XXXVIII IAH Congress

Groundwater Quality Sustainability

Krakow, 12–17 September 2010

Extended Abstracts

Editors:
Andrzej Zuber
Jarosław Kania
Ewa Kmiecik

University
of Silesia
Press 2010

abstract id: **258**

topic: **1**
Groundwater quality sustainability

1.6
Groundwater monitoring

title:

Optimization of groundwater quality monitoring network using information theory and simulated annealing algorithm

author(s): **Wiktor Treichel**
Environmental Engineering Faculty, Warsaw University of Technology, Poland,
wiktor.treichel@is.pw.edu.pl

Małgorzata Kucharek
Environmental Engineering Faculty, Warsaw University of Technology, Poland,
malgorzata.kucharek@is.pw.edu.pl

keywords: monitoring network, information theory, optimization, entropy, simulated annealing

INTRODUCTION

Assessment and optimization of groundwater quality monitoring networks is an important and difficult task and should be carried out in terms of different criteria. While the problem of assessing the cost of network operation do not create any problems from the methodological point of view, a choice of quantitative criterion for assessing the network quality is not as clear. The main goal of the monitoring system is to produce data for statistical analysis. Thus, one of the evaluation criteria should be the amount of information that the monitoring network is able to provide to the control system. The network should be evaluated by the test that measures whether the amount of information obtained from monitoring meets the expectations. If we assume that the monitoring network is a signal communication system capable of providing environmental information, we can use the entropy-based criteria, derived from the Shannon information theory (Shannon, Weaver, 1949). The fundamental criteria derived from this theory are: (1) the value of marginal information entropy, which is a measure of the amount of information containing in the data in a location of sampling point, and (2) the value of transinformation (mutual information) which measures the amount of information shared between each of two sampling points. Marginal information entropy uses probability distribution functions to measure the randomness (or uncertainty) of a random variable. Transinformation can be interpreted as an index of the stochastic dependence between the random variables corresponding to groundwater quality data recorded in different sampling points of monitoring network and shows the reduction of uncertainty included in one variable due to the knowledge of the other variable.

Some methods relating to Shannon information theory were developed to assess monitoring networks. Harmancioglu and Alpaslan (Harmancioglu, Alpaslan, 1992) have shown application of the information theory into water quality monitoring network design in the context of multiobjective optimization. The results were highly promising as the benefits of a monitoring network were defined quantitatively in the terms of information gain measured by entropy. Mogheir and Singh (Mogheir, Singh, 2002) used the entropy-based criteria to quantify the information produced by ground water monitoring network and combined it in the cost-effectiveness analysis. Recently Masoumi and Kerachian (Mogheir, Singh, 2002) used the discrete entropy theory, C-means clustering method and fuzzy set theory to optimal redesign of groundwater quality monitoring network of the Tehran aquifer. The measure of transinformation was used to find the optimal distance between the monitoring wells.

This paper presents a methodology of assessing and optimizing groundwater quality monitoring networks which takes into account the value of transinformation. This criterion allows to assess the redundant information in the network containing in the series of the same water quality parameter observed at different control points. Since the formulated problem of the monitoring network optimization is a complex combinatorial problem, which is hardly solvable by means of classical algorithms, a heuristic algorithm of simulated annealing is proposed, which allows one to find a satisfactory sub-optimal solution. The proposed methodology was applied to optimize the groundwater monitoring network of contaminant reservoir "Żelazny Most" located in the West-South part of Poland which receives post-flotation contaminants originating from copper ore treatment (Duda, Witczak, 2003; Kucharek, Treichel, 2007). This reservoir has been classified as one of the worlds biggest industrial waste disposal site.

INFORMATION ENTROPY MEASURES

The base term of the information theory introduced by Shannon (Shannon, Weaver, 1949) is entropy *H(X)*. It allows to describe quantity of information coming from random variable. If *X* is a discrete random variable with the probability distribution *p(x$_i$)*, *i* = 1, 2, …, *N* then marginal entropy *H(X),* that measures information quantity which comes from observation of *X*, can be calculated as follow:

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log p(x_i) \tag{1}$$

If the probabilities *p(x$_i$)* are low, the entropy value is high. The maximum value of entropy equal to *log(N)* is reached for uniform probability distribution *p(x$_i$) = 1/N* for *i* = 1, 2, …, *N*.

