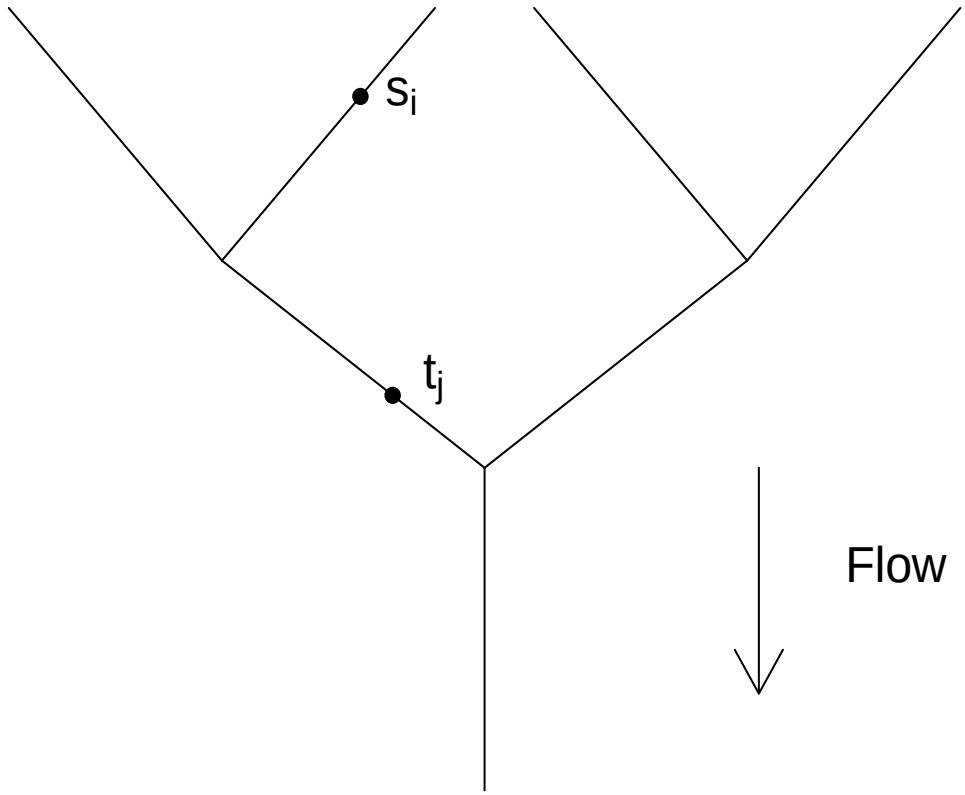
Fluvial variography

- Applications of geostatistics to stream network variables date back to 2003
- Many stream ecologists simply substituted Euclidean distance in the semivariogram model with *stream distance*, i.e. distance along the stream network
- However, valid semivariogram models in Euclidean space are not necessarily valid on a stream network
- Ver Hoef et al. (2006, *EES*) and Cressie et al. (2006, *JABES*) introduced the first valid families of models on stream networks; another important family of valid models was added by Ver Hoef and Peterson (2010, *JASA*).

Fluvial variography, continued

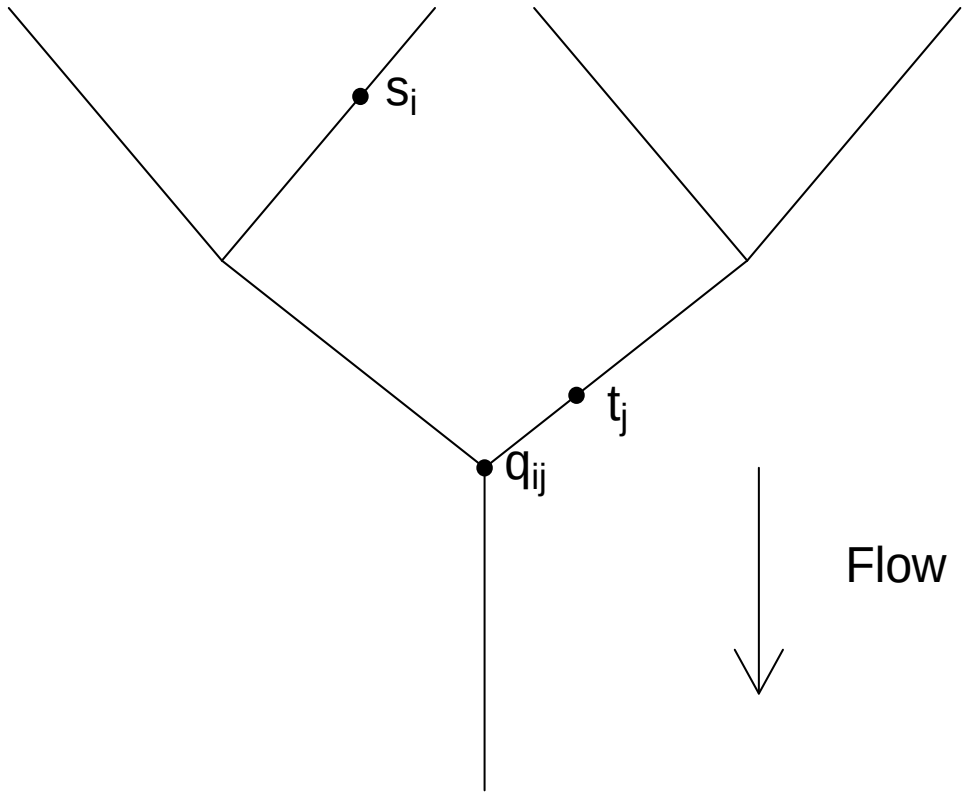
- These models can now be fit to data using the Spatial Modeling on Stream Networks (SSN) package in R
- However, little attention has been given to the development of graphical tools for stream-network variography analogous to the Euclidean distance-based empirical semivariogram
- I will introduce such a tool, called the *Torgegram*, which is an assemblage of four empirical semivariograms

