



---

# Spatial Assessment and Optimization of the Synoptic Sampling Network in the Great Smoky Mountains National Park using Multivariate Techniques

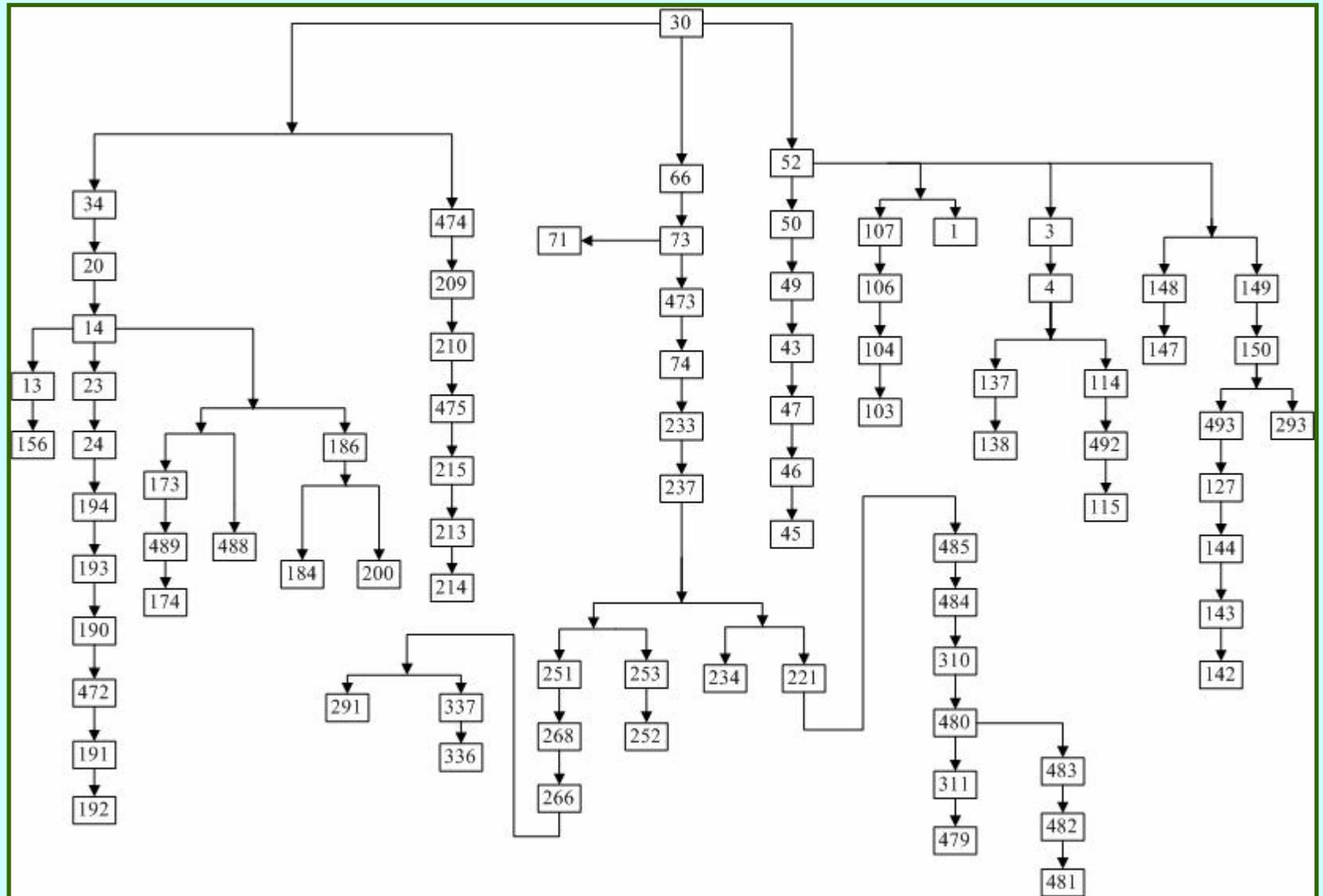
2004 National Water Quality Monitoring Conference  
Chattanooga, Tennessee

Kenneth R. Odom, PhD, PE  
R.B. Robinson, PhD, PE

# Objectives

- Determine the data-redundancy of water quality variables between the sampling sites using Principal Components Analysis (PCA), Clustering Analysis (CA), and Discriminant Analysis (DA).
- Assess the similarities of the sampling sites based on the geology, morphology, and vegetation of the watersheds in which the sampling sites are located using PCA, CA, and DA.
- Assign benefits to collocated sampling sites where auxiliary information on fish and benthic organisms is collected.
- Develop a simulated annealing (SA) optimization algorithm that will integrate the results of the above analyses to maximize the benefits and minimize the costs of the network by identifying sites that could be discontinued without a significant loss of information.