There are three additional types of entropy measures associated with stochastic dependency between two random variables *X* and *Y* (Harmancioglu, Alpaslan, 1992; Kucharek, Treichel, 2007; Mogheir, Singh, 2002): joint entropy, conditional entropy and mutual entropy called transinformation. The joint entropy *H(X, Y)* measures a total information content in both *X* and *Y* , and is a function of the joint probability distribution *p(x$_i$, y$_j$)*. The total entropy of two independent random variables is equal to the sum of their marginal entropies. When *X* and *Y* are stochastically dependent, their joint entropy is less than the total entropy of these variables. Conditional entropy *H(X | Y)* is a measure of the information content of *X* which is not contained in the random variable *Y*. It represents the uncertainty remaining in *X* when *Y* is known. The transinformation *T(X, Y)* is another entropy measures which measures the redundant or mutual information between *X* and *Y*. It is defined as the information content of *X* which is contained in *Y*. It can also be interpreted as the reduction of uncertainty in *X*, due to knowledge of variable *Y*.

$$T(X,Y) = T(Y,X) = \sum_{i=1}^{N} \sum_{j=1}^{N} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i) p(y_j)} \tag{2}$$

where *X* and *Y* are two discrete random variables defined in the same probability space with probability *p(x$_i$)* and *p(y$_j$)*, respectively.

The approach developed here consists in assessing the reduction in the joint entropy of two or more variables due to the presence of stochastic dependence between them. This reduction corresponds to the redundant information in the series of the same water quality parameter observed at different control points. Thus a criterion of evaluation of the groundwater quality monitoring network will be the value of transinformation. Minimization of this objective function could be achieved by an appropriate choice of the number and location of sampling points.

GENERAL CHARACTERISTICS OF THE DATA SET

The reservoir "Żelazny Most" was establish in 1974 as a landfill of copper ores flotation tailings and nowadays is one of the world's largest industrial waste disposals. It occupies an area of approximately 1400 hectares. The landfill is surrounded by a protective zone ranging from about 500 to about 1500 meters from the dams. Groundwater quality monitoring network includes 278 data points in which the water quality of first groundwater level is observed. Depending on the plan at some points the measurement is conducted three times a year and at others once every four years.

Following an analysis of the data set, three of variables were chosen for further study: Cu [mg/dm^3], Na [mg/dm^3] and Cl [mg/dm^3]. Data include 55 measurement points, where tests were performed for 10 years, from 1996 to 2005. In the first stage of the analysis the results of measurements are used to calculate basic descriptive statistics (see Table 1 for some examples). Based on analysis, it was found that as time increases the average concentrations of pollutants in groundwater increases and higher maximum values are met. Data distributions are asymmetrical, as evidenced by the coefficient of skewness and significant difference between the mean and median. In addition, each year a number of outliers is detected. The calculated basic statistical parameters are confirmation of the general upward trend due to the continuing expansion of the landfill.

Table 1. The descriptive statistics for the chloride contamination [mg/dm^3] in ground-water of the first water level around the disposal site "Żelazny Most" in 2002–2005.

Statistics	Cl_2002	Cl_2002a	Cl_2003	Cl_2003a	Cl_2003b	Cl_2004	Cl_2004a	Cl_2005
Mean	2 785.9	2 732.8	3 138.6	3 060.6	3 292.6	3 793.3	3 630.0	3 673.3
Standard error	403.8	396.4	417.0	421.0	416.8	447.9	427.2	503.0
Quartile1 (25%)	71.3	60.15	68.8	72.6	72.85	80.4	139.95	144.85
Median	1 865.0	1 556.0	2 584.0	1 971.0	2 669.0	3 967.0	3 921.0	2 746.0
Quartile 3 (75%)	4991.5	4962	5266.5	5378.5	5467	6116	6048	6145.5
Standard deviation	2 994.8	2 940.1	3 092.3	3 122.1	3 091.0	3 321.4	3 168.1	3 730.4
Kurtosis	0.004	0.124	−0.397	−0.362	−0.603	−1.088	−0.793	0.205
Skewness	0.945	0.970	0.730	0.793	0.614	0.331	0.445	0.879
Range	10 206.6	10 201.1	10 328.3	10 329.1	10 143.4	10 992.3	10 505.0	15 216.3
Minimum	10.4	7.4	17.7	16.9	15.6	6.7	10.0	7.7
Maximum	10 217.0	10 208.5	10 346.0	10 346.0	10 159.0	10 999.0	10 515.0	15 224.0
Number of items	55	55	55	55	55	55	55	55
Confidence level (95.0%)	809.6	794.8	836.0	844.0	835.6	897.9	856.5	1 008.5

