IAH 2010 Krakow

BIULETYN PAŃSTWOWEGO INSTYTUTU GEOLOGICZNEGO 441: 193–200, 2010 R.

OPTIMIZATION OF GROUNDWATER QUALITY MONITORING NETWORK USING INFORMATION THEORY AND SIMULATED ANNEALING ALGORITHM

WIKTOR TREICHEL[1], MAŁGORZATA KUCHAREK[1]

Abstract. In this paper a methodology of assessment and optimization of groundwater quality monitoring network which takes into account the evaluation criteria derived from the Shannon information theory is presented. 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. 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. In the optimization problem the objective function involving the value of transinformation of the investigated water quality parameters (Cl, Cu, Na) is minimized. To minimize the objective function the simulated annealing algorithm, which allows to find a satisfactory sub-optimal solution, was used. The proposed methodology was applied to optimize the groundwater monitoring network of contaminant reservoir Żelazny Most, one of the worlds biggest industrial waste disposal site, which collects post-flotation contaminants originating from copper ore treatment. The results show an increase in the effectiveness of the monitoring network by reducing the number of sampling points while maintaining an acceptable amount of information available in the network.

Key words: simulated annealing, monitoring network, information theory, optimization, entropy.

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 (1992) have shown application of the information theory into water quality monitoring network design in

[1] Warsaw University of Technology, Faculty of Environmental Engineering, Division of Informatics and Environment Quality Research, Nowowiejska 20, 00-653 Warszawa, Poland; e-mail: wiktor.treichel@is.pw.edu.pl, malgorzata.kucharek@is.pw.edu.pl

the context of multi-objective optimization. They have developed temporal, spatial and combined temporal/spatial design criteria based on entropy. 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 (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 (2009) 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, a criterion derived from the Shannon information theory. Transinformation 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 was 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, 2006). 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 $H(X)$. This term allows to describe quantity of information coming from random variable. Entropy is a quantitative measure of the information content of a series of data since reduction of uncertainty by making observations equals the same amount of gain in information. According to Shannon, information is attained only when there is uncertainty about an event, which implies the presence of alternative results the event may assume. The name "entropy" is used since the mathematical expression for this concept is analogous to that of entropy in statistical mechanics.

If X is a discrete random variable with the probability distribution $p(x_i)$, $i = 1, 2, \ldots, N$ then marginal entropy $H(X)$, that measures information quantity which comes from observation of X, can be calculated as follow:

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

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 $i = 1, 2, \ldots, N$. There are three additional types of entropy measures associated with stochastic dependency between two random variables X and Y (Harmancioglu, Alpaslan, 1992; Mogheir, Singh, 2002; Kucharek, Treichel, 2006): 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 $H(X \mid Y)$ 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.

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

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 transinformation may be expressed as the difference between the total entropy and the joint entropy of the two dependent random variables X and Y:

$$\begin{aligned} T(X,Y) &= H(X) + H(Y) - H(X,Y) = \\ &= H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \end{aligned} \qquad [3]$$

For discrete random variables the transinformation can be calculated using contingency tables (Mogheir *et al.*, 2003).

The approach developed here consists in assessing the reduction in the joint entropy of two or more variables (i.e. two or more sampling points where a particular water quality variable is observed) 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.

Therefore the basis for the assessment of whether the existing monitoring network provides sufficient non-redundant

information to the system may be the average value of mutual information (transinformation) calculated for the different scenarios that define the location of measurement points in the monitoring network. On the basis of this criterion it is possible to propose a revision of the existing monitoring network in the most efficient way.

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 1,400 hectares. The landfill is surrounded by a protective zone ranging from about 500 to about 1,500 m 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. In the piezometers the concentrations of chemical variables in the form of ions are measured: calcium (Ca), cadmium (Cd), chlorine (Cl), chromium (Cr), copper (Cu), iron (Fe), potassium (K), magnesium (Mg), manganese (Mn), sodium (Na), nickel (Ni), ammonium (NH_4), nitrite nitrogen (NO_2), nitrate nitrogen (NO_3), lead (Pb), sulphates (SO_4), and additionally color, pH, conductivity, hardness, alkalinity and total dissolved substances.