Concepts and notation: flow-connected sites



Stream distance = s − t

Concepts and notation: flow-unconnected sites



Stream distance = $(s - q_{ij}) + (t - q_{ij})$

q_{ij} = upstream distance of 'common junction' of segments i and j

Considerations for covariance models on streams

- As in Euclidean settings, limited data may lead us to make simplifying assumptions akin to stationarity and isotropy
- For some variables, it would seem that values at flow-unconnected sites should be less correlated than values at flow-connected sites, and perhaps they should even be modeled as completely uncorrelated. Examples: water temperature, levels of point-source pollutants
- For those same variables, it would seem that differential flow volumes in tributaries may have an effect on dependence.
- For other variables, e.g. cutthroat trout abundance, perhaps we can ignore flow-connectedness and volume.

Valid covariance models

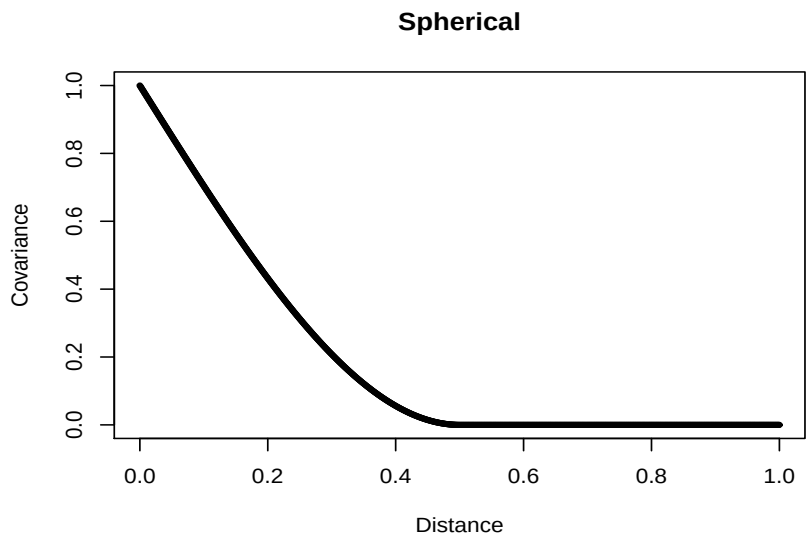
- A classical approach to the development of covariance functions on the real line is to create model residuals as integrations of a moving-average function over a white-noise random process. i.e.,

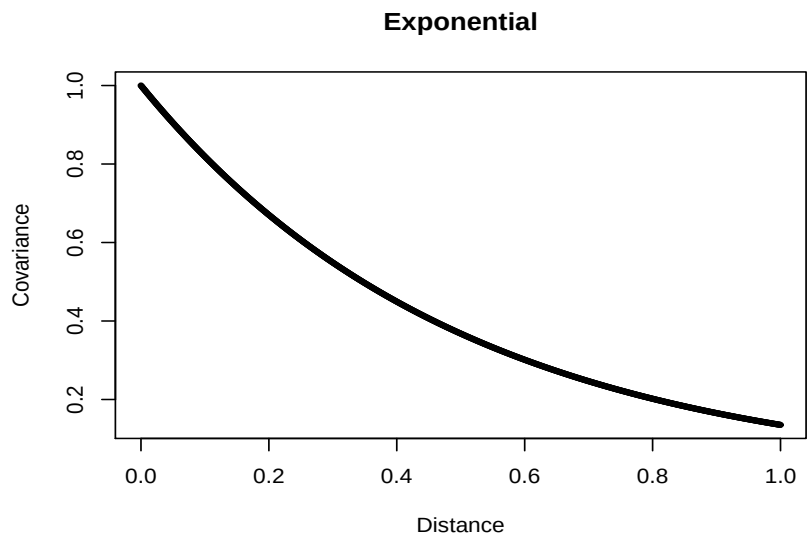
$$\varepsilon(s|\theta) = \int_{-\infty}^{\infty} g(x-s|\theta)dW(x),$$

