Data Mining for Water Resource Management
Part 2 – Methods and Approaches to Solving Contemporary Problems

Edwin A. Roehl, Jr.[1] and Paul A. Conrads[2]

AUTHORS: [1]Chief Technical Officer, Advanced Data Mining Intl, 3620 Pelham Rd., PMB 351, Greenville, SC 29615; 864-201-8679; ed.roehl@advdmi.com; [2]Surface Water Specialist, U.S. Geological Survey South Carolina Water Science Center, Gracern Rd, Suite 129, Columbia, SC 29210; 803-750-6141; pconrads@usgs.gov

REFERENCE: *Proceedings of the 2010 South Carolina Water Resources Conference,* held October 13-14, 2010, at the Columbia Metropolitan Convention Center.

Abstract. This is the second of two papers that describe how data mining can aid natural-resource managers with the difficult problem of controlling the interactions between hydrologic and man-made systems. Data mining is a new science that assists scientists in converting large databases into knowledge, and is uniquely able to leverage the large amounts of real-time, multivariate data now being collected for hydrologic systems. Part 1 gives a high-level overview of data mining, and describes several applications that have addressed major water resource issues in South Carolina. This Part 2 paper describes how various data mining methods are integrated to produce predictive models for controlling surface- and ground-water hydraulics and quality. The methods include:

- signal processing to remove noise and decompose complex signals into simpler components;
- time series clustering that optimally groups hundreds of signals into "classes" that behave similarly for data reduction and (or) divide-and-conquer problem solving;
- classification which optimally matches new data to behavioral classes;
- artificial neural networks which optimally fit multivariate data to create predictive models;
- model response surface visualization that greatly aids in understanding data and physical processes; and,
- decision support systems that integrate data, models, and graphics into a single package that is easy to use.

INTRODUCTION

Data mining is a relatively new science that assists in converting large databases into knowledge (Weiss and Indurkhya, 1997), and is uniquely able to leverage the real-time, multivariate data now being collected for hydrologic systems. In side-by-side comparisons with state-of-the-art physics-based hydrologic models, data-mining solutions have been substantially more accurate, less time consuming to develop (Conrads and Roehl, 1999; Conrads and Greenfield, 2010), and embeddable into spreadsheets and sophisticated decision support systems, making them easy to use by regulators and stakeholders.

This is the second of two papers that describe how data mining can aid natural-resource managers with the difficult problem of controlling the interactions between hydrologic and man-made systems in ways that preserve resources while optimally meeting the needs of disparate stakeholders. Part 1 gives a high-level overview of data mining, and describes several applications in South Carolina. Part 2 describes how various data mining methods are integrated to produce predictive models for controlling surface- and groundwater systems.

DATA MINING CONCEPTS AND METHODS

Periodicity, Chaos, Noise and Signal Decomposition

Process signals exhibit three types of behavior - periodic, chaotic, and noise (random) that are superposed. For example, coastal water level exhibits "multiply-periodic" behaviors caused by the gravitational interactions of the earth, moon, and sun. It also is affected by chaotic and random influences such as wind and storms. Theoretically, periodic behavior repeats itself perfectly, making it perfectly predictable. Examples of periodic behavior include diurnal (24-hour) and seasonal ambient temperature cycling, and human impacts on water resources controlled by the workday, the workweek, and the seasons, for example, irrigation and power generation.

Chaos Theory (Abarbanel, 1996) studies physical processes that are highly sensitive to small changes in boundary conditions. These processes can flip-flop between different behaviors with little apparent cause. Weather is a chaotic process that affects nearly everything, including industrial processes such as water and wastewater treatment, making them also chaotic. Chaotic processes are somewhat predictable and special methods have been developed for analyzing and modeling them.

