An environmental sensor network to determine drinking water quality and security

Anastassia Ailamaki
School of Computer Science
Carnegie Mellon University
natassa@cmu.edu

Christos Faloutsos
School of Computer Science
Carnegie Mellon University
christos@cs.cmu.edu

Paul S. Fischbeck
Engineering and Public Policy and Social and Decision Sciences
Carnegie Mellon University
pf12@andrew.cmu.edu u

Mitchell J. Small
Civil & Environmental Engineering and Engineering & Public Policy
Carnegie Mellon University
ms35@andrew.cmu.edu

Jeanne VanBriesen
Civil and Environmental Engineering and Biomedical Engineering
Carnegie Mellon University
jeanne@cmu.edu

Abstract

Finding patterns in large, real, spatio/temporal data continues to attract high interest (e.g., sales of products over space and time, patterns in mobile phone users; sensor networks collecting **o***perational data from automobiles, or even from humans with wearable computers). In this paper, we describe an interdisciplinary research effort to couple knowledge discovery in large environmental databases with biological and chemical sensor networks, in order to revolutionize drinking water quality and security decision making. We describe a distribution and operation protocol for the placement and utilization of in situ environmental sensors by combining (1) new algorithms for spatial-temporal data mining, (2) new methods to model water quality and security dynamics, and (3) a sophisticated decision-analysis framework. The project was recently funded by NSF and represents application of these research areas to the critical current issue of ensuring safe and secure drinking water to the population of the United States.*

1. INTRODUCTION

The need for research to link the power of information technology and decision making with the complexity of environmental problems is compelling in its logic and motivation, but quite challenging in its demand for interdisciplinary skills and knowledge. Environmental problems involving water quality and security have a wide scope of the entire planet, a daunting complexity at both the microscopic and ecosystem level, and a profound relevance to our daily lives – clean waterways and secure water supply are our best protection from communicable disease and the effects of chemical and biological contaminants either accidentally or intentionally released to our environment. These centrally important problems, with their high degree of knowledge domain inter-connectedness and varying scales of spatial and temporal aggregation, are greatly in need of, and ideally suited for, new techniques in analysis and interpretation that are emerging in the area of information technology and decision making.

To address these problems, joint research is needed between environmental science and data analysis for decision-making. This research is made possible by three trends. First, there has been a rapid expansion of large environmental databases, easily accessible, from the U.S. Environmental Protection Agency (EPA), U.S. Geological Society (USGS), and other agencies. Second, key developments in IT research and fast, scalable implementation of older tools from machine learning and statistics enable more extensive inferences to be drawn from these data, thus improving decision-making potential. Third, the pace of development in biological and chemical sensing technology indicates that low-cost, easily emplaced environmental sensors will be available soon. The first two trends make possible, and the third makes imperative, research to develop a distribution protocol for these sensors, a data-collection and storage protocol for the potential data that they will supply, and an integrated model to interpret and "mine" the data made available by this extensive "pulse-taking" of the natural environment. Further, security concerns regarding detection of intentional contamination of drinking water have focused our attention on the need for enhanced sensing, data evaluation, and decision-making within drinking water distribution systems.

The goal of the research presented in this paper is to "leapfrog" over the current limitations in sensor design, capabilities, and cost and to look forward to the day when these technologies are ready for deployment. Instead of additional research in the already crowded field of those developing better, cheaper, longer-lasting sensors, we evaluate current information on drinking water treatment and distribution systems in order to inform the eventual placement and use of these sensing technologies. The scenario that we want to avoid is the sub-optimal placement and uncoordinated use of advanced sensing technologies within the natural and engineered environment. In addition to current cost limitations, in situ sensor placement always carries with it the potential to affect the environment during sensor placement or malfunction. Likewise, inefficient or disorganized monitoring, data-storage, and data-evaluation protocols may result in data that are unusable for regulatory decision-making, planning, or rapid response to an external threat. Our goal is to avoid these start-up problems that may be associated with poor distribution and utilization of *in situ* sensing technologies. We produce a distribution and operation protocol for the placement and utilization of in situ environmental sensors by combining (1) new algorithms for spatial-temporal data mining, (2) new methods to model water quality and security dynamics, and (3) a sophisticated decision-analysis framework.

