**Sequential Design of Optimal Stream Monitoring**

**Networks Using SPARROW**

**Principal Investigator: Susan Colarullo**

**Abstract**

The SPARROW model is unique in its ability to predict regional fluxes of nutrient loads from spatially-distributed sources and transport parameters using a closed-form solution to the watershed transport problem. The analytical form of the SPARROW watershed regression equation lends itself to first-order second-moment uncertainty analysis, which relates uncertainty in independent source terms to uncertainty in predicted stream loads. Based on the analysis, a mixed-integer objective function for sequential stream monitoring network design, stochastically conditioned on loads monitored in the existing stream network, is developed. The objective function uses results from the SPARROW bootstrapping module to systematically identify reaches where expansion of the existing network will most reduce load prediction error.

**The SPARROW Transport Equation**

The SPARROW model uses the following watershed transport equation to predict distributed nutrient loads resulting from both land surface and instream nutrient losses (Smith *et al*., 1997):

  i ∈I [1]

where:

Li = load in reach i

sn,j = nutrient mass from the nth source in drainage to reach j

Zj = vector of land-surface characteristics associated with drainage to reach j

Ti,j = vector of channel transport characteristics between reaches i and j

βn = source regression parameter

α = vector of land-to-water delivery regression parameters

δ = vector of instream loss regression parameters

J = the set of reaches upgradient of and including reach i, excluding reaches containing or upgradient of upstream monitored sources

The analytical form of equation [1] lends itself to statistical analysis using first-order, second-moment techniques, which provide a formal mechanism for relating uncertainty in independent variables to uncertainty in the dependent load variable.

We’re interested in reducing error in our estimate of Li by reducing uncertainty in any of the variables it depends on. Of particular interest are errors stemming from monitored point source terms, which can be reduced by simply expanding the network of sites at which instream load samples are collected.

**First-Order Approximation of Load**

Classical multivariate statistical methods of minimizing load or concentration prediction error do not explicitly account for the physical characteristics of the transport system under investigation. These methods also rely on marginal and joint distributions of input and output random variables to design an optimal strategy of measuring source terms, without regard for how spatial variations in those terms might influence nutrient transport. To account for the dynamics of distributed transport, we analyze the SPARROW transport equation using a first-order second-moment approach. Such an approach provides a formal mechanism for relating uncertain monitored sources to uncertain load predictions.

To find the monitored source location that most reduces load prediction uncertainty, we assume a linear relationship between instream source and load using the first-order Taylor series expansion about mean load. Assuming that variation in load is predominantly caused by variation in the vector of monitored sources, load in reach i can be expressed as:

  [2]

where is the sensitivity of load in reach i with respect to the instream monitored source vector **s**, conditioned on some estimate of the monitored source ensemble mean vector, **sd** is a vector of distributed sources, and **Z** and **T** are previously defined. Note that, while load depends on distributed source variables **sd** as well as on transport variables **Z** and **T,** this particular analysis focuses on how uncertainty in instream sources affects load prediction error. Thus, **sd**, **Z**, and **T** are not considered to behave as random variables, and do not fluctuate about their means. For notational convenience, the dependence of load on distributed source and transport variables will be implied in equations to follow.

The above equation essentially expands solution of the transport equation about the deterministic solution, adding deviations that are attributable to random, or unexplained, fluctuations about the mean source. Because it omits instream source terms of second- and higher- order, it represents an approximation of the relation between load and instream monitored sources. Furthermore, if variations in load are large relative to mean load or if load and monitored sources are not approximately linearly related in the vicinity of mean load, the approximation may become invalid.

The sensitivities ∂Li/∂sj, i = 1, 2, 3,…., I, j = 1, 2, 3,…., J define an I-by-J Jacobian matrix of load sensitivities:



  [3]

 



where J is the number of potential monitored sources that are presently unmonitored in the existing network. In practice, the number of reaches contained in set J depends on the downgradient reach index i. In this context, J can be viewed as the set of all unmonitored reaches upgradient of the entire set of monitored reaches, excluding reaches upgradient of upgradient monitored reaches, with zero-valued entries for reaches not contained in the ith local set.

