Optimizing Long Term Monitoring at a BP Site Using Multi-Objective Optimization

Barbara Minsker (University of Illinois and RiverGlass Inc.), Peter Groves (RiverGlass Inc.), and Dennis Beckmann (BP)

Abstract

BP (formerly British Petroleum) incurs significant costs associated with monitoring subsurface remediation sites. The purpose of this project is to evaluate whether these costs could be reduced by identifying and eliminating both spatial and temporal redundancies in the monitoring data at a BP site without significantly increasing monitoring errors. The project also aims to demonstrate the potential for multi-objective optimization approaches to improve monitoring decision making at the many sites at BP and elsewhere with long-term monitoring records.

The first step in the optimization process is to identify monitoring objectives and constraints, and express them in mathematical form. In this case, the initial objectives were to minimize the number of samples collected and to minimize relative BTEX interpolation error. The BTEX interpolation error for trial sets of sampling plans are calculated by comparing the concentrations interpolated using all sampling locations and times with those interpolated using only reduced sampling frequencies or locations. Historical data from the wells that are currently being sampled are used to develop a suite of interpolation models, which are then tested using a cross-validation approach. Adaptive Environmental Monitoring System (AEMS) software, developed at the University of Illinois and RiverGlass Inc., is then used to search through the billions of sampling plans to identify the optimal tradeoffs between the number of samples collected and the relative error.

Introduction

Routine groundwater monitoring can account for a significant portion of the lifecycle spending for many remediation projects. Some estimates put long-term monitoring costs at 30% of an overall environmental restoration project budget [EWRI 2003]. Numerous monitoring optimization approaches exist (see EWRI 2003 for an overview), but most recent work has focused on reducing spatial and temporal redundancy in monitoring plans. Spatial redundancy focuses on eliminating wells whose sampling data can be estimated from surrounding wells without introducing significant errors. Temporal redundancy focuses on reducing sampling frequencies to eliminate redundant data at the same well over time. This paper proposes a new approach to optimizing for both spatial and temporal redundancy simultaneously using evolutionary multi-objective optimization. The approach is demonstrated at a BP site in New Jersey.

Site Description

The BP site is a 100-acre terminal built in the late 1920s with both dense and light nonaqueous phase liquid (DNAPL and LNAPL) from chlorinated solvents and BTEX. The geology of the site is complex, with a multi-layered sand and gravel aquifer and a

fluctuating water table. This project focuses on the upper-layer aquifer only, specifically the shallow wells that monitor the BTEX plume. This plume is considered stable and is relatively well understood, whereas the DNAPL plume, which extends 1 mile off site, is still being characterized. The two constituents of concern (COC's) of most interest were benzene and chlorobenzene. Of the 110 wells that were sampled for these constituents in the upper aquifer, a number of wells contained free product and were not sampled during at least one of the 7 semi-annual sampling periods considered in this analysis. Also, 22 wells were declared sentinel wells and were not eligible to be removed from the sampling regimen. These are wells on the perimeter of the site that must be sampled regularly to insure that the plume does not expand beyond its known boundaries.

Methodology

To perform the spatio-temporal monitoring optimization, we used a software package called Adaptive Environmental Monitoring System (AEMS), developed at the University of Illinois and RiverGlass Inc. AEMS is being built as a complete online data analysis and optimization package using Data to Knowledge (D2K) [Welge et al. 2003] technology. AEMS contains a set of data-analytics model building components, automated data tracking components, and optimization components that can be combined as needed for a particular monitoring application.

Using AEMS for spatio-temporal monitoring optimization, the first step is to create a suitable interpolation model using data from all wells and sampling periods to be included in the analysis. An interpolation model estimates contaminant concentrations at all spatial locations of interest at a site and in all monitoring periods of interest in a spatio-temporal approach, using a number of measured data points. This approach is similar to the spatial redundancy approach that has been used in many previous studies, but is extended here to both space and time.

In this study, we propose a local linear regression model to perform the space-time interpolation. In this model, a decision tree is used that recursively splits the data into a tree of subsets [Matheus 1990]. The subsets are formed based on features (in this case, space and time coordinates) that are most predictive of the outputs (in this case, concentrations). The decision tree thus splits the data into local clusters at each terminal node of the tree. At each node, a stepwise linear regression model is then fit to the local data. Stepwise linear regression involves iteratively adding or removing features that improve predictive capability of the model until the best linear model is found.

This novel approach was compared with instance-based weighting [Frink 1994], which is a standard procedure for interpolating data. In this case, concentrations at a particular point in space and time are estimated using a multidimensional weighted mean based on concentrations at the nearest neighbors. The weighting is scaled by the distance to each neighbor in both space and time. The number of neighbors included in the estimation and the weighting factor are selected by the learning machine to minimize cross-validation error.

Both models were tested with and without a quantile transformation of the data prior to model fitting. In a quantile transformation, the largest data point is assigned a value of 1, the smallest point a value of 0, and the remaining points are scaled between 0 and 1 (e.g., the 50^{th} percentile data point would have a value of 0.5). Figures 1 and 2 shows the cumulative cross-validation error for each of the 770 benzene and chlorobenzene data points, sorted by increasing error. A number of data points have quite high cross-validation errors because they cannot be estimated from nearby data (indicating no redundancy). The results show that the local linear regression has lower cross-validation errors for both COC's, particularly with the quantile transformation, so this model was chosen for the redundancy analysis.

Once the interpolation model is created, a redundancy analysis is undertaken in which one or more data points are removed from the data set, concentrations are re-interpolated, and the interpolated concentrations are compared before and after the data were removed. If the interpolated concentrations change insignificantly, then the data were redundant and further sampling at those well locations (or at those times) may not be necessary.