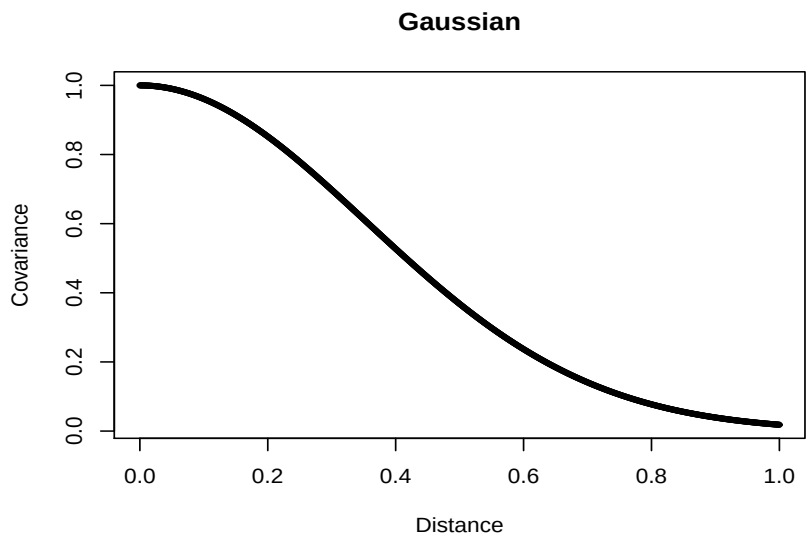
  where x and s are locations, and $g(x|\theta)$ is a square-integrable moving average function, defined on the real line.

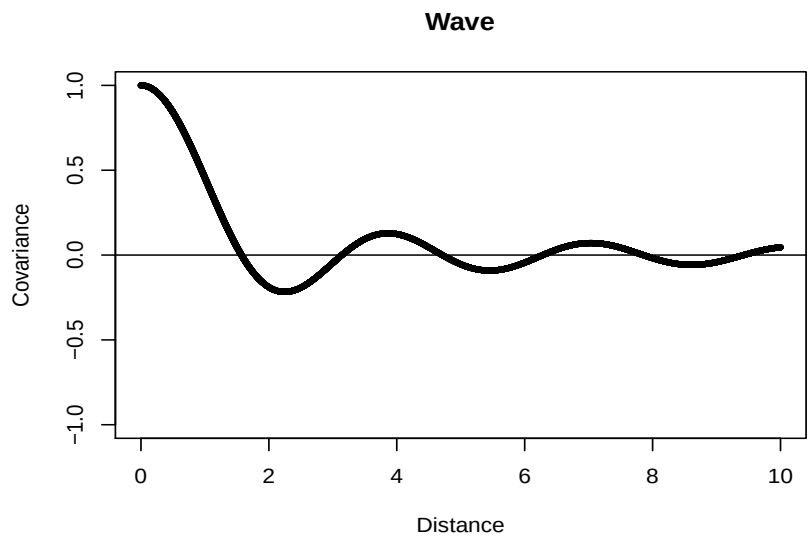
- The covariance between between $\varepsilon(s)$ and $\varepsilon(s+h)$ so defined is given by