From the SPARROW transport equation [1], sensitivity of load Li to monitored source sj is given by:

  [4]

where  for instream monitored sources and Gij is the ij-th entry in a matrix of instream decay loss fractions evaluated at the current monitored source estimate :

 

**G** =  [5]



 

It should be noted that, while there is some dependency between s and the regression parameter δ, this dependence stems from the SPARROW regression, is not explicit, and does not enter into the differentiation. The load sensitivities embody the unique physics of the SPARROW transport problem, and are therefore a crucial element of the monitoring network design strategy. Equation [5] suggests that sensitivity of predicted load to monitored sources will be greatest in reaches where the fraction of nutrient remaining in the stream after instream losses is high.

Equations [2] and [4] provide us with a framework for systematically relating downstream load to upstream monitored sources. However, we are far more interested in how the statistical moments of the monitored source random variable translate into moments of the predicted load variable. Specifically, we’d like to know how uncertainty, or variance, in the monitored source variable contributes to uncertainty in predicted load. Based on the first-order approximation given by equation [2], we perform a second-moment analysis to obtain explicit expressions for moments of loads as functions of moments of monitored sources. Taking the expectation of equation [2],

  [6]

Equation [6] states that to first order, the expectation of the load is simply the load conditioned on the ensemble mean of the monitored source vector.

**Second-Moment Analysis**

We are interested in some measure of the uncertainty in load caused by uncertainty in instream monitored sources. To obtain an estimate of load uncertainty as a function of monitored source uncertainty, we evaluate the load covariances from equation [2] as follows (Dettinger and Wilson, 1981):

 

 

 

  [7]

where **Gi** is the ith row of Jacobian of conditional load sensitivities evaluated at the source ensemble mean, and Cov(**s**) is the conditional source variance vector. The load covariance given by equation [7] is what we’re truly interested in, because it quantifies the uncertainty in predicted load caused by uncertainty in the monitored source term of the SPARROW transport equation. It states that uncertainty in load will depend not just on uncertainty in **s**, but also on how changes in monitored sources translate into changes in load, as given by sensitivity coefficients ∂Li/∂sj. In monitored source reaches where predicted load is sensitive to upstream monitored source magnitude and where the source covariance is also large, the source contributes the greatest variance, or uncertainty, to the load prediction.

Our goal is to identify the single unmonitored reach where **s** contributes the most variance to predicted load in reach i, Li. Information gained by measuring that particular element of **s** will best inform the conditional load estimate.

From equation [7], taking the derivative of  with respect to Cov(**s**) yields:

  [8]

For a single instream monitored source location j,  and the above equation reduces to:

  [9]

We’re interested in formulating an objective function that maximizes the change in conditional load variance,  , over all previously unmonitored reaches.

**The Design Objective Function**

A network design algorithm can easily be developed using standard integer-programming techniques that ‘turn on’ sampling at any one of J potential measurement locations to yield the most additional information about load, providing us with a more reliable predictive model.

Termwise expansion of equation [2] for load, and weighting of each term by a binary decision variable uj, yields:

  [10]

where the uj is either 1 to denote upgradient instream monitoring or 0 otherwise. We want to maximize total conditional load variance across the study area by searching over the set of unmonitored reach indices J. Assuming that measured loads are independent among all reaches, the objective function can be expressed as:



 

 [11]

Subject to:

uj = 0 or 1 j ∈ J  [11a]

where sensitivities are evaluated at some prior estimate of the unmonitored source ensemble mean. Since we are conditioning on an unknown mean, the presence of an outlier entry in **s** can produce distorted sensitivity estimates and yield an extremely suboptimal network design, particularly during early stages of monitoring when the existing network may be small and little information about source magnitude is available. To guard against that possibility, sequential design is introduced via the additional constraint:

  [11b]

The sequential design constraint forces the model to choose the single reach where uncertainty in the monitored source contributes the most uncertainty to predicted load L. It protects the design against prior uncertainty in the source vector by restricting the network design to choose only a single reach at a time. As a natural consequence of the sequential design constraint, the off-diagonal terms of the Cov(**s**) matrix, which account for information shared by two source measurements, become equal to zero. The objective function is further simplified by noting that u2j = uj for all j∈J. Together with the simplification provided by dropping off-diagonal terms, the objective function reduces to:

  [12]

where the load sensitivities are conditioned on the current estimate of the source ensemble mean using equation [4]. Both ensemble mean and variance are easily estimated from standard nonparametric moment-estimating techniques such as bootstrapping.