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. The number of elements for each measurement is different and is as follows:

– for the chlorine ions (Cl^-) is a sequence of 17 measurements taken at a frequency of once a year in 1996–1999, and similarly in 2005, two times a year in the years 2001, 2002 and 2004 and three times a year in the years 2000 and 2003;

– for the copper (Cu^{2+}) and sodium (Na^+) ions, data set is a sequence of 13 measurements made with the frequency once a year in 1996–1999 and 2005 and two times a year in the years from 2000 to 2002 and 2004; in 2003 no measurements were made.

In the first stage of the analysis the results of measurements are used to calculate basic descriptive statistics (mean, minimum, maximum, standard deviation, skewness, kurtosis). Based on analysis, it was found that the average of the chlorine ion concentration increases since 1996 (2065 mg/dm^3) by the year 2005 (3673 mg/dm^3). The minimum value is in the range 4.0–17.7 mg/dm^3. Depending on the year in which the measurement was performed, 25% of the measurements of chlorine ion concentration is lower than the 50.5 mg/dm^3 in 1997 and 144.9 mg/dm^3 in 2005. 75% of the measurement of chlorine ion concentration reaches a value of not more than 3516.5 mg/dm^3 in 1996 while this figure is increasing steadily and in 2005 a turnover of 6145.5 mg/dm^3. Histograms of chloride concentration, the coefficient of skewness and significant difference between the mean and median show the high asymmetrical data distribution. In addition, the standard deviation in the range of 2310.3 to 3730.4, and the difference between the upper and lower quartiles show a large dispersion of data around the averages Table 1.

By examining the concentration of other ions we observed a similar trends as in the case described above concerning the concentration of chlorine ions. As time increases the average concentrations of pollutants in groundwater

Table 1

The descriptive statistics for the chloride contamination [mg/dm^3] in groundwater 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%)	4 991.5	4 962	5 266.5	5 378.5	5 467	6 116	6 048	6 145.5
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

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.

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: chlorine, copper and sodium 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 sodium ions, chlorine ions and copper ions:

$$J = \frac{1}{3M\,(M-1)} \sum_{s=1}^{3} \sum_{n \neq m} T_s\,(X_n, X_m) \qquad [4]$$

where:

s – index of the investigated variable (Cl, Cu, Na),
n and m – indices of sampling points,
M – number of sampling points.

Because 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. Therefore, the application of transinformation criterion allows to select those sampling control points which contain the least information on other points, while itself deliver to the system a significant amount of unique information.

SIMULATED ANNEALING OPTIMIZATION ALGORITHM

The primary objective of the optimization is to improve the efficiency of the network, which may involve the removal of poorly located measuring points (if we are talking about reducing the network), or make additional points (if a problem of network expansion is considered). In addition to the classical optimization methods, such as the simplex method for linear programming and Newton's method for nonlinear programming, a new class of optimization methods based on heuristic search techniques is recently successfully developed. This group of techniques is known as global optimization techniques and is allowing to find the global optimum without using the gradient of the objective function. This group includes the genetic algorithms, simulated annealing algorithms, tabu search algorithm and different evolutionary algorithms. The idea of heuristic algorithms is to evaluate some new solutions of optimization problem from the neighborhood of the current solution, but the procedure for checking all possible solutions is replaced with a process that works according to a specific heuristic. It mainly allows to obtain a result within a reasonable time. But it should be kept in mind that heuristic algorithms do not guarantee to find optimal solutions but provide a good enough solution, close to the optimal one.

Since the formulated problem of minimization of objective function [4] belongs to a class of complex combinatorial problems, which are hardly solvable by means of classical algorithms, a heuristic algorithm of simulated annealing 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.

Simulated annealing algorithm is an extension of the local search algorithm. A novelty in this method is that there is possible that particles (solutions) move to the state of higher energy (higher value of objective function) and it occurs with a certain probability. Movement to a lower energy state (lower value of objective function) is always allowed.

Simulated annealing algorithm starts working on the solution generated from the set of feasible solutions, then each subsequent solution is selected from a set of neighboring solutions, and the objective function is calculated. If the new solution is better than the previous one then the new solution overwrites the previous solution. Otherwise, if the new solution is worse than the previous one, it could be accepted with some probability. The probability to accept a worse solution is determined from the following relationship, called Metropolis rule (Kirkpatrick *et al.*, 1983):

$$P = e^{-\frac{f(x_{cur}) - f(x_{new})}{k \cdot T}} \qquad [5]$$

where:

k – Boltzmann constant ($k = 1$),
T – temperature (described by cooling schedule),
x_{cur