USGS
science for a changing world

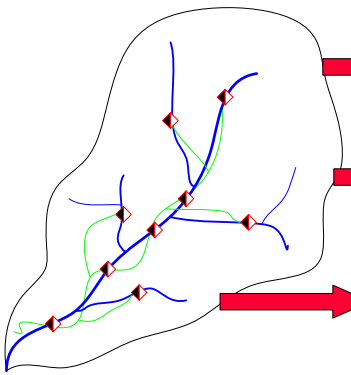
Heuristic Approach to Monitoring Network Optimization using Simulated Annealing and Multivariate Statistics

by Kenneth R. Odom, Ph.D., P.E., USGS Kentucky Water Science Center

1. Optimization Objectives and Data Needs

Why optimize this network?
- Reduce duplicated efforts by identifying highly redundant sampling sites.
- Provide a "smart" plan for discontinuation if budget constraints eventually force a reduction in network size.
- Assess the informational content of the historical data.
- Identify unique locations where additional monitoring may be needed.

What data are available to perform an optimization?



Primary Site Data
- Water quality
- Geology
- Morphology
- Vegetation
- Etc.

→ Principal components analysis (PCA)
Cluster Analysis (CA)
Discriminant Analysis (DA)

Ancillary Site Data
- Collocation
- Length of record
- Etc.

→ Applied directly to the composite score equation

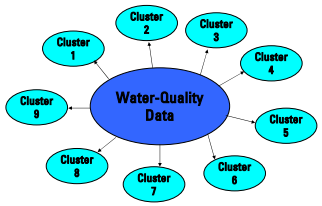
Cost data
- Trail distances
- Road distances
- Linkages between sites
- Sampling and lab costs
- Etc.

→ Calculate network costs

2. Measuring Benefits

METHODS: Principal Components Analysis (PCA), Cluster Analysis (CA), Discriminant Analysis (DA)

STEP 1. PCA is applied individually to each set of primary site data.
STEP 2. CA is performed on PC scores and on raw data for each data set. A mixture of hierarchical, non-hierarchical, and fuzzy clustering methods are used to ensure the proper number of clusters.
STEP 3. DA is performed using the cross-validation procedure on the final clusters using PC scores and raw data to test the discriminating ability of the clusters.
STEP 4. The members of each cluster are ranked. The lowest rank is assigned to the member that is closest to the cluster centroid.



Cluster Ranks

- Water Quality
- Geology
- Morphology
- Vegetation

Composite Score Equation

Water Quality, Geology, Morphology, Vegetation, Collocation (ancillary benefit)

$$\Psi_i = \omega_1 W_i + \omega_2 G_i + \omega_3 M_i + \omega_4 V_i + \omega_5 C_i$$

Apportionment Equation

$$BENEFIT_i = \frac{\Psi_i}{\Psi_{TOTAL}} \times BENEFIT_{TOTAL}$$

NOTE: A total monetary BENEFIT must be determined and scaled so that the total BENEFIT is greater than the total COST for the network.

3. Optimization: Simulated Annealing

SIMULATED ANNEALING – A heuristic process where the temperature of a body is heated to a temperature so that all molecular bonds have been broken. A controlled cooling is then slowly applied so that the molecules can rearrange themselves in a state of minimal energy.

PSEUDO-CODE
- Start with a network (P1). Calculate the objective function (OF) for P1.
- Randomly choose one site called Z from the collection of all sites in the network.
- If Z is in the network P1, test for removal of Z by naming a new network called P2 that includes all sites from P1 except Z. Calculate OF for the new network P2.
- If Z is out of the network P1, test for addition of Z by naming a new network called P2 that includes all sites from P1 plus Z. Calculate OF for the new network P2.
- If OF(P2) > OF(P1), then P2 replaces P1 as the current network.
- If OF(P2) < OF(P1), P2 might still replace P1 according to the Boltzmann probability. As temperature gets lower it becomes harder for a network to be accepted using the Boltzmann probability.

$$P(\Delta E) = e^{\frac{-\Delta E}{k_b T}}$$

- Return until the termination loop is satisfied.

Maximize the Objective Function

$$NETBENEFIT_p = \sum_p BENEFIT - \sum_p COST$$

A SIMPLE EXAMPLE for a network consisting of 10 sites 1,2,3,4,5,6,7,8,9,10

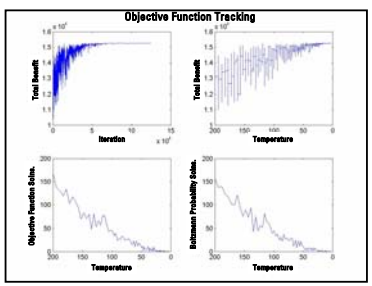
P1 = {4,5,7,10} Calculate OF(P1)

Randomly select a site from the existing network base (e.g. 1)

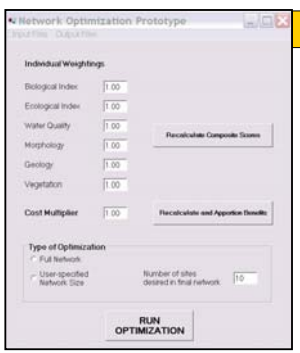
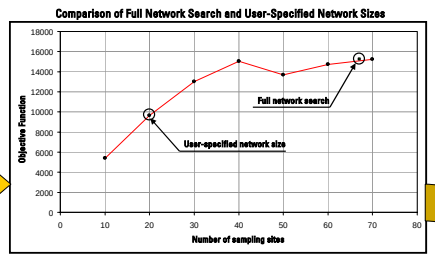
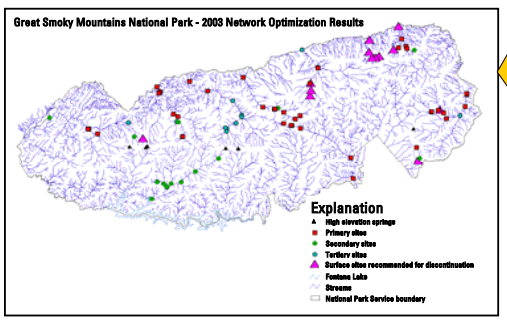
P2 = { ,4,5,7,10} Calculate OF(P2)

Is OF(P2) > OF(P1)?
If yes, then P2 replaces P1 as the current network.
If no, can P2 replace P1 based on the Boltzmann probability?
If yes, then P2 replaces P1 as the current network.
If no, then P1 remains the current network.

Return to beginning of process.



4. Optimization: Controls and Results

Simulated Annealing (SA) Controls

SA controls allows the modeler to quickly and easily perform multiple model runs for different weighting schemes and cost multipliers. A full network search can be selected or the modeler can specify the number of sites desired in the final network. This interface also streamlines the sensitivity analysis process.