$$C(h|\theta) = \int_{-\infty}^{\infty} g(x|\theta)g(x-h|\theta)dx.$$


Spherical
Covariance
0.0
0.2
0.4
0.6
0.8
1.0
Distance

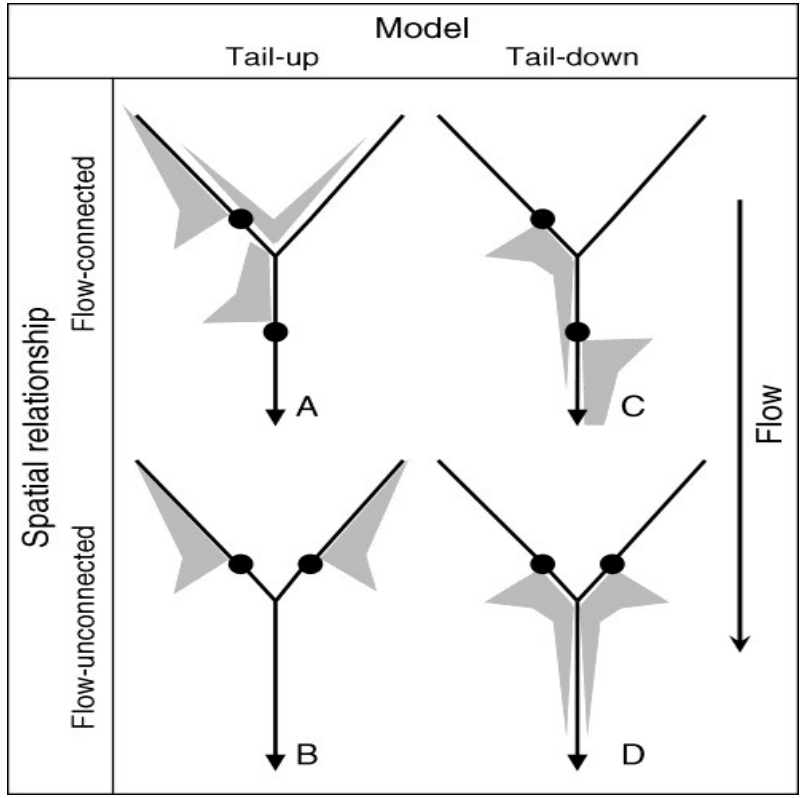

Exponential
Covariance
0.2
0.4
0.6
0.8
1.0
Distance


Gaussian
Covariance
0.0
0.2
0.4
0.6
0.8
1.0
Distance


Wave
Covariance
−1.0
−0.5
0.0
0.5
1.0
0
2
4
6
8
10
Distance

Adaptation to stream networks

- This approach to obtain covariance models can be adapted for use with stream networks
- The models proposed in the aforementioned publications are *unilateral*, i.e. they take $g(\cdot)$ to be positive in only one direction (either upstream or downstream)
- Positive only upstream $\Rightarrow$ "tail-up" models
- Positive only downstream $\Rightarrow$ "tail-down" models

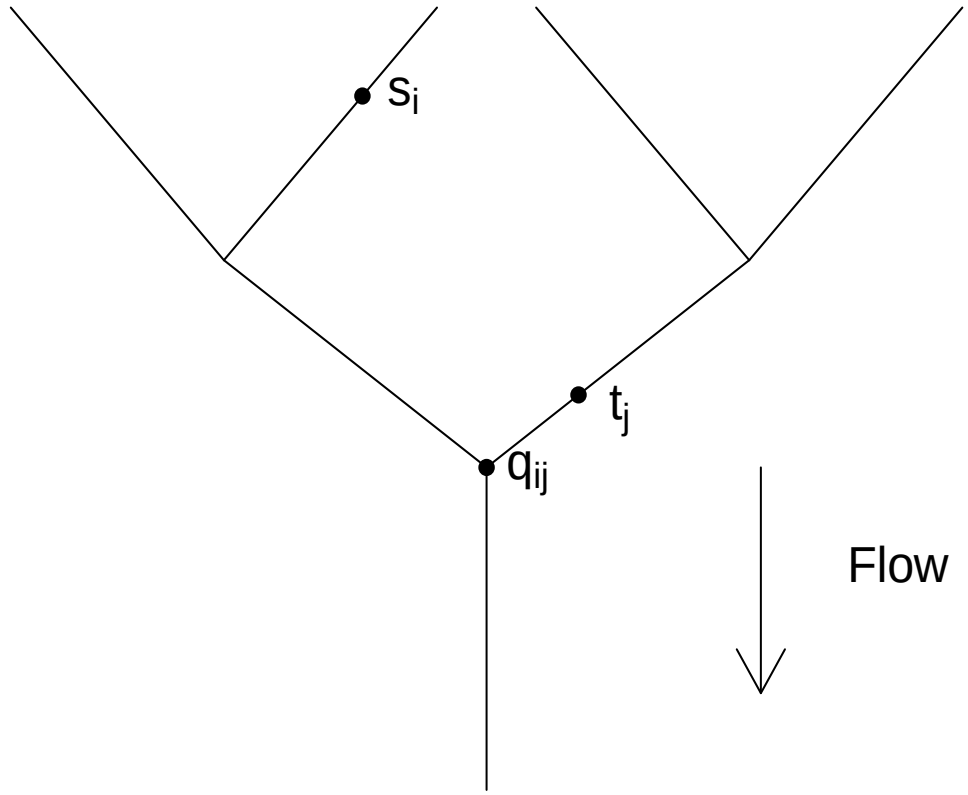

Model
Tail-up
Tail-down
Spatial relationship
Flow-connected
Flow-unconnected
A
B
C
D
Flow

Tail-up models

$$C_{tu}(s_i, t_j | \{\pi_{ij}\}, \theta) = \begin{cases} \pi_{ij} C_{uw}(s - t | \theta) & \text{if } s_i \geq t_j \text{ are f/c} \\ 0 & \text{otherwise,} \end{cases}$$

where the π_{ij}'s are flow-volume weights chosen to preserve variance stationarity, and $C_{uw}(\cdot)$ is a valid covariance function in one dimension.

- Tail-up models account for flow-connectedness (so that the correlation is zero when sites are flow-unconnected) and for differential flow volumes on coalescing stream segments
- Might be appropriate for such things as point-source pollutants



Stream distance = $(s - q_{ij}) + (t - q_{ij})$

q_{ij} = upstream distance of 'common junction' of segments i and j

Tail-down models

$$C_{td}(s_i, t_j|\theta) = \begin{cases} C_{fc}(s-t|\theta) & \text{if } s_i \geq t_j \text{ are flow-connected,} \\ C_{fu}(s-q_{ij}, t-q_{ij}|\theta) & \text{otherwise,} \end{cases}$$

where $C_{fc}(\cdot)$ and $C_{fu}(\cdot)$ are valid covariance functions of one and two variables, respectively (and are related to each other through their functional dependence on the same moving average function).