Mathematical optimization is used to efficiently sort through the many combinations of sampling locations, times, and constituents that could be considered and identify the optimal set of wells to meet the user's objectives and constraints. The optimization model, in this case a multi-objective genetic algorithm, becomes a wrapper around the interpolation model. The optimization model uses the interpolation model to test candidate monitoring designs to determine how much error each design would create relative to sampling all of the well locations and/or monitoring periods. Using this information, the optimization will identify the best set of wells to meet the user's objectives and constraints. With multiple objectives (e.g., minimizing cost and contaminant estimation error), a set of optimal solutions will be identified showing tradeoffs among objectives. The multi-objective genetic algorithm used in this work is based on the NSGA-II (Nondominated Sorted Genetic Algorithm – II), which was proposed by Deb (2002).

Optimization Results

Results of the optimization will be presented at the conference.

Conclusions

This paper proposes a spatio-temporal monitoring redundancy analysis approach using local linear regression interpolation models and multi-objective optimization. The local linear regression models are shown to have lower cross-validation errors than inverse-distance weighting models.

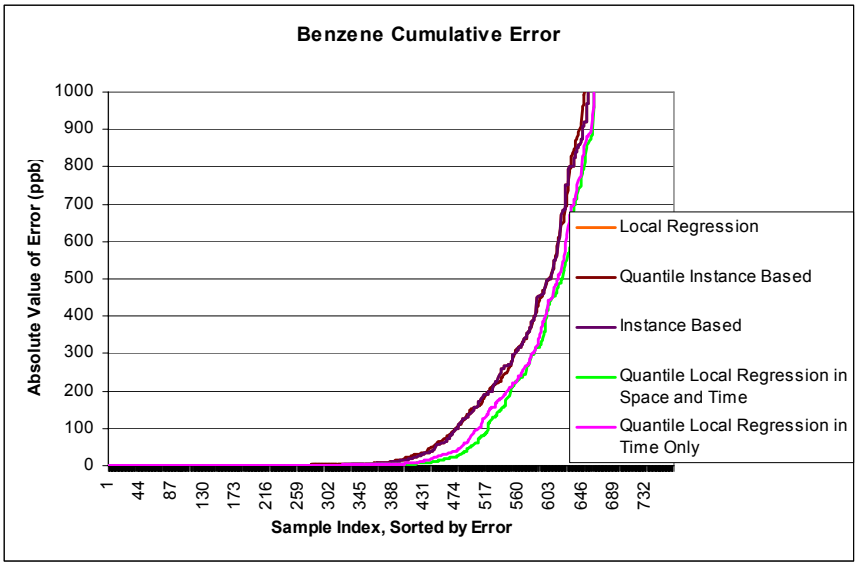

Figure 1. Cumulative benzene cross-validation error for each interpolation model.

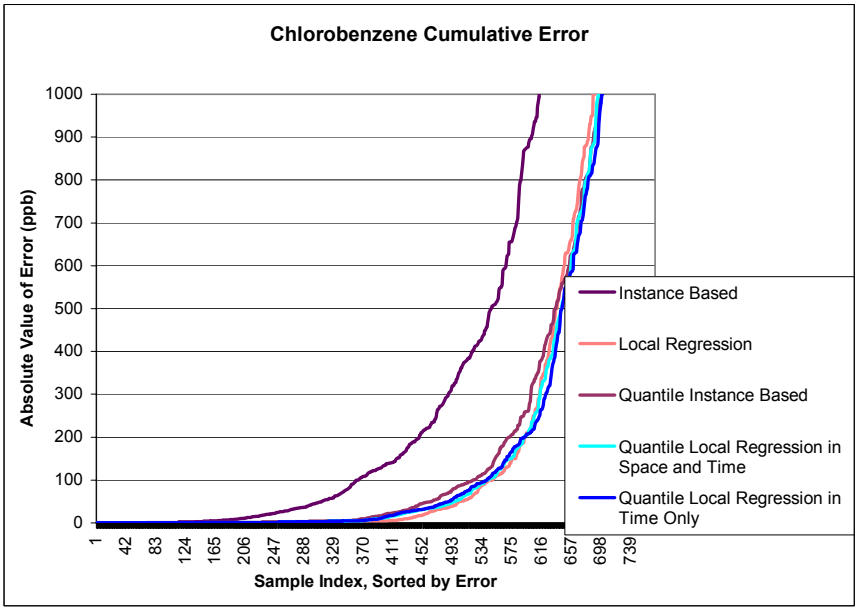

Figure 2. Cumulative chlorobenzene cross-validation error for each interpolation model.

References

Deb, K., et al., A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II, *IEEE Trans. Evol. Computation*, 6(2), 182-197, 2002.

EWRI Task Committee on the State of the Art in Long-Term Groundwater Monitoring Design, *Long-Term Groundwater Monitoring: The State of the Art*, ed. by B.S. Minsker, American Society of Civil Engineers, Reston, VA, 2003.

Frink, N.T., Recent progress toward a three-dimensional unstructured navier-stokes flow solver, AIAA Paper 94-0061, 1994.

Matheus, C. J., Feature Construction: Analytical Framework and an Application to Decision Trees, Ph.D. Thesis, University of Illinois, Urbana, 1990.

Welge, M., L. Auvil, A. Shirk, C. Bushell, P. Bajcsy, D. Cai, T. Redman, D. Clutter, R. Aydt, and D. Tcheng, *Data to Knowledge (D2K)* An Automated Learning Group Report, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, 2003. Available at http://alg.ncsa.uiuc.edu.