Objective function [12], subject to constraints [11a] and [11b], is a mixed integer programming problem that maximizes the gain in conditional load information from making a single instream source measurement. It establishes a quantitative framework for sequential network design that optimizes expansion of the existing stream monitoring network. As might be anticipated solely on the basis of intuition, information gain will be greatest by monitoring in a reach upgradient from the reach where load is most sensitive to instream sources and where source uncertainty is the largest, because it is at these locations where conditioning on a new source measurement will most improve our understanding of load. After the source is monitored in the optimal reach, it becomes known during subsequent iterations of the design algorithm () and drops from the objective function because it is no longer a potential source of uncertainty. During subsequent iterations of the design algorithm, estimates of load sensitivity and source variance are conditioned on the newly-acquired monitored source magnitude, as well as on source measurements from the original network.

**Sequential Design Procedure**

The design algorithm requires estimates of  and  for each unmonitored reach in the set J to condition the objective function [12]. Ensemble source mean and variance for each unmonitored reach can be obtained by conducting a large number of probabilistic experiments using the same bootstrapping techniques as outlined by Smith *et al.* (1997). The proposed sequential design proceeds iteratively as follows:

1. Using the existing set of monitored sources, generate random samples of all SPARROW regression parameters using bootstrapping techniques.
2. For each set of regression parameters, use bootstrapping to generate multiple realizations of unmonitored sources. The number of realizations in the ensemble depends on the degree of uncertainty in the current set of monitored sources. Since a higher degree of uncertainty increases the likelihood of an outlier source realization and a suboptimal design, more realizations are needed when there is a high degree of source uncertainty present in the existing network.
3. Using the multiple unmonitored source realizations generated in the step above, build local conditional cumulative density functions (ccdfs) for the instream source in each unmonitored reach. From the local ccdfs, estimate source ensemble mean and variance in each unmonitored reach and evaluate load sensitivity in all downgradient reaches, conditioned on the ensemble mean.
4. Calculate the terms of objective function Z for all i and j, and rank them from highest to lowest. Choose the reach index j = j\* associated with the highest value as the ‘optimal’ reach for the next instream source measurement.
5. Monitor the source in reach j\*.
6. Repeat steps 1 through 5, each time conditioning the new ensemble of unmonitored source realizations on all measured instream source magnitudes, including the most recently acquired source measurement, to complete the iterative sequential design.

**Design Extensions**

The proposed methodology could be expanded to include semi-quantitative or ‘soft’ information related to whether the load in any given unmonitored reach is above or below some threshold load. For example, if it is known that a nutrient source at a particular unmonitored location is below some load threshold, the downstream load predictions could easily be conditioned on this information by including an indicator variable for that reach. Such indicator techniques would help reduce downstream load prediction error at a fraction of the cost required to actually monitor the source in that reach.

The proposed design algorithm could also be easily reformulated to account for uncertainty in distributed sources and transport parameters Zj and Ti,j. The sequential design constraint would eliminate the need to consider strong spatial dependencies that invariably occur in such distributed processes. However, the computational burden associated with generating multiple realizations of areal sources or parameters would likely prove too large to justify application of the proposed uncertainty analysis to anything but the smallest of study areas.

**References**

Dettinger, M.D. and J.L. Wilson, First-order analysis of uncertainty in numerical models of groundwater flow, 1. Mathematical development, *Water Resources Research*, 17, 149-161, 1981.

Smith, R.A., G.E. Schwartz, and R.B. Alexander, Regional interpretation of water-quality monitoring data, *Water Resources Research*, 33, 2781-2798, 1997.