# Schematic Map of Sampling Sites





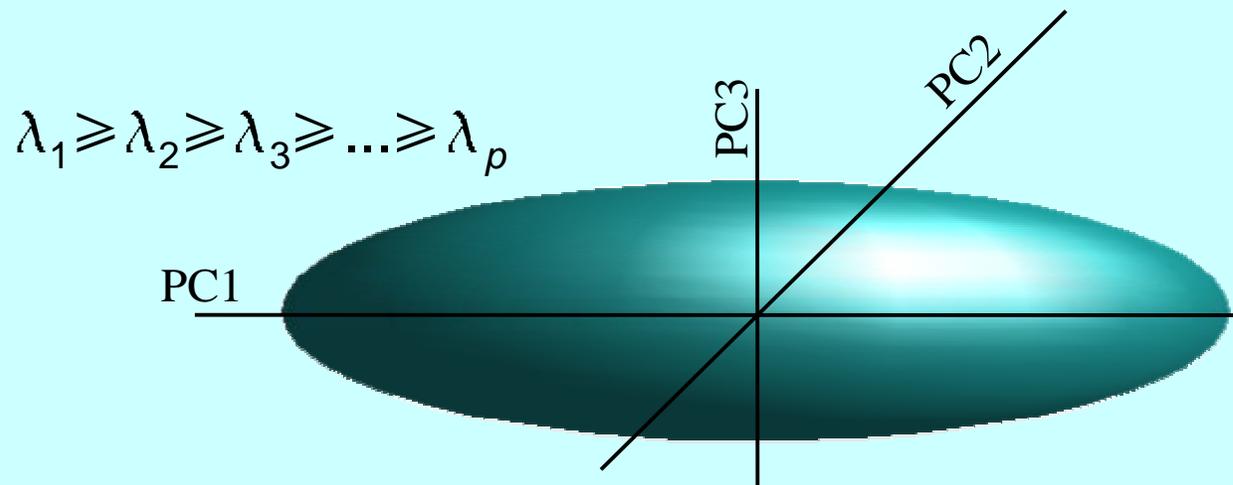
# Data

---

- **Water Quality – pH, ANC, conductivity, nitrate, sulfate, chloride, sodium, and potassium**
- **Quarterly grab samples**
- **Means for each of 83 sampling sites for the period from 1996-2001**
- **Watershed characteristics**
  - **Geology**
  - **Stream morphology**
  - **Vegetation**
- **Collocated sites (sites used by the National Park Service for other studies)**

# Multivariate Statistical Methods

- **Principal components analysis**
  - Transforms a set of correlated variables into a set of uncorrelated variables called principal components
  - Eigenanalysis performed on the correlation matrix
  - Eigenvalues quantify the variability that is explained by each principal component
  - $p$ -variables =  $p$ -principal components



# Multivariate Statistical Methods

## ➤ Cluster analysis

- Between cluster variance maximized
- Within cluster variance minimized
- Non-parametric “nearest neighbor”
- Cluster centroid distances are the “key”

## ➤ Discriminant analysis

- Tests the discriminating ability of the clusters
- Cross-validation method used ( not enough data available for a holdout sample)
  - Removes one observation at a time
  - Develops a new set of discriminant rules
  - Tests the removed observation to see if it can be classified into the original cluster



# Data Screening

---

- **Univariate and multivariate normality**
- **Notable outliers: sites 156, 174, 237, and 489**
  - **Identified using univariate boxplots and robust principal components analysis**
  - **Induced and masked multicollinearity**
- **Significant correlations identified in water quality, geology, stream morphology, and vegetation data**

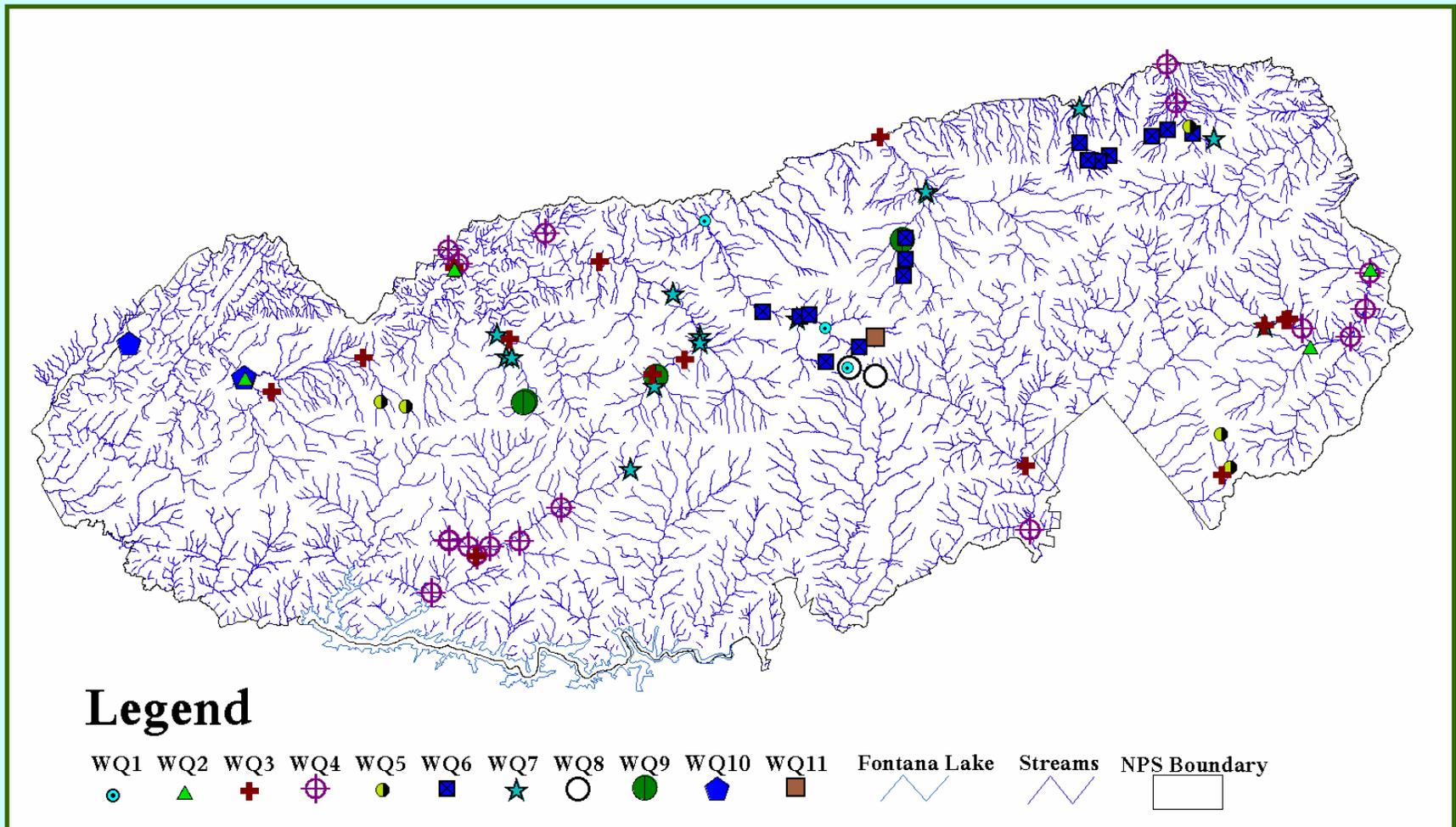


# Water Quality Data

---

- **Sampling sites 147, 156, 237, and 489 removed**
  - **Sites 147, 156, and 489 assigned to cluster 10**
  - **Sites 237 assigned to cluster 11**
- **Multivariate analysis performed on 79 of 83 sampling sites**
  - **PCA explained 86.4 percent of variability using PC's 1, 2, and 3**
  - **CA identified 9 clusters**
  - **DA correctly classified 90 and 95 percent of sites using the principal component scores and the original data**