"Signal decomposition" provides a quantitative accounting of the predictable and unpredictable. Figure 1 illustrates signal decomposition, which filters raw signals into "components" that manifest a signal's different behaviors. At upper right, a fast Fourier transform (FFT) generates a "spectral signature" of a tidally forced, water-level signal. The spectral “peaks” at ½, 1, and 2 times the 12.4-hour tidal cycle result from the earth’s rotation and the 28-day lunar orbit. An FFT-based filter was then used to split the raw signal into high and low frequency (*f*) components. The high *f* components are predominantly the ones identified in the spectral signature. The midsize “humps” in the low *f* components occur every 7 days and are caused by upstream hydroelectric generation. The remaining low *f* components are predominantly chaos and noise. Chaotic components are separated from noise with more difficulty using empirical model-based filters. Removing all of the predictable periodic and chaotic components leaves behind the unpredictable noise.

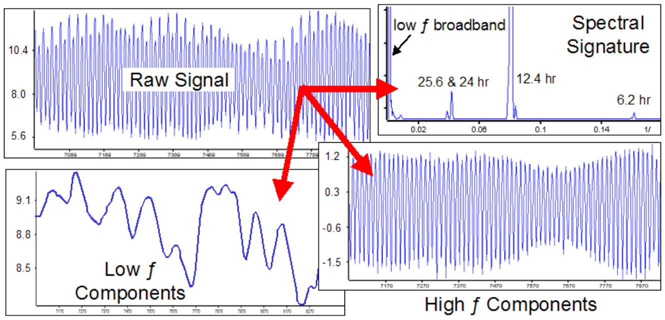

Figure 1. Signal decomposition reveals causes of estuary water-level variability.

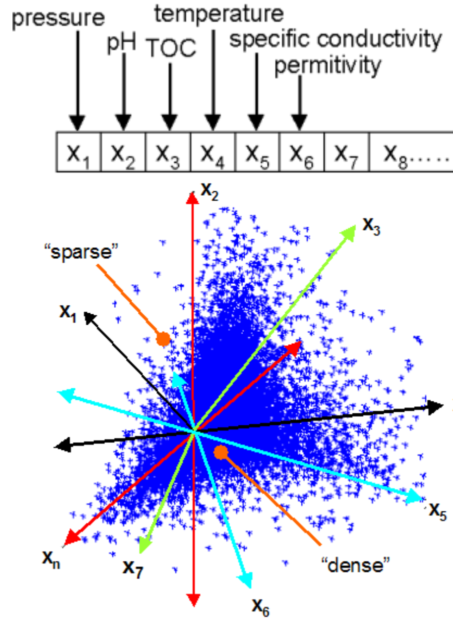

Figure 2: n-dimensional, multivariate state vector (top) lying in an n-dimensional state space.

State Vectors and State Space

Typically, the behavior of a variable x(t) is represented by a time series of values measured at constant time intervals, for example, once per minute. Trend plotting shows how x(t) changes over time. A value of x can be forecasted at a future time by fitting a line or curve to recent measurements and extrapolating forward. The problem with this “univariate” approach is that it is weakly analytical and employs only one variable at a time.

Chaos Theory, like trending, uses multiple measurements to characterize process behavior. Chaotic processes are said to transit from one unique “state” to another in time, whereas periodic processes repeat the same states. Figure 2 shows that a state is characterized by a “state vector”. The vector’s “features” x_n can represent one or more variables. Multiple measurements from the same variable can be assigned to different features to represent its trend. Each vector feature represents a different dimension in an “n-dimensional state space”. Densely populated regions of state space are generally more understandable and modeled than sparse regions.

The three dimensional (3D) scatter plot in Figure 3 shows that as a process changes in time, it leaves a “track” of points in n-space, representing a “state history”. The history can be used to develop an empirical, predictive process model by curve fitting an n-dimensional function. Here, of interest are the salinity intrusion events on the Lower Savannah River Estuary indicated by spiking specific conductance (SC) at an inland gage. The spikes are seen to coincide with low freshwater flow (Q) and high ocean water level (WL).