- Allow for positive correlation among both flow-connected and flow-unconnected site-pairs
- Might be appropriate for such things as counts of fish or insects
- Generally, $C_{fu}(\cdot)$ is <u>not</u> a function of $(s-q_{ij})+(t-q_{ij})$. Exception: exponential case

Mixed models

Some stream network variables may have covariance structures that are neither pure tail-up nor pure tail down. For such variables researchers have adopted a mixed linear model approach, which leads to the following covariance structure:

$$var(\mathbf{Y}) = \sigma^2_{tu}\mathbf{R}_{tu}(\rho_{tu}) + \sigma^2_{td}\mathbf{R}_{td}(\rho_{td}) + \sigma^2_{nu}\mathbf{I},$$

where $\mathbf{R}_{tu}(\rho_{tu})$ is a matrix of autocorrelation values from the tail-up component; $\mathbf{R}_{td}(\rho_{td})$ is a matrix of autocorrelation values from the tail-down component; $\mathbf{I}$ is an identity matrix; σ^2_{tu}, σ^2_{td}, and σ^2_{nu} (the nugget effect) are variance components; and ρ_{tu} and ρ_{td} are vectors of correlation parameters.

The Torgegram: a stream network version of the empirical semivariogram

- For unilateral models and mixes thereof, correlations may depend not only on stream distance but also on:
  - flow connectedness
  - flow volume
  - distances to a common junction

  Thus, for diagnostic purposes one empirical semivariogram is not adequate.
- Instead, four are needed. We call this quadripartite collection the *Torgegram*, in honor of stream ecologist Christian Torgerson.

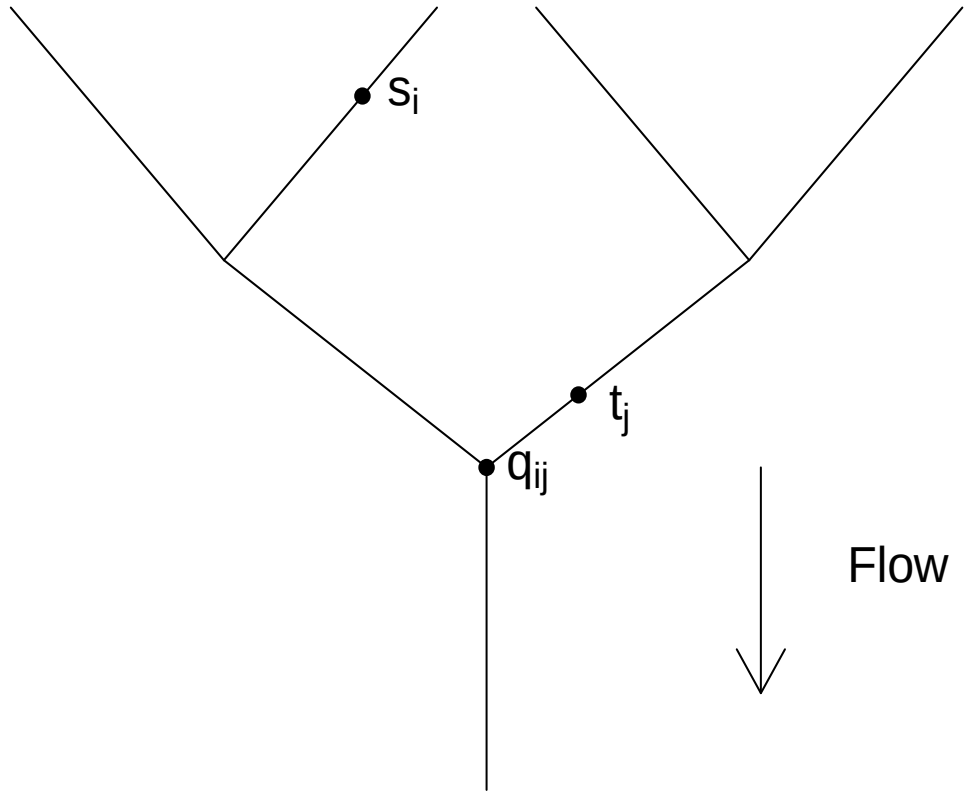
Four components of the Torgegram

1. The flow-unconnected stream-distance (FUSD) semivariogram:

$$\hat{\gamma}_{FUSD}(h_k) = \frac{1}{2N(U_k)} \sum_{(s_i,t_j)\in U_k} \Big(Y(s_i) - Y(t_j) \Big)^2,$$

where the U_k's partition the site-pairs on f/u segments into stream distance bins.

- If $Y(\cdot)$ is pure tail-up, then $\hat{\gamma}_{FUSD}(h_k)$ is unbiased for the f/u portion of its semivariogram, which is flat.
- If $Y(\cdot)$ is not pure tail-up but has exponential semivariogram, then $\hat{\gamma}_{FUSD}(h_k)$ is unbiased for the f/u portion.
- Otherwise, this component has no clean interpretation.