In this work, we describe the nature of the sensor placement and data structure challenge associated with water security monitoring. Further, we describe the integrated research approach we are undertaking to advance spatial-temporal mining of environmental data.

2. RELATED WORK

2.1 Spatial-Temporal Data Mining

Spatial-temporal data mining is an emerging field. Traditional data mining works on Association Rules[1], looking for patterns of the form: "Customers who buy bread, also buy milk, with probability x%." There is significant literature in this arena (see for example, the recent book by Han and Kamber[2]). Extensions to temporal and spatial patterns are limited, typically looking for rules of the form "customers who buy cars now, will buy tires in 2 years."[3] Algorithms for spatial patterns are very limited, using ad-hoc thresholds and neighborhood radii. We use multi-resolution algorithms that will search for patterns for all scales of time and space, using fractals. Related work includes similarity search in time sequences, where the typical algorithm is to approximate each time sequence with its first few Fourier coefficients[4,5] or with a piece-wise linear approximation.[6] Our work will use more sophisticated matching algorithms to detect correlated sequences. Specifically, we will use (a) multivariate linear regression, for co-evolving time sequences and (b) non-linear models, to capture non-linear phenomena (like bursty/auto-catalytic/self-similar traffic patterns, as well as time sequences that exhibit non-linear correlations).

2.2 Water-quality Monitoring and Modeling

The National Academy of Engineering declared water and wastewater treatment as one of the top five engineering accomplishments of the 20th century. Water-quality improvements related to sewage treatment awakened society to the potential for water transmission of biological and chemical contaminants, and amendments to the Safe Drinking Water Act continue to increase the number of chemicals that we monitor in our drinking water. The recent security concerns regarding drinking water have added to the complexity of monitoring and decision-making; biological detection and early warning are critical for water security. Monitoring water quality has always been a labor-intensive process, and most critical water-quality parameters (e.g., presence of pathogenic microorganisms) require sampling and ex situ analysis. These data are collected within the distribution system by individual water providers for regulatory compliance. Rather than duplicating the data collection or organization provided by these agencies, we focus on knowledge development through analyzing existing databases using classic data-mining algorithms and the development of new algorithms designed to handle the unique spatial and temporal variability in environmental datasets.

3. PROBLEM STATEMENT

The goal of the present project is the development of new data-mining techniques for knowledge discovery in water-quality databases and the design of an implementation strategy for using this knowledge discovery, related watershed and water distribution models, and a decision framework, to inform the development and placement of in situ sensor networks. Thus, our goal is not specific to one environmental problem, but will produce a generalizable solution to the issue of interpreting and using environmental data. For the purposes of demonstrating the types of issues that might benefit from this work and for validation we present a possible scenario of use within a water distribution system in this section. This example is only one of many important relationships in water systems that we will address in order to inform sensor placement and policy decision making.

Consider in this case that we want to detect the presence of pathogenic microorganisms in a water distribution system. Pathogens are generally small and easily transported in water. Sources of pathogens to a water distribution system include: (1) source water contamination followed by improper or insufficient treatment, (2) regrowth of organisms due to insufficient disinfectant residual in the distribution system, (3) contamination due to transient pressure drops leading to infiltration of groundwater into water pipes, (4) contamination due to incorrect cross-connections with sewer lines, and (5) intentional addition of pathogenic organisms at the treatment plant or in the distribution system.

Several treatment methods are used to remove pathogens from source water. Drinking water treatment involves coagulation, settling and filtration as well as chemical disinfection. Drinking water regulations require finished water to be free of pathogenic organisms to the best of our detection methods. When detection methods are insufficient, standards require drinking water systems to meet critical treatment goals that are associated with known removal levels.

The major removal mechanism for pathogens in the distribution system is the maintenance of a disinfectant residual from the treatment plant to the consumer's tap. Many pathogenic organisms are very sensitive to chlorine, which is routinely added to treated water before release into the distribution system. This chlorine residual in the water is designed to inhibit regrowth of organisms in the pipes and destroy any organisms that enter the distribution system accidentally. Whether the disinfection residual would also destroy organisms released to the system intentionally would depend on the type and concentration of organisms and where in the system they were introduced. Additional factors that can affect the impact of introduction of pathogens into the distribution system include (1) water temperature, pH, turbidity and oxygen concentration, (2) water demand in the system, and (3) distribution system configuration. Many of these conditions will vary in time and space in the distribution system, and pathogens are likely to be affected in a synergistic manner. The relationships between varying distribution system conditions, varying loading of pathogens to the system, and the survival and persistence of specific organisms -- indicator organisms or high-risk human pathogens -- is clearly an area where extensive monitoring coupled with modeling can be used to establish effective early detection systems.