OBJECTIVE FUNCTION

In the next step of data analysis the values of transinformation were calculated. For all control points in the monitoring network and for three variables: Cl, Cu and Na transinformation was calculated using equation (2). The computing of marginal and joint probability distributions for each sampling points and for each variables was carried out by the mean of contingency tables (Mogheir et al., 2003). To take into consideration in the optimization problem all the investigated variables (Cl^- , Cu^{2+}, Na^+) the objective function was defined as the average of transinformation values determined for the pairs of sampling control points and for the subsequent concentration of all three ions Cl^- , Cu^{2+}, Na^+ :

$$J = \frac{1}{3M(M-1)} \sum_{s=1}^{3} \sum_{n \neq m} T_s(X_n, X_m) \qquad (3)$$

where *s* is an index of the investigated variable (Cl^- , Cu^{2+}, Na^+), *n* and *m* are indices of sampling points, *M* is a number of sampling points.

Because the transinformation defines the amount of information contained in one variable (sampling point), which is also contained in another, the use of this criterion allows us to remove redundant information from the groundwater quality control system.

SIMULATED ANNEALING OPTIMIZATION ALGORITHM

Since the formulated problem of minimization of objective function (3) belongs to a class of complex combinatorial problems, which are hardly solvable by means of classical algorithms, a heuristic algorithm of simulated annealing (Kirkpatrick et al., 1983) was proposed, which allows one to find a satisfactory sub-optimal solution. This is a technique that attracted significant attention as suitable for optimization problems of large scale. At the heart of this method is an analogy with thermodynamics, specifically with the way that metals cool and anneal. The key parameter of the algorithm is the cooling schedule. Sufficiently high initial temperature in the initial phase allows the search through the entire search space. However, the speed and the way of lowering the temperature determines the speed of the algorithm. Too slow decrease in temperature can hamper the identification of the optimum by leaving too much freedom to search during a large number of iterations. On the other hand, if the temperature drops too quickly, the algorithm can easily stay at a local optimum, it will not have enough iterations to effectively search through the entire search space.

RESULTS OF OPTIMIZATION

In the order to improve the efficiency of the monitoring network around the disposal site "Żelazny Most" a number of optimization was performed. Calculations were carried out in several variations. We performed calculations for the scenario of whole groundwater monitoring network and for the option the network is divided into zones having regard to the ability or inability of the impact on each individual sampling points. Given the complexity of the phenomenon and the possibility of interactions between the various points the best results were obtained when considering all three ions (Cl^- , Cu^{2+}, Na^+), but maintaining the distinction between the eastern and western forefield.

In each variant of optimization the number of piezometers in the network was successively decreased. For each variant of reduction of the number of sampling points the simulated annealing algorithm calculated the optimal value of the objective function, which evaluates the informational value of the monitoring network, and defined the optimal configuration of the remaining network. Figure 1 shows the optimal configuration of the network designated in the optimization process for the variant of reduction of the network by 20%. It is worth to note that the criterion of transinformation for evaluating the amount of redundant information in the network ensures the stability of the solutions obtained in sequential variants. Sampling points removed from the monitoring network in the variant of a reduction by 10% are also removed from the monitoring network in the variant of a reduction of 20% and were still removed in the variant of a reduction of 30%. This means that the monitoring network reduction procedure can be performed sequentially.