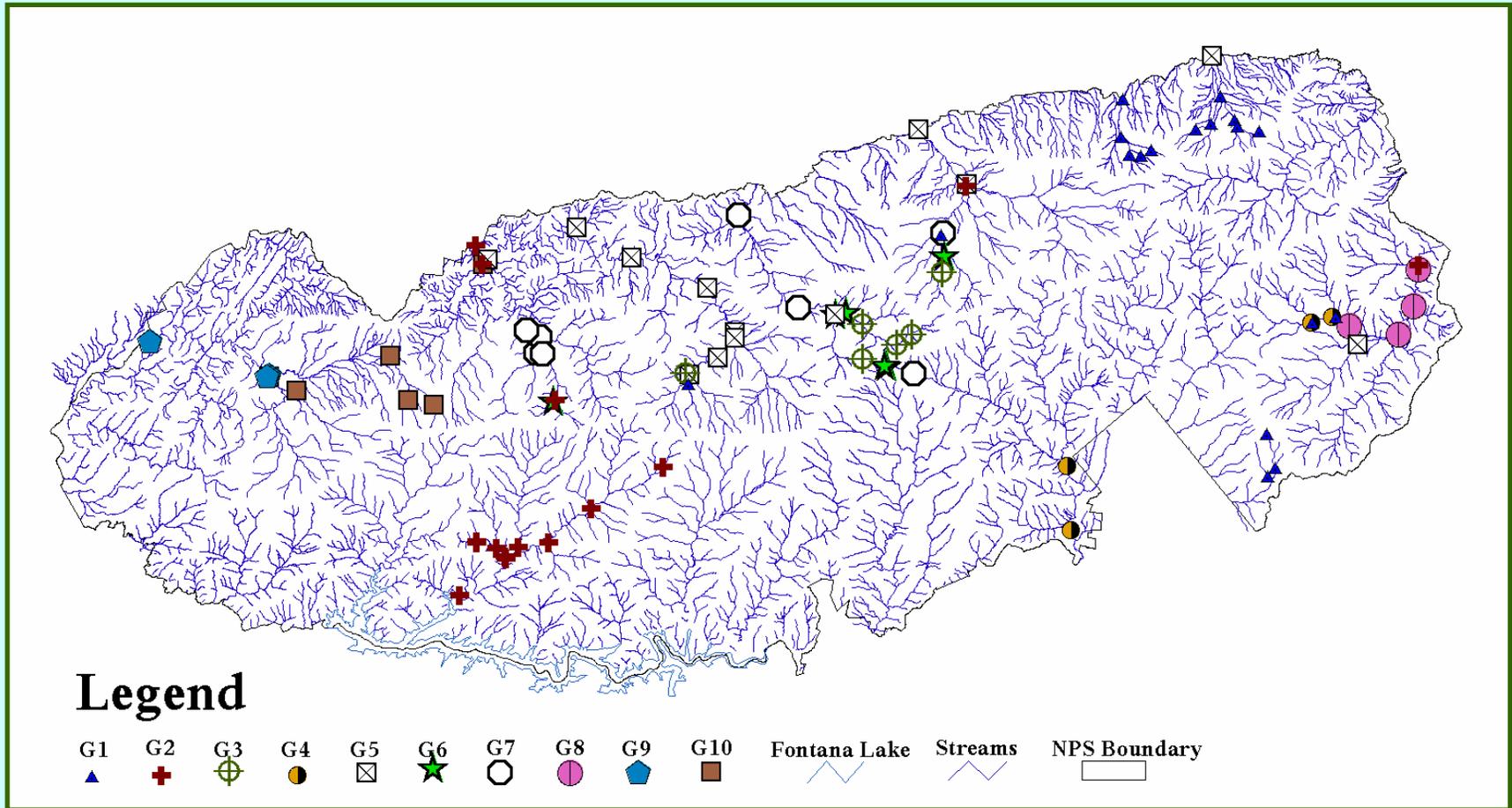
# Map of Water Quality Clusters



# Geology Data

- Initial PCA identified 5 PCs with eigenvalues greater than 0.7
- Principal variable analysis and multiple regression (using the 5 PCs from the initial PCA) resulted in removal of Great Smoky group
- Second PCA identified 4 PCs
- CA (*MODECLUS* and *FASTCLUS*) identified 10 clusters
- DA correctly classified 98.8 percent using the PCs and the original variables