4. RESEARCH APPROACH

We expect that data-mining techniques coupled with expert knowledge of water-quality parameters will result in knowledge discovery from the wealth of water-quality data currently available. Coupling prior domain knowledge of mechanistic energy and mass-balance relationships with the discovered knowledge of observed relationships between water- and sediment-quality parameters, stream-ecosystem parameters, time and flow characteristics, and anthropogenic effects on watersheds helps develop a holistic watershed model, fully informed by all available data. This watershed model provides information on source water for drinking water treatment and is coupled with models for water distribution systems, also informed by knowledge discovered through data mining. The models developed through knowledge discovery in databases allow us to postulate suitable locations and monitoring timetables for arrays of environmental sensors placed to most effectively monitor watersheds and water distribution systems.

4.1 Scenario of use

Our work in the water-quality databases begins with single-source, single-attribute retrievals of the form "get all the values of pathogens reported in source water in a single watershed from a single data source (i.e., EPA's STORET) for the range of positions $x+Dx$ and times $t+Dt$." We then move on to multi-attribute or multi-source retrievals of the form "get all the values of pathogens and turbidity and flow in all watersheds that experienced rainfall greater than 2 inches at locations $x+Dx$ and times $t+Dt$" or "get all the values of organism growth in the distribution system and levels of dissolved organic carbon (DOC) in the source and finished water for positions $x+Dx$ and time $t+Dt$." These queries are likely to require multiple databases (e.g., the EPA Storet database of chemical and biological parameters and the water treatment system monitoring data for microorganisms). Following development of the necessary spatio/temporal multi-source retrieval system, we search for patterns in the source water data of the form "if rainfall exceeds 0.5 inches in the watershed, within four hours pathogen levels at locations x, $x_1...x_i$ will exceed regulatory allowances by k, $k_1...k_i$% for y, $y_1...y_i$ hours, thus necessitating increased treatment diligence at the plant." We also search for patterns in the distribution water quality of the form "if DOC in the source water exceeds 2 mg/L, regrowth of nonpathogenic organisms within the distribution system at locations x, $x_1...x_i$ impairs detection of pathogen incursions by k, $k_1...k_i$% for y, $y_1...y_i$ hours in our monitoring system." Rainfall and pathogen levels have a well-known (but not quantified) relationship. Likewise, DOC and regrowth of organisms has a well-known relationship, although the affect of this on our ability to design a sensor network for pathogen intrusion detection is unclear. We also mine for novel relationships within the source watershed and the distribution system. Are natural pathogens in the watershed related to high nutrient levels (because fertilizer runoff from farms and pathogen runoff from feedlots are co-incident)? If they are correlated, does the correlation persist in time and space? Are organisms more likely to regrow in the distribution system because of a certain type of DOC in the water? Understanding these relationships assists with locating sensors and identifying pathogen inputs into the source water *and* with locating sensors within the distribution system to monitor routine organism regrowth and to distinguish this from intentional pathogen loading to the system.

Following pattern discovery, we use the newly developed patterns to modify the currently used "quasi-mechanistic" models with statistical models that relate key environmental conditions with pathogen die-off, transport, and pathogen loading in the watershed and the distribution system. We use the new informed model to ask "what-if" scenarios, generating large "pseudo-datasets" to evaluate differences in natural and intentional pathogen loading to watersheds, possible future long-term changes in natural pathogens (e.g., due to changing land use patterns), possible short-term changes related to intentional introduction of pathogens to the watershed or the distribution system, , and possible sensor network locations and query schedules. We use these pseudo-datasets to hypothesize about the effects of unpredictable natural and anthropogenic events (e.g., a flood, an increase in wildlife concentrations or illness rates, the intentional release of pathogens) and possible future scenarios (e.g., higher populations, improved water treatment methods, alternative distribution systems). For example, given a distribution system with two critical reservoirs, would covering one of the reservoirs or adding in-system post-treatment chlorine boosters lead to a need for more or fewer sensors to evaluate water safety. Analyzing these scenarios will inform decisions regarding monitoring locations for a network of sensors to provide "early warning" to a natural or intentional pathogen event.