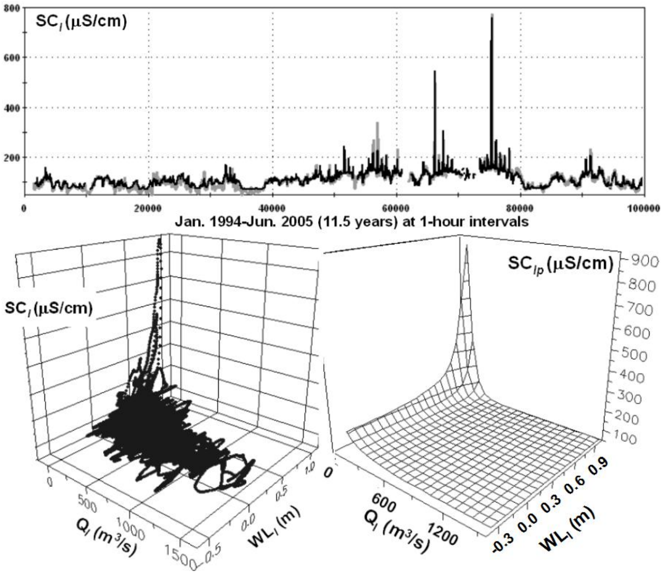

Figure 3. Modeling Savannah River Estuary seawater intrusion. Lower left – three dimensional scatter plot of specific conductance (SC) versus with freshwater flow (Q) and sea water level (WL). Above - measured (gray) and ANN-predicted (black) SC. Lower right - ANN response surface fitted to scatter plot data.

Developing Accurate, Predictive Process Models

To answer complex questions such as, “Which variables most affect the process?”, or “What will happen if I do this instead of that?”, one needs a virtual process, also known as a model, to poke and probe for answers. Modeling is the development of a mathematical function that, for vectors of input values, will calculate a set of output predictions.

Calibrating a model involves fitting vectors with predetermined or synthesized mathematical functions. Examples of predetermined functions are lines, n-dimensional hyperplanes, and physics-based models, whose coefficients are manipulated to provide the best possible fit. A benefit of predetermined functions is their rigorous mathematical foundation, but their downside is poor accuracy when they are functionally unable to fit the data. Synthesized functions employ "machine learning" methods such as multivariate adaptive regression splines (Friedman, 1991), and the method discussed here - multi-layer perceptron artificial neural networks (ANN; Jensen, 1994), to better fit calibration data with nonlinear "hypersurfaces", making them more accurate predictors for some problems. Above in figure 3 are plots of the aforementioned measured SC data with ANN predictions having an R^2 more than 8 times higher than a state-of-the-art physics-based model of the same system. At lower right in figure 3 is a 3D projection, called a "response surface", of the higher dimensional, nonlinear hypersurface fitted by an ANN to the salinity intrusion data. Response surfaces clearly reveal the relations among variables learned by the ANN to provide knowledge about a process's physics.

Modeling is the inverse of signal decomposition because the goal is to synthesize a new signal, a prediction of an output variable, from multiple input signals. Modeling is an iterative process involving multiple steps. Raw signals are cleaned up to remove unreliable measurements. Complex signals are then “decomposed” into multiple, simpler components whose behaviors can be ascribed to identifiable causes. Candidate input components are checked for relative independence and culled or decorrelated if necessary before being used in models. Because output signals themselves have multiple components, they are seldom modeled with a single empirical function. Instead, a “sub-model” is used to model each output component using the most appropriate input components. In most cases, a sub-model will have only a single output. As shown in figure 4, the outputs of sub-models are combined into “super-models” to synthesize an overall prediction of the output variable. This systematic “divide-and-conquer” approach reveals the intricacies of a process, and enforces rigor to ensure that the super-model is as accurate and representative of the actual process as possible.