# Map of Geology Clusters



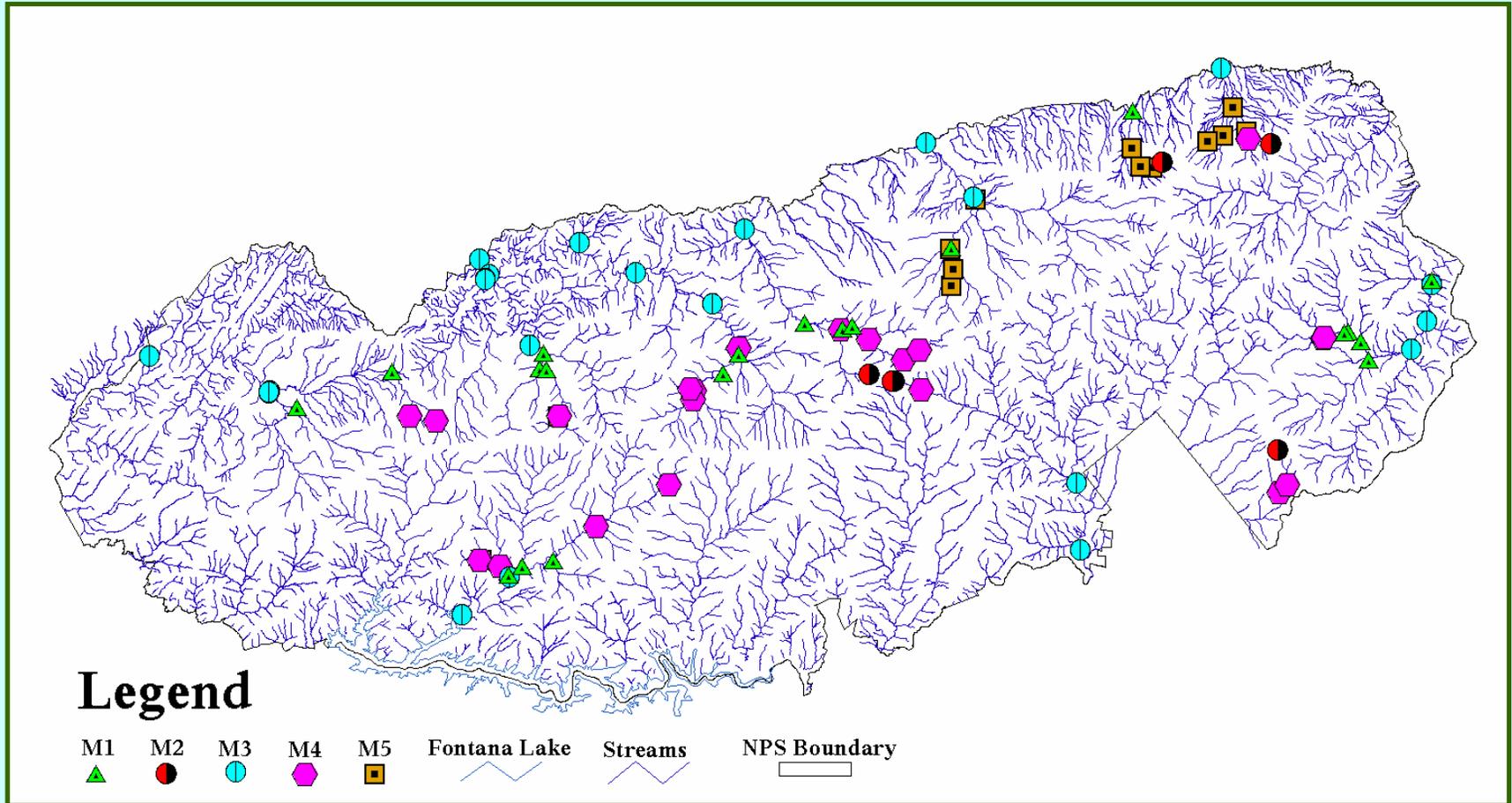


# Morphology Data

---

- **Initial PCA produced 3 PCs explaining 84.7 percent of the variability**
- **Second PCA using a smaller set of variables produced 3 PCs explaining 85.3 percent of the variability**
- **5 clusters were identified**
- **DA produced 98 and 90 percent positive classification rates using the PCs and the original variables**