Stream distance = $(s - q_{ij}) + (t - q_{ij})$

q_{ij} = upstream distance of 'common junction' of segments i and j

Four components of the Torgegram

2. The flow-unconnected distance-to-common-junction (FUDJ) semivariogram:

$$\hat{\gamma}_{FUDJ}(j_k, j_l) = \frac{1}{2N(J_k, J_l)} \sum_{s_i \in J_k, t_j \in J_l} \Big(Y(s_i) - Y(t_j) \Big)^2,$$

where the (J_k, J_l)'s partition the site-pairs on f/u segments into bins on the basis of distances to common junction.

- Unbiased for the f/u portion of the semivariogram, without qualifications.

Four components of the Torgegram

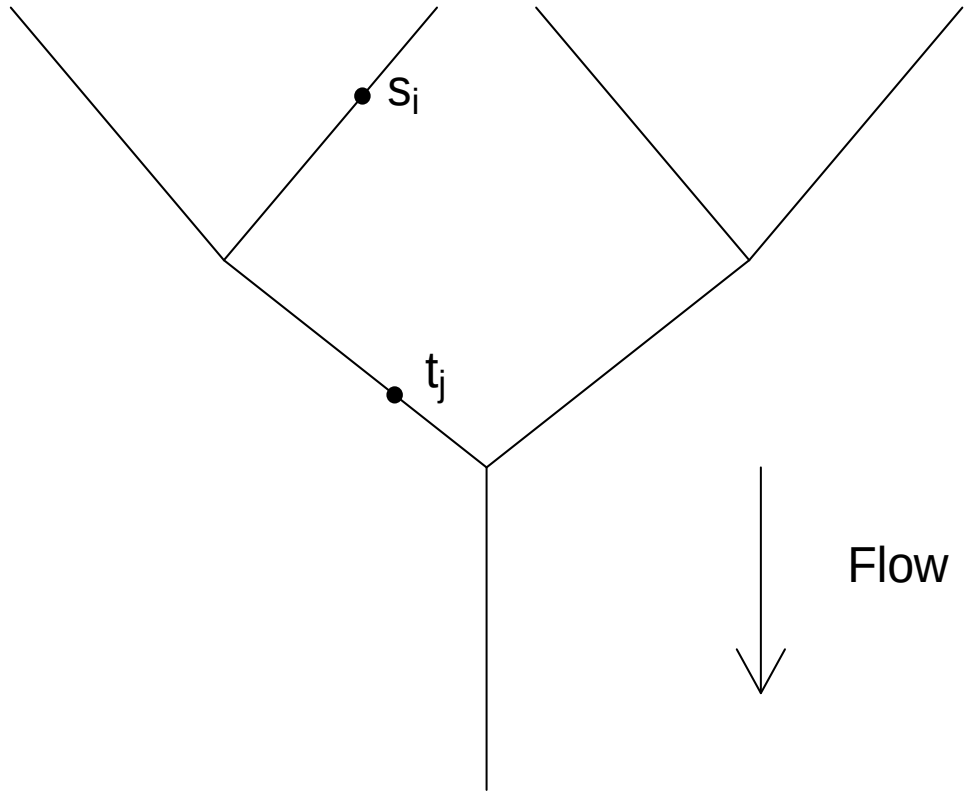
3. The flow-connected stream-distance (FCSD) semivariogram:

$$\hat{\gamma}_{FCSD}(h_k) = \frac{1}{2N(C_k)} \sum_{(s_i,t_j)\in C_k} \Big(Y(s_i) - Y(t_j) \Big)^2,$$

where the C_k's partition the site-pairs on f/c segments into stream distance bins.

- If $Y(\cdot)$ is pure tail-down, then $\hat{\gamma}_{FCSD}(h_k)$ is unbiased for the f/c portion of its semivariogram.
- Otherwise, this component is positively biased for the f/c component (because it does not account for flow volume) and has no clean interpretation.

- However, if the FCSD semivariogram is computed from only those site-pairs for which both sites lie on the same segment, then it is unbiased for the f/c semivariogram without qualification (analogous, in a sense, to a pure error mean square). We call this the pure-correlation FCSD semivariogram.



Stream distance = s − t

Four components of the Torgegram

4. The flow-connected volume-adjusted (FCVA) semivariogram:

$$\hat{\gamma}_{FCVA}(h_k) = \overline{\gamma}_{FUSD} - \frac{1}{2N(C_k)} \sum_{(s_i,t_j)\in C_k} \frac{2\overline{\gamma}_{FUSD} - [Y(s_i) - Y(t_j)]^2}{\pi_{ij}}$$

where $\overline{\gamma}_{FUSD} = \frac{\sum_k N(U_k)\hat{\gamma}_{FUSD}(h_k)}{\sum_k N(U_k)}$.