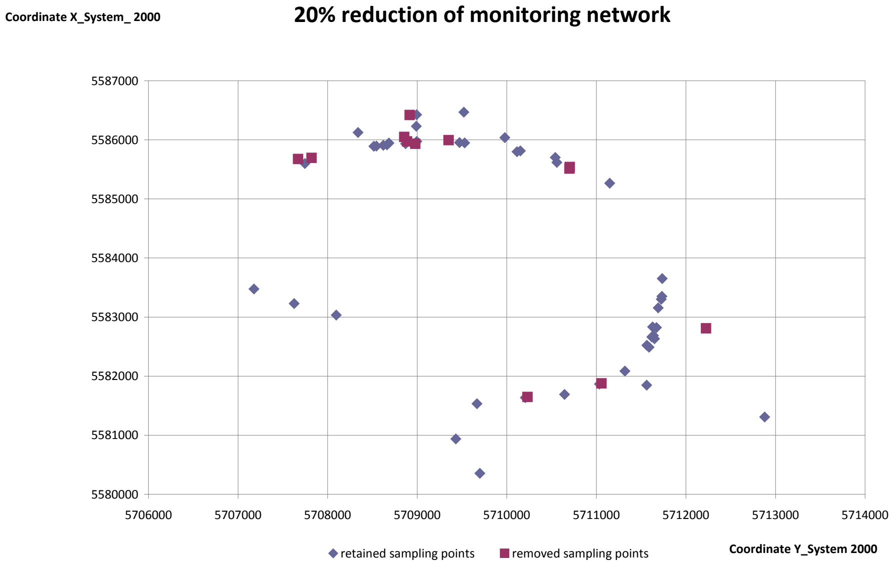

Figure 1. Removed and retained sampling points while reducing the monitoring network by 20%.

SUMMARY AND CONLUSIONS

The aim of this study was to increase the effectiveness of the groundwater quality monitoring network by reducing the number of piezometers while maintaining an acceptable amount of information available in the network. The methodology was applied to the monitoring network of the contaminant reservoir "Żelazny Most" which collects post-flotation contaminants originating from copper ore treatment. When analyzing the information value of the monitoring network, the data on ion concentration of chlorine, sodium and copper in the groundwater of the first water level are used. Different combinations of a number and locations of sampling points were evaluated using the measure of redundant information called transinformation. The simulated annealing algorithm was used to find a sub-optimal solution of the optimization problem. The results show that the proposed methodology can be effectively used for redesign and reorganization of the existing monitoring network and the best combination of sampling points considering minimal redundant information in the system could be selected.

ACKNOWLEDGEMENTS

This research was partially financed by Polish Ministry of Science and Higher Education, project number 1 T09D 010 30. The authors acknowledge KGHM Polska Miedź for providing data for this study.

REFERENCES

Duda R., Witczak S., 2003: *Modeling of the transport of contaminants from the Żelazny Most flotation tailings dam.* Gospodarka Surowcami Mineralnymi, vol. 19, No. 4, pp. 69–88.

Harmancioglu N.B., Alpaslan N., 1992: *Water quality monitoring network design: A problem of multi-objective decision making.* Water Resource Bulletin., vol.28, No.1, pp. 179–192.

Kirkpatrick S., Gelatt C.D., Vecchi M.P., 1983: *Optimization by simulated annealing.* Science, vol. 220, No. 4598, pp. 671–680.

Kucharek M., Treichel W., 2007: *Assessment of groundwater quality monitoring network based on information theory, in Recent Advances in Stochastic Modeling and Data Analysis.* C. H. Skiadas (ed.), World Scientific, pp. 636–644.

Masoumi F., Kerachian R., 2009: *Optimal redesign of groundwater quality monitoring networks: a case study.* Environmental Monitoring and Assessment, Springer, doi:10.1007/s10661-008-0742-3.

Mogheir Y., Singh V.P., 2002: *Application of information theory to groundwater quality monitoring networks.* Water Resources Management, vol. 16, No. 1, pp. 37–49.

Mogheir Y., De Lima J.L., Singh V.P., 2003: *Assessment of spatial structure of groundwater quality variables based on the entropy theory.* Hydrology and Earth System Sciences, vol. 7, No. 5, pp. 707–721.

Shannon C.E., Weaver W., 1949: *The mathematical theory of communication.* The University of Illinois Press, Urbana, Illinois.

International Association of Hydrogeologists

AGH University of Science and Technology

2-vol. set + CD

ISSN 0208-6336

ISBN 978-83-226-1979-0