# Map of Morphology Clusters



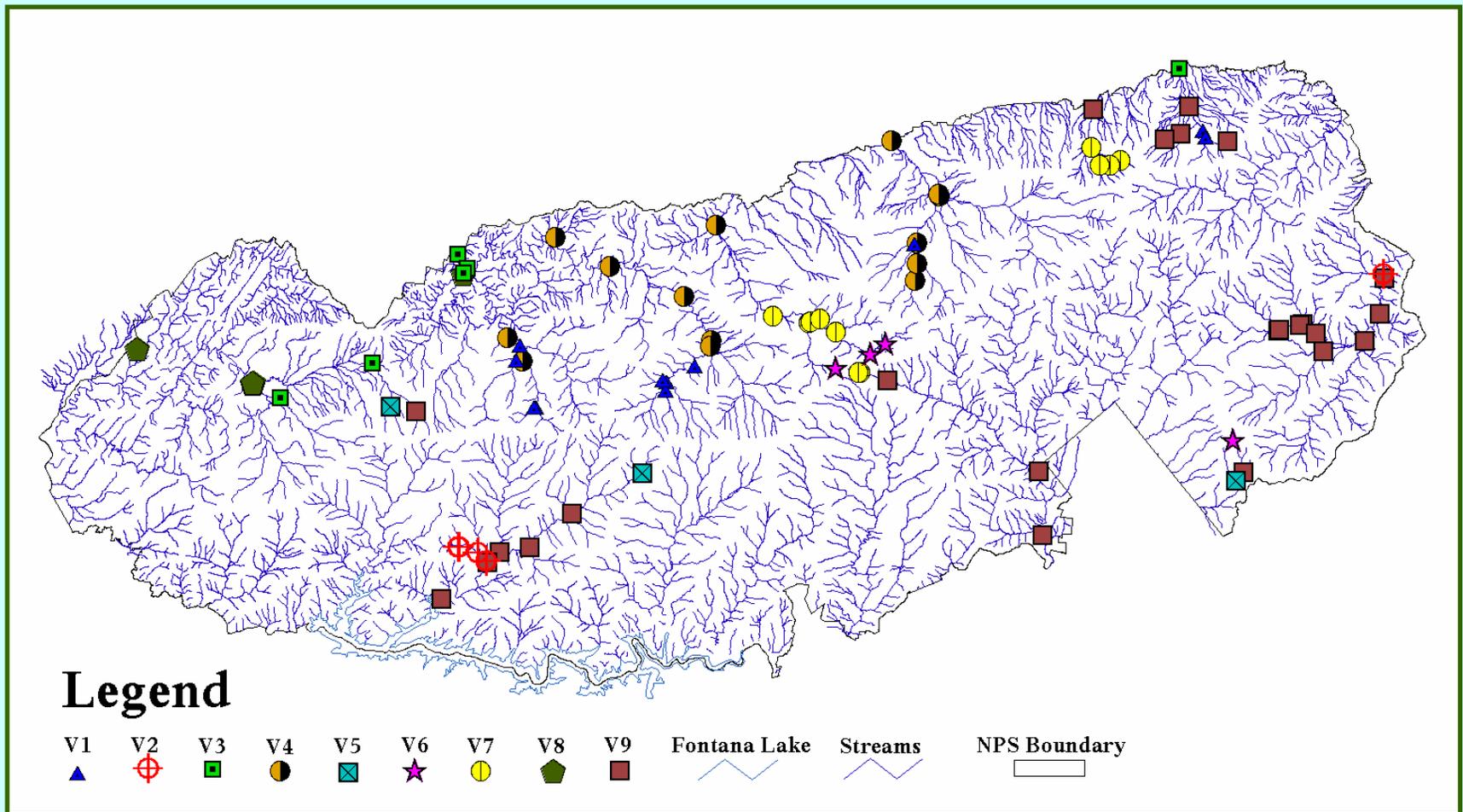


# Vegetation Data

---

- **PCA was used but in the end produced severely overlapping clusters**
- **Group-average hierarchical method and k-means were applied to 4 and 5 variable models identified by the PCA, principal variable analysis, and regression**
- **9 clusters were identified**
- **DA correctly classified 95 percent of the sites**

# Map of Vegetation Clusters



# Calculating Benefit Scores

- Sites with the greatest distance from their respective cluster centroid explains more of the variability than a site nearest the centroid
- Sites were ranked according to their distance from the centroid (greater distance = greater score)
- Site ranking was relative to the largest cluster to preserve small clusters
- Benefit scores are calculated by:

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

# Determining Costs of the Network

- Total network cost of \$69,200
- \$19,200 per year for access and sampling time (640 man-hours X \$30/man-hour)
  - Hiking
  - Driving
  - All-terrain vehicle
- \$50,000 per year for laboratory, technical, administration, and overhead (approx. \$602 per site/year)
- Cost of p-sites:

$$COST_p = \sum_p LABCOST + \sum_p ACCESS$$

# Determining Benefits of the Network

- **Total Benefit = 1.2 X \$69,200 = \$83,040**
  - **Basis: Benefit should outweigh cost**
  - **Basis: 20 percent return is a modest expectation**
  - ***BENEFIT*<sub>TOTAL</sub> = \$83,040**
- **Site benefit is calculated by:**

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

# Network Optimization

## ➤ Simulated annealing

- Heuristic method based on the thermodynamics of heating a body to a temperature such that all bonds have been broken between molecules
- Controlled cooling is then applied such that the molecules can arrange themselves to a minimal energy state
- Process is controlled by applying an annealing schedule
- Minimize or maximize an objective function

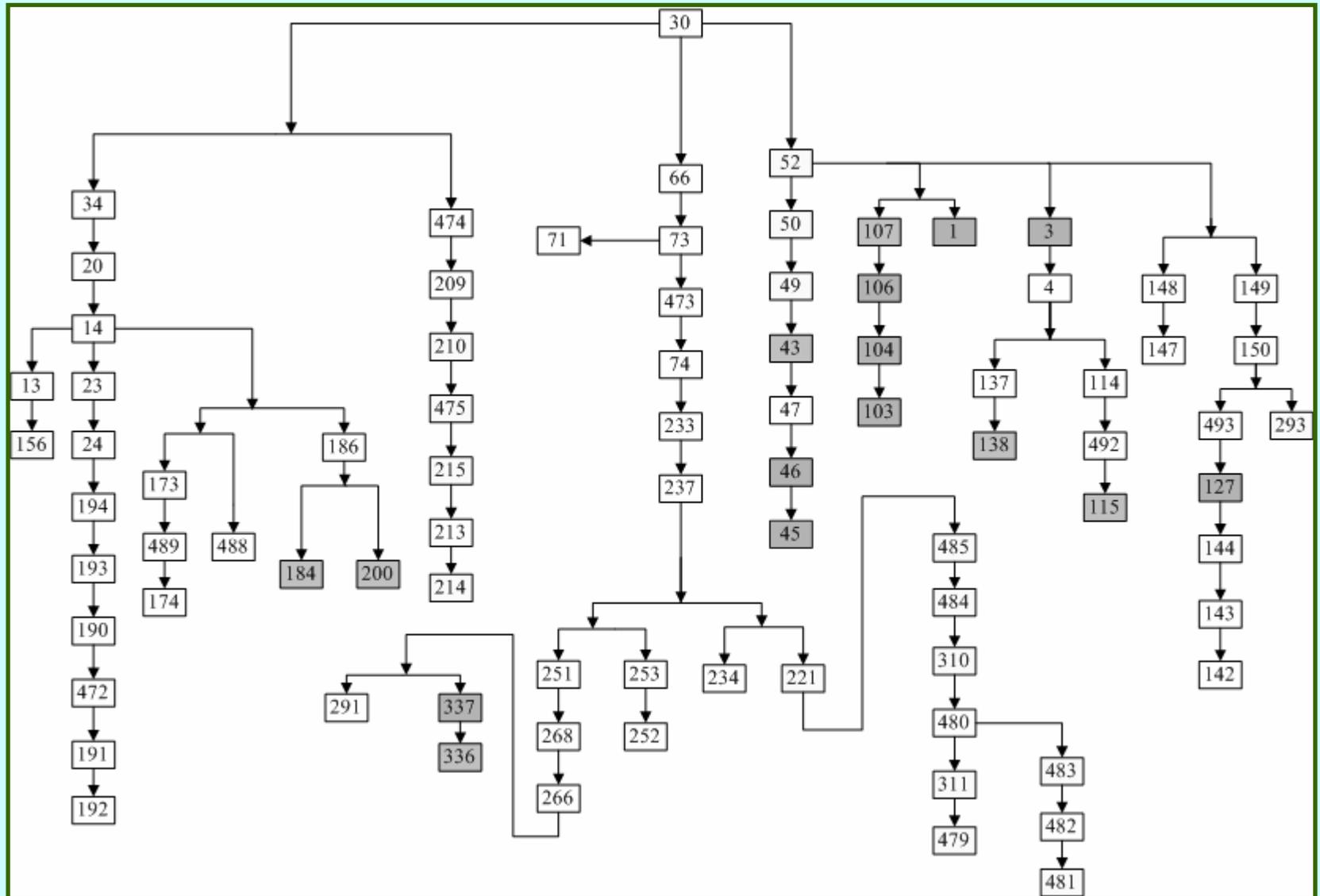
$$NETBENEFIT_p = \sum_p BENEFIT - \sum_p COST$$