- $\hat{\gamma}_{FCVA}(h_k)$ is unbiased for the unweighted f/c portion of the semivariogram of a pure tail-up $Y(\cdot)$
- Negatively biased otherwise

A strategy for fluvial variography

In practice, the covariance structure of $Y(\cdot)$ is unknown to the analyst, but the Torgegram may be used to identify a plausible structure, using the following strategy:

1. Examine the FUSD semivariogram. If it appears to be flat, adopt a pure tail-up model and use the FUSD and FCVA semivariograms unambiguously to determine the model's attributes. Otherwise, conclude that a tail-up model is not adequate and proceed to the next step.

2. Compare the "pure correlation" FCSD semivariogram to the "remainder" of the FCSD semivariogram. If they are similar, adopt a pure tail-down model and use the FCSD and FUDJ semivariograms unambiguously to determine the model's at-

tributes. Otherwise, proceed to the next step.

3. Adopt a mixed model and estimate it via formal methods.

Formal test for pure tail-up dependence

- Adaptation of Diblasi-Bowman test for spatial independence in Euclidean geostatistics (2001, *Biometrics*)
- Define $\hat{\gamma}_{ij} = (Y_i - Y_j)^2$ for f/u sites s_i and t_j
- Test statistic

$$T = \frac{\sum_{i<j} \left(\hat{\gamma}_{ij} - \overline{\gamma}_{FUSD}\right)^2 - \sum_{i<j} \left(\hat{\gamma}_{ij} - \tilde{\gamma}_{ij}\right)^2}{\sum_{i<j} \left(\hat{\gamma}_{ij} - \tilde{\gamma}_{ij}\right)^2}$$

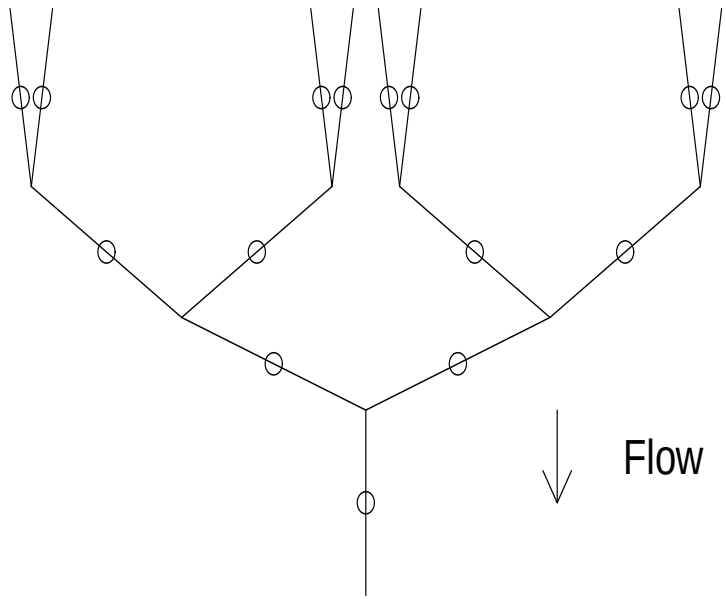
  where $\tilde{\gamma}_{ij}$ is a nonparametric (kernel-smoothed) estimate of the f/u component.

- Can assess significance via random permutation

Simulation study

- Dyadic branching stream network of Shreve order 4, 5, or 6 (15, 31, or 63 segments of length 1.0 unit)
- Null model: Tail-up exponential model $C_{tu}(h) = \exp(-\theta h) \equiv \rho^h$; equal weighting ($\sqrt{0.5}$) at each node
- Alternative model: Tail-down exponential model
- Sample each segment at its midpoint
- 1000 simulations from Gaussian process for each combination of sample size and ρ
- For each simulation, determine whether T is among the top 5 T-values when considered with those from 99 independently randomized permutations of the data ($\Rightarrow$ size-.05 test)