4.2 Steps towards achieving the goal

In this section we describe the various research components necessary to achieve the project's goal.

4.2.1 Assembling Relevant Water-quality Data

Research conducted by our team has explored the use of Bayesian methods and Bayesian Belief Networks (BBN) to evaluate whether chemical evidence at monitoring wells is suggestive of a landfill leak[7] or of natural biochemical reactions of contaminants in ground water[8]. Similar methods can be applied to surface waters. These results suggest application of KDD to water-quality databases can provide predictive insight even using limited, though well-chosen, characteristics. We expect use of a more complete database of water-quality parameters and a more extensive understanding of issues in water quality to yield improved understanding of the complex relationships inherent in water quality and watershed ecosystems.

To design an appropriate database schema for the data involved in this research, we examine the data available in various databases around the country related to source and finished water quality. EPA, for example, maintains water-quality information for the nation's waters in two major database management systems (DBMSs): LDC (data before 1998) and STORET (data after 1998). STORET is managed by an Oracle DBMS, and stores data in three tables. Environmental locations are indexed by stream reach number, which can be correlated with other data stored in GIS format. Based on these observations, our schema incorporates (a) historical and current data from the above databases and (b) potential data from the in situ biological and chemical sensors that are likely to be distributed in the environment. The developed schema:

- groups information in an optimal way to obtain fast answers to data-mining queries,
- supports translation to and from eXtensible Markup Language (XML), in order to be able to incorporate XML data from the Internet and for compatibility and comparability of the results,
- is reinforced with active rules that, following the model explained in [9], supports scientific workflow design for experimental studies.

4.2.2 Data Mining and Knowledge Discovery

Discovering knowledge in environmental data involves temporal data mining, pattern discovery, and spatial data mining. Using feature extraction, we can map each time sequence into a low dimensionality vector, by, e.g., keeping the first few Fourier or wavelet coefficients.[4,10] In addition, MUSCLES [11] involves data mining for co-evolving time sequences, like network-traffic data or currency-exchange data. MUSCLES works well and gives significantly lower reconstruction errors for real network data when there are linear correlations. In that case we can spot correlated time sequences, e.g., a spike in chemical "x" is followed by a spike in chemical "y" after 3 days, and a dip of organism "w" after 5 days. In the water-quality setting, however, several of the governing differential equations might be non-linear. Our approach in that case is twofold: first, if the distribution is "80/20" [12] we expect 80% (or p) of an organism or pollutant to appear in 1-p of the time slots. Second, we use lag-plots: a sub-sequence t(k+1), ..., t(k+m) corresponds to an m-dimensional vector. Nearby vectors will be clustered, with the so-called Spatial Access Methods (SAMs), like the R-trees[13], and thus can be quickly retrieved for nearest neighbor searches.

Spatial data mining operates on a set of n-dimensional points or regions, and reports clusters, patterns and outliers. In this context, we look for rules of the form "whenever we see chemical 'x' in the finished water, we also see chemical 'y' in the source water, and organism 'w' in the distribution system." The task is closely related to temporal data mining, with the extra difficulty that we have two or three spatial dimensions instead of just one. Spatio/temporal data mining is even harder, because of a subtle difficulty: the relative importance of the time dimension versus the spatial ones are unclear. Therefore, we extend the traditional "Association Rules"[14] to handle spatial attributes in the way we need it. To study separability, we extend previous algorithms[15] to handle several sets of points in space (e.g., one set of points for each chemical/organism in high concentration), in order to group chemicals and organisms that appear in nearby positions, and report them to the analyst. Finally, we use the so-called "correlation integral" from physics and fractals to performa clustering: if the log plot of the cumulative distribution function of the pair-wise distances among the points of the set is straight with slope "s", the dataset is self-similar (= fractal), with intrinsic dimensionality "s" and it is pointless to look for clusters. If the plot has plateaus, they could indicate clusters.