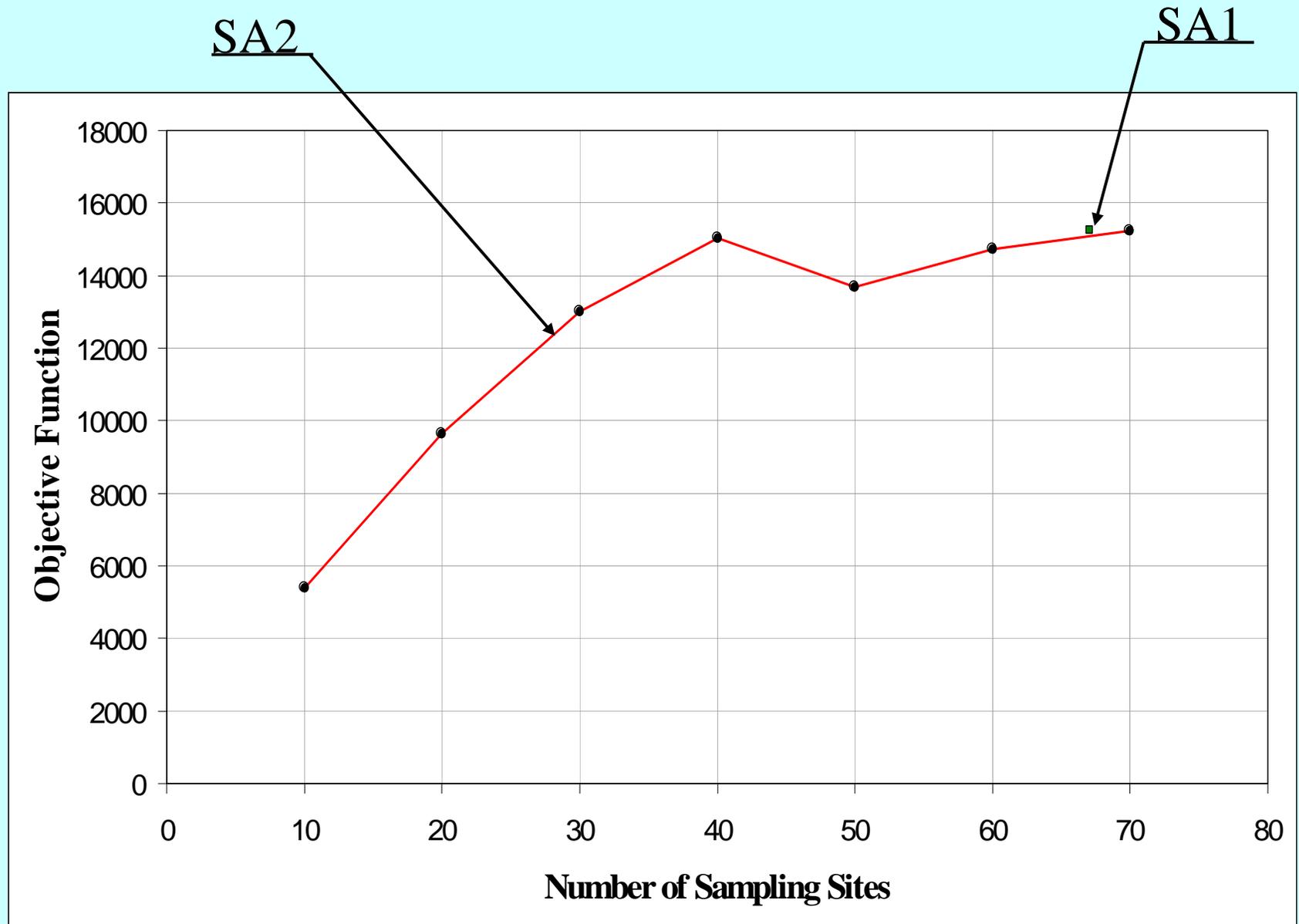
# Network optimization

- **Simulated annealing written in Matlab for two cases of optimization**
  - **First case (SA1) – Simulated annealing is performed on the network to determine the overall optimum network configuration.**
  - **Second case (SA2) – Simulated annealing is performed on the network using a user-specified (n) number of sites desired in the final network. The optimized network will contain exactly n-sites.**
    - **Provides a validation for SA1 results**
    - **Provides a logical format for considering other sampling sites to be retained or discontinued**