Clustering and Classification

Modeling spatially expansive natural systems is difficult because behaviors vary discontinuously both spatially and in time. These problems require the integration of large numbers of categorical and time-series variables, and reducing them to a select set with maximum predictive capability, preferably without subjectivity in a numerically optimized way. Figure 5 shows that hydrographs of monitoring wells in the Floridan aquifer system can vary greatly over short distances, indicating differences in their underlying process physics. Often the causes of such differences are unknown, making the employment of physics-based models problematic.

Roehl and others (2006-1) describe another divide-and-conquer modeling approach that employs "time-series clustering" as a method for optimally clustering large numbers of signals into "classes", whose "members" behave similarly. Figure 6 shows the hydrographs of two of the 12 classes used. Note how much alike the member hydrographs of a class are, and how dissimilar they are class-to-class. Each class is then modeled with a "spatially-interpolating" ANN sub-model that incorporates categorical inputs, such as monitoring site descriptors and spatial coordinates, and dynamic inputs derived from signals such as rainfall. The super-model of the entire system is composed of the class sub-models. Predictions at a new site, not used in model development, are made by first assigning it to a class using a "classification algorithm" that employs the site categorical variable descriptions, and then running the appropriate sub-model.

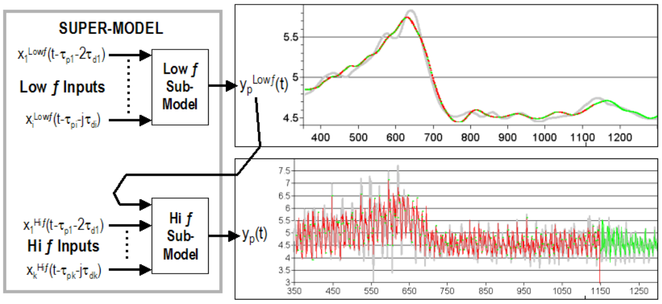

Figure 4. Super-model composed of two sub-models. Gray trends at right are measurements and the red and green trends are predictions made using calibration and testing data, respectively. At upper left, a Low frequency (f) sub-model predicts low f components $y_p^{Low f}$ of the output variable y from input low f components $x^{Low f}$. $y_p^{Low f}$ is then input with other high f components $x^{Hi f}$ to the Hi f sub-model to predict y_p.

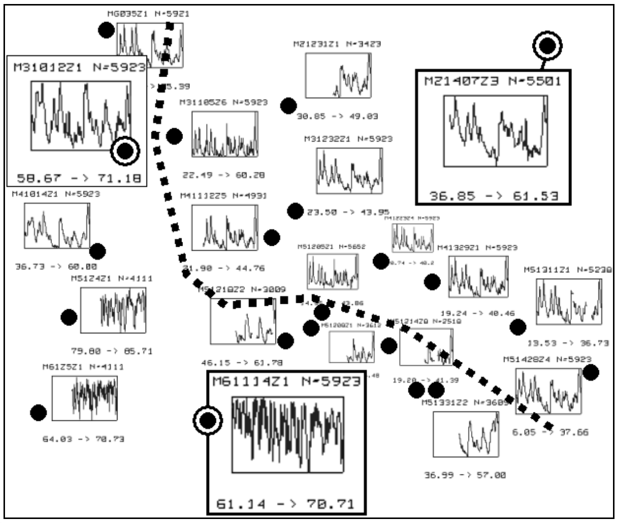

Figure 5. 18-year hydrographs in the Floridan aquifer system. The wells shown cover a 30x50 square kilometer sub-region of an approximate 100x100 km^2 monitoring network. Dotted line marks the Suwannee River.

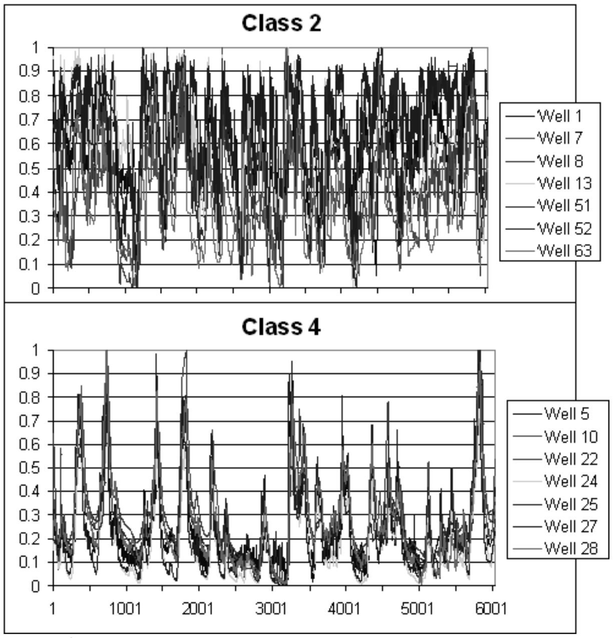

Figure 6. Normalized WLs for Classes 2 and 4 of 12 total classes. x-axis is approximately 18 years of days.

Decision Support Systems

The collective interests and computer skills of resource managers, scientists, and other stakeholders can be quite varied. A decision support system (DSS) provides a means to effectively transform arcane databases and models into information that is equally accessible to all stakeholders for cooperative, informed decision-making (Roehl and others, 2006-2). Important features of DSSs include:

- accurate predictive models;
- databases that describe historical behaviors;
- model controls for running *"What if?"* scenarios;
- graphical user-interfaces that integrate the DSS components with user controls and graphical output;
- "constrained optimization" that couples a search routine to the model to determine the input scenario that provides the best predicted outcome; and
- expert knowledge such as water-quality standards and expert hydrology rules.

CONCLUSIONS

Data mining methods constitute a divide-and-conquer approach to solving complex hydrologic and water-quality problems. They have been successfully employed on many projects in South Carolina and elsewhere, and can provide the knowledge and tools to solve problems that are unsolvable by other means. The solutions they provide are inherently adaptive and easily updated when new data become available. They are easily deployed to end-users in spreadsheets, DSSs, or other types of off- or on-line computer program.

LITERATURE CITED

Abarbanel, H.D.I., 1996, *Analysis of Observed Chaotic Data*, Springer-Verlag New York, Inc., New York, 4-12, 39.

Conrads, P.A. and E.A. Roehl, 1999, Comparing physics-based and neural network models for predicting salinity, water temperature, and dissolved oxygen concentration in a complex tidally affected river basin, South Carolina Environmental Conference, Myrtle Beach, March 1999, 7p.

Conrads, P.A., and Greenfield, J.A., 2010, Potential mitigation approach to minimize salinity intrusion in the lower Savannah River Estuary due to reduced controlled releases from Lake Thurmond, Conference Proceedings Paper for the 4th Federal Interagency Hydrologic Modeling Conference Las Vegas, NV June 2010, 9p.

Friedman, J.H., 1991, "Multivariate Adaptive Regression Splines", Annals of Statistics, 19, 1-141.

Jensen, B.A., 1994, Expert Systems - Neural Networks, *Instrument Engineers' Handbook Third Edition: Chilton*, Radnor PA.

Roehl, E., Risley, J., Stewart, J., Mitro, M., 2006-1, Numerically optimized empirical modeling of highly dynamic, spatially expansive, and behaviorally heterogeneous hydrologic systems – part 1, iEMSs 2006 Summit on Environmental Modelling and Software, Burlington VT, June,2006, 6p.

Roehl, E., Risley, J., Stewart, J., Mitro, M., 2006-2, Features of advanced decision support systems for environmental studies, management, and regulation, iEMSs 2006 Summit on Environmental Modelling and Software, Burlington VT, June,2006, 6p.

Weiss, S.M. and Indurkhya, N., 1997, *Predictive Data Mining: A Practical Guide*, Morgan Kaufmann.