Stream network (order-4 case) for simulation study



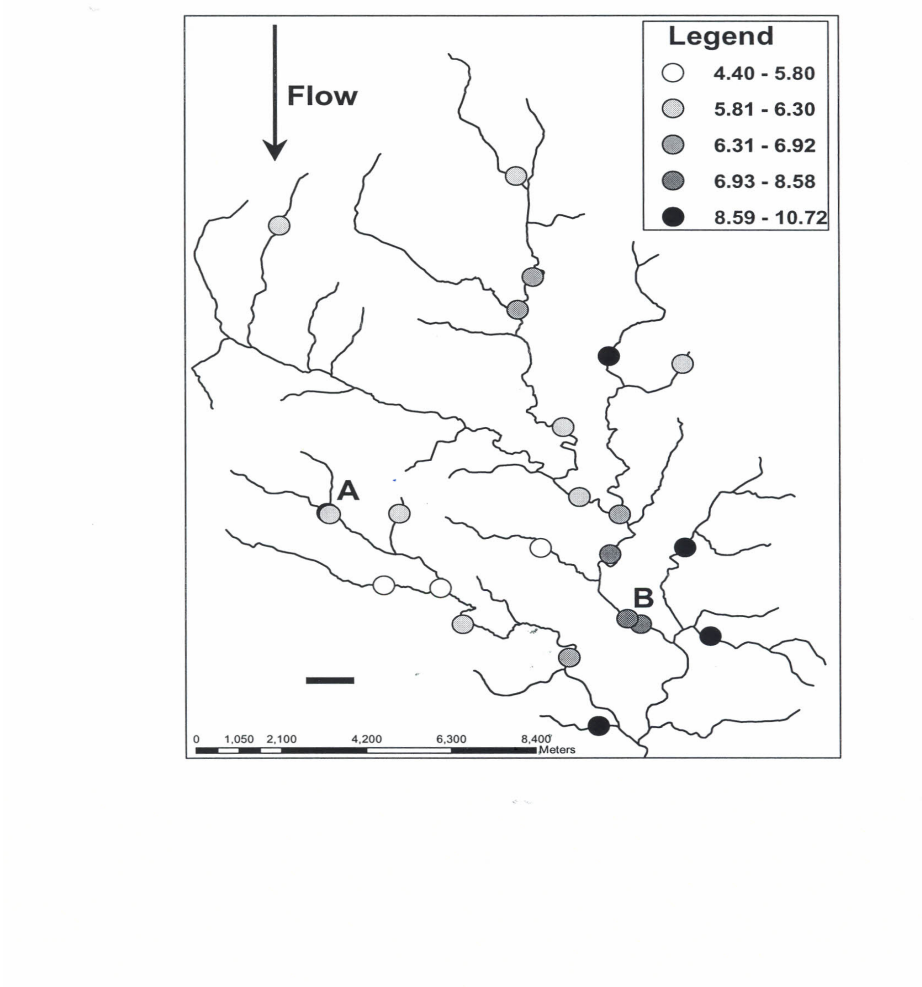
Simulation study results

Power (and size):

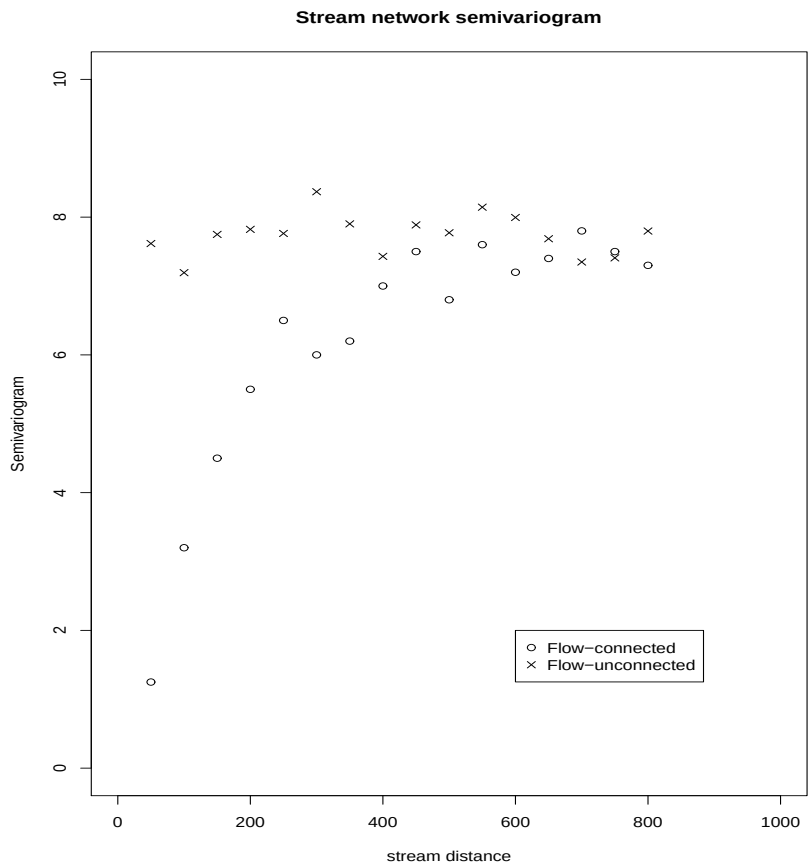
Order	n	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.90$
4	15	.172 (.049)	.266 (.039)	.378 (.010)
5	31	.254 (.055)	.478 (.046)	.629 (.023)
6	63	.473 (.061)	.790 (.059)	.957 (.047)

The test behaves as expected; size and power are reasonable.

Application to Maryland SO_4 data



FUSD and FCVA semivariograms for Maryland SO_4 data



The test for pure tail-up dependence is not rejected ($P = 0.23$).

Conclusions and ongoing work

- The Torgegram is the graphical equivalent of the empirical semivariogram for characterizing spatial dependence on stream networks.
- A coherent strategy for stream network covariance model selection can be based on it, but would be enhanced by several accompanying hypothesis tests.
- We've developed a test for pure tail-up dependence, and are currently working on tests for
  - pure tail-down dependence
  - exponentiality of the tail-down component
  - variance stationarity, both within and across watersheds