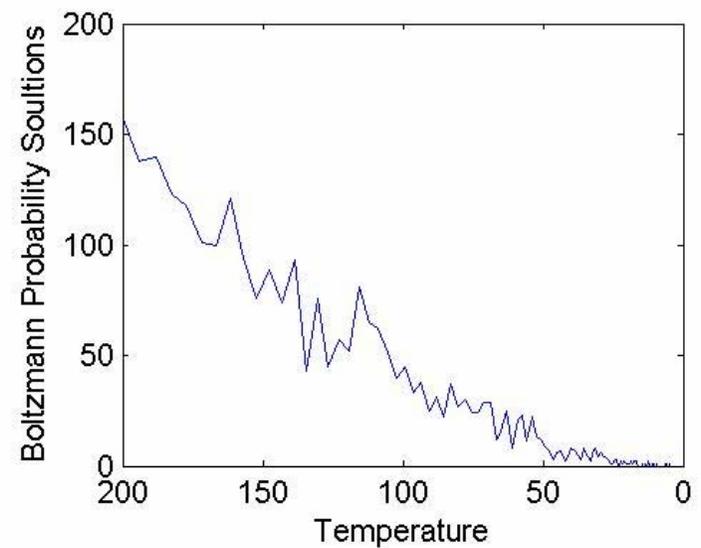
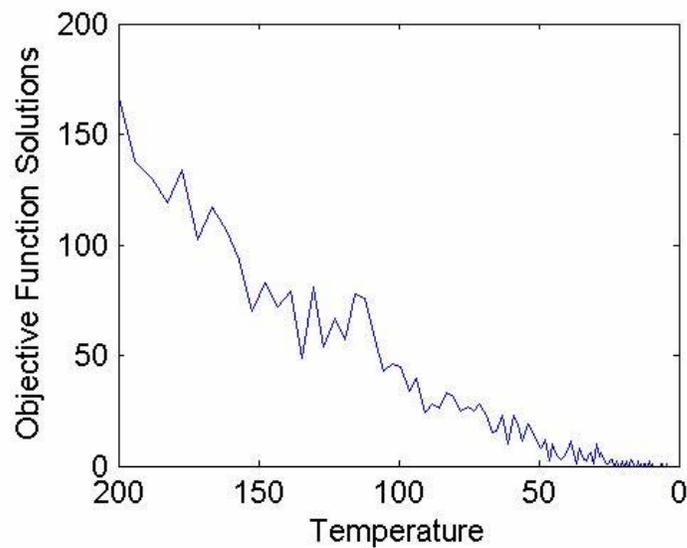
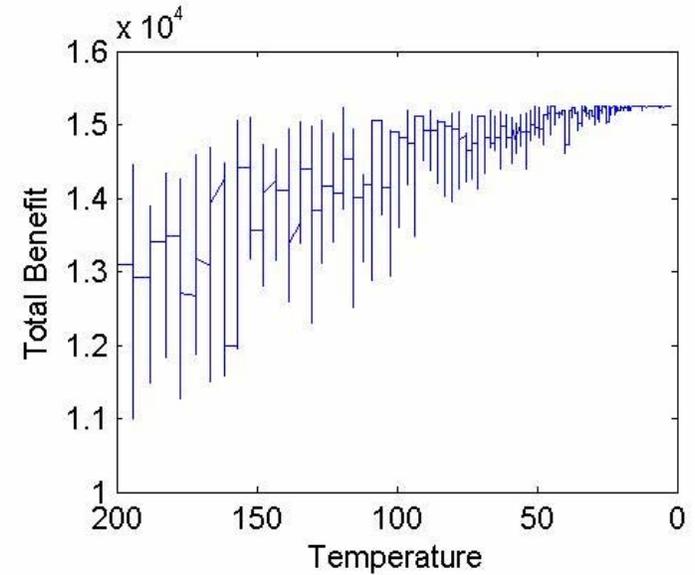
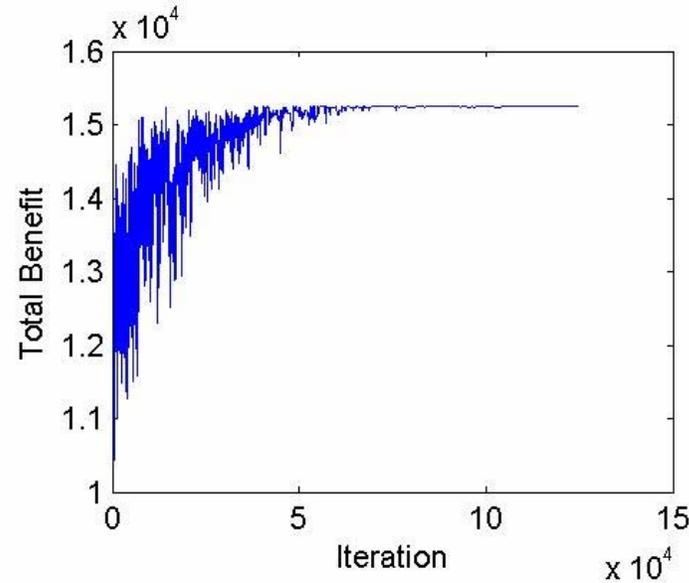
# SA1 results