Finally, as with nearly every set of experimental data, reconstruction of missing values is often needed in this work. "Linear regularization"[6] gives consistently lower reconstruction error than the uniformity assumption, and is linear on the length of the time sequence. We extend linear regularization to (a) apply to higher dimensions while keeping it fast (because the required matrices grow quickly with the dimensionality), (b) find the solution that satisfies differential (or difference) equations, as opposed to being smooth and (c) reward reconstructions that follow these correlations among sequences.

4.2.3 A Dynamic Water-quality Model

Environmental models are developed for scientific purposes and as tools for applied policy development, implementation, and management. Models provide an organizing and integrating framework for fundamental knowledge on environmental processes and interactions. When properly formulated, tested, and corroborated with observed data,[16,17] models can provide a foundation and focus for decision support in the development of environmental policy.

Watershed and distribution system temporal or spatial detail modeling involves complex transport and reactions conditions. Existing models often make assumptions that are unacceptable within the sensor deployment framework. For example, many models assume complete mixing over the cross-sectional area of a stream in areas where characterization of horizontal or vertical gradients in concentration are important for interpretation of sensor monitoring data. Our model enhancement work focuses on expanding existing models for watershed and water distribution systems to address these complexities as well as to add new mechanistic representations for multiple biological reactions and the effects of chemical and biological mixtures. By coupling our model development with knowledge discovery in existing environmental databases, we can incorporate statistically relevant microbial population dynamics and mixture effects for which mechanistic understanding is still unavailable.

The importance and role of sensitivity and uncertainty analysis is now recognized in virtually all domains of model application.[18-20] We systematically study model sensitivity and uncertainty as part of the parameter estimation process to identify which aspects of model structure can be resolved with available laboratory and field data, and provide guidance on those additional data and studies with the most potential information value (i.e., those most likely to reduce key uncertainties in model predictions). Uncertainty analysis is conducted using advanced Bayesian methods and Monte Carlo procedures that combine prior scientific knowledge with the information in the available datasets.[21] Numerical methods that combine classical parameter-estimation procedures and Bayesian simulation techniques, such as Markov Chain Monte Carlo and efficient sampling techniques that quickly find and span the posterior parameter space, are used and are being advanced for this purpose.[22]

4.2.4 Data Sensing Array Decision Making

Our decision making research focuses on large sets of hypothetical data. First, reconstructed data for past conditions is used to test the models' predictions against the historical data. Second, we generate hypothetical data for a wide range of initial conditions and potential future trajectories. In these hypothetical worlds, we evaluate different sensor-deployment plans. Each environment (watershed or distribution system) has multiple possible development trajectories (scenarios), each with significant uncertainty, requiring the generation of thousands of pseudo-data sets. These datasets are used to evaluate sensor development, deployment, and querying schedules

For example, we determine the spatial and temporal data needs for identifying an intentional introduction of a pathogenic organism into a water-distribution system prior to large-scale human health effects. The location and timing of data collection is affected by the sensitivity of the *in situ* sensors to changes in the concentration of controlling parameters (e.g., flow, disinfectant residual concentration, and organism concentration). Thus, our work informs sensor development by demonstrating the performance trade-offs inherent in sensor sensitivity and selectivity design decisions. By coupling cost/benefit modeling with our sensor-deployment plans,[23] we are able to evaluate when the use of new, but potentially more expensive, *in situ* techniques provides more value than traditional monitoring. Hypothetical datasets also allow evaluation of methods to plan and query sensor networks for environmental monitoring. This methodology is necessary because existing optimization processes for network sensor locations are limited to linear processes[24-26] and the water-quality dynamics are highly nonlinear. [25,26]

5. SUMMARY

Widespread distribution of sensor networks in the natural and built environment will eventually allow for extensive "pulse-taking" of our world. Large datasets produced by these sensor networks will require sophisticated data mining and modeling to enable optimal decision-making. Water quality sensing is an area ideally suited to testing new algorithms for spatial-temporal data mining and new methods to model coupled non-linear processes. Security concerns regarding detection of intentional contamination of drinking water have focused attention on the need for enhanced sensing, data evaluation, and decision making within drinking-water distribution systems. The work described here links development of new knowledge discovery and numerical modeling methods with decision making research designed to evaluate and optimize sensor deployment plans.

6. REFERENCES

[1] Agrawal, R.; Imielinski, T.; Swami, A. *IEEE Transactionon Knowledge and Data Engineering;* 1993; p 6.

[2] Han, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Morgan Kaufmann: 2000.

[3] Agrawal, R.; Srikant, R. Mining Sequential Patterns. *ICDE* **1995**, 3-14.

[4] Faloutsos, C.; Ranganathan, M.; Manolopoulos, Y. Fast Subsequence Matching in Time-Series Databases. *SIGMOD Conference* **1994**, 419-429.

[5] Rafiei, D.; Mendelzon, A. O. Similarity Based queries for time series data. *SIGMOD Conference* **1997**.

[6] Faloutsos, C.; Jagadish, H. V.; Sidiropoulos, N. D. Recovering Information from summary data. *Journal of Very Large Databases* **1997**, *August*.

[7] Small, M. J. Groundwater detecion monitoring using combined information from multiple constituents. *Water Resources Research* **1997**, *33*(5), 957-969.

[8] Stiber, N.; Pantazidou, M.; Small, M. J. Expert Systems Methodology for evaluating reductive dechlorination at TCE sites. *Environmental Science and Technology* **1999**, *37*(17), 3012-3020.

[9] Ailamaki, A.; Ioannidis, Y.; Livny, M. *Proceedings of the 10th International Conference on Scientific and Statistical Database Management*.

[10] Yi, B.-K.; Faloutsos, C. Fast Time Sequence Indexing for Arbitrary Lp Norms. *VLDB* **2000**, 385-394.

[11] Yi, B.-K.; Sidiroupoulos, N.; Johnson, T.; Jagadish, H. V.; Faloutsos, C.; Biliris, A. Online Data Mining for Co-Evolving Time Sequences. *ICDE* **2000**, 13-22.

[12] Faloutsos, C.; Matias, Y.; Silberschatz, A. *VLDB;* 1996.

[13] Guttman, A. A. R-Trees: A synamic index structure for spatial searching. *Proceedings ACM SIGMOD* **1984**, 47-57.

[14] Agrawal, R.; Imielinski, T.; Swami, A. *Proceedings of ACM SIGMOD;* 1993.

[15] Faloutsos, C.; Seeger, B.; Taina, A. J. M.; Traina, C. *Journal of the SIGMOD Conference;* 2000.

[16] U.S. Environmental Protection Agency (EPA) EPA-SAB-EEC-89-012; Resolution on the Use of Mathematical Models by EPA for Regulatory Assessment and Decision-Making Science Advisory Board (SAB), E. E. C., Washington DC, 1989.

[17] Small, M. J. Show me the Data. *Journal of Industrial Ecology* **1997**, *1*(4), 9-12.

[18] Morgan, M. G.; Henrion, M. *Uncertainty: a Guide to Dealing With Uncertainty in Quantitative Risk and Policy Analysis*; Cambridge University Press: Cambridge UK, 1990.

[19] Skaggs, T. H.; Barry, D. A. Sensitivity methods for time-continous, spatially discrete groundwater contminant transport models. *Water Resources Research* **1996**, *32*(8), 2409-2420.

[20] Turanyi, T. Sensitivity analysis of complex kinetic systems: tools and applications. *Journal of Mathematical Chemistry* **1990**, *5*, 203-248.

[21] Lockwood, J. R.; Schervish, M. J.; Gurian, P. L.; Small, M. J. Characterization of Arsenic Occurrence in US Drinking Water Treatment Facility Source Waters Carnegie Mellon University, Department of Statistics Technical Report No. 700.

[22] Reichert, P.; Schervish, M.; Small, M. J. An efficient sampling technique for Bayesian Inference with Computationally Demanding Models. *Revised and Resubmitted* **2001**.

[23] Bagajewicz, M.; Sanchez, M. Cost-optimal design of reliable sensor networks. *Computers and Chemical Engineering* **2000**, *23*(11-12), 1757-1762.

[24] Sen, S.; Narasimhan, S.; Deb, K. Sensor network design of linear processes using genetic algorithms. *Computers and Chemical Engineering* **1998**, *22*(3).

[25] Ali, Y.; Narasimhan, S. Sensor Network design for maximizing reliability of lienar processes. *AICHE Journal* **1993**, *39*(5), 820-828.

[26] Ali, Y.; Narasimhan, S. Redundant sensor network design for linear processes. *AICHE Journal* **1995**, *41*(10), 2237-2249.