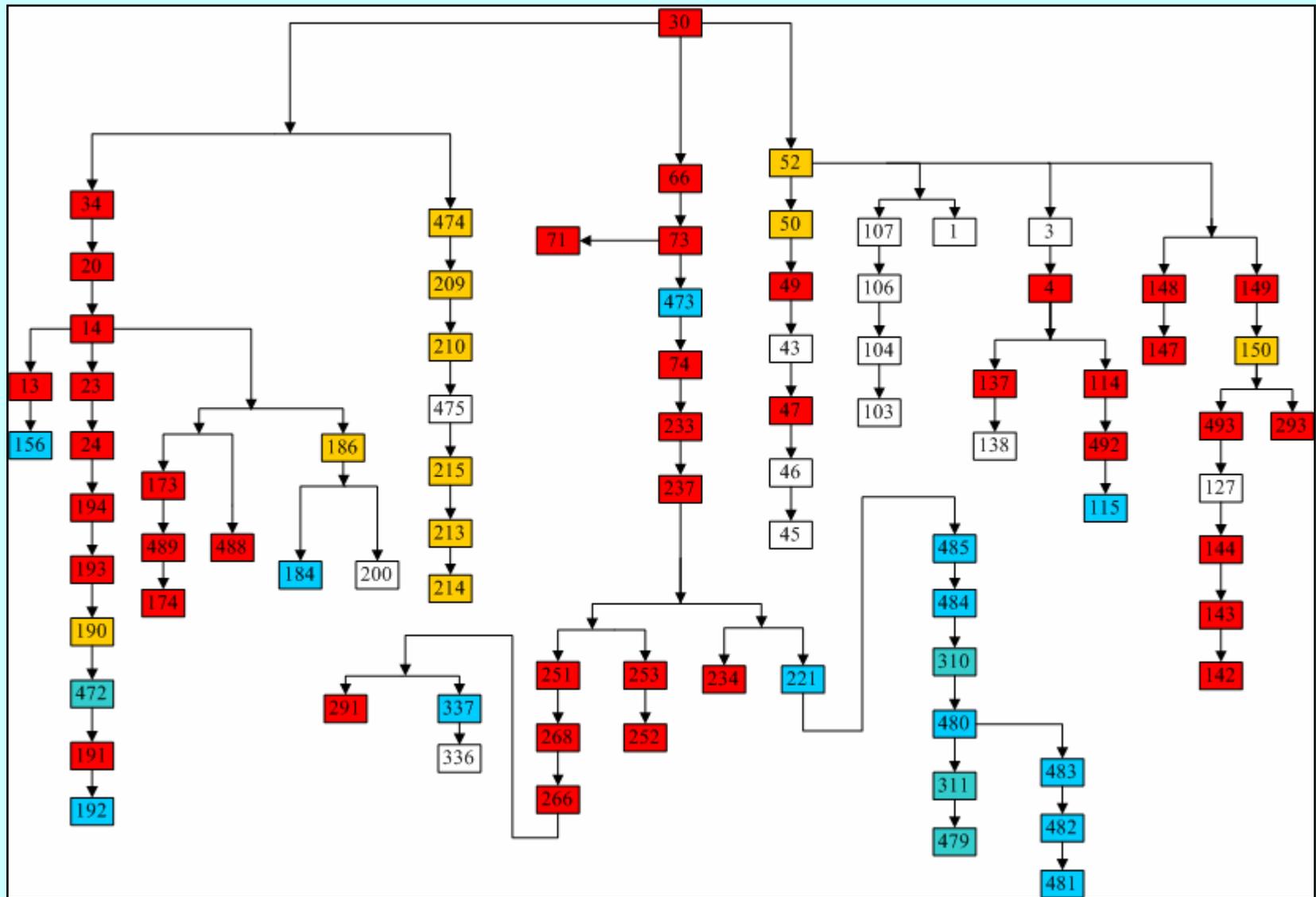
# SA2 results – $n$ best sites



# SA2 results – objective function tracking for $n=70$



# Schematic of the Redesigned Network



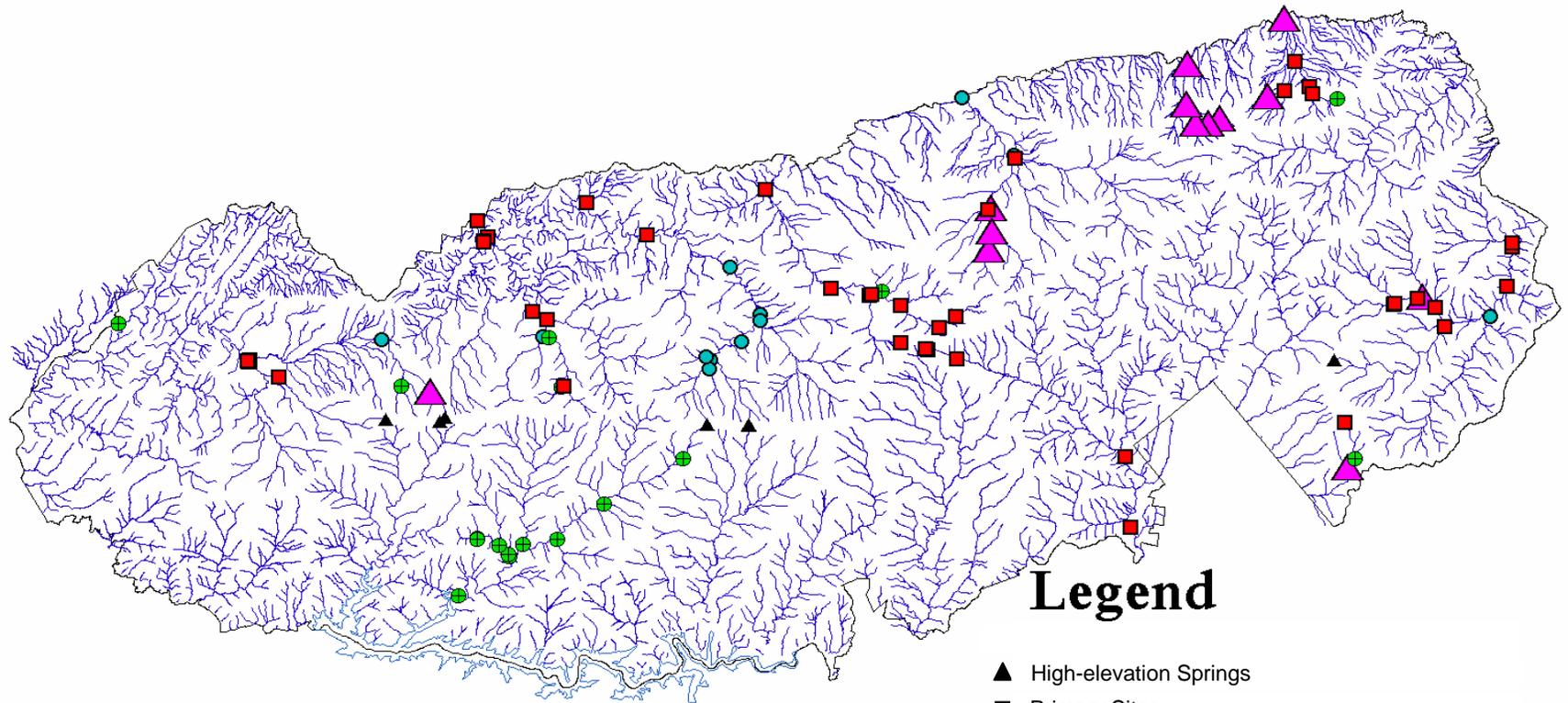
Primary

Tertiary

Secondary

Discontinue these sites first if needed

# Redesigned Network



## Legend

- ▲ High-elevation Springs
- Primary Sites
- ⊕ Secondary Sites
- Tertiary Sites
- ▲ Sites considered for discontinuation if necessary
- ∩ Fontana lake
- ∩ Streams
- NPS Boundary



# Final Considerations

---

- **Small clusters should remain intact – only clusters with large memberships should be targeted**
- **Ensure that all water quality, geology, morphology, and vegetation clusters are represented in the final network**
- **Each site represents a unique record of historical data. Careful consideration should be given before discontinuing any sampling site.**

# ACKNOWLEDGEMENTS

---

Great Smoky Mountains National Park Service

Steve Moore

Matt Kulp

University of Tennessee

Bruce Robinson, Ph.D.

Chris Cox, Ph.D.

Bill Seaver, Ph.D.

Bruce Tschantz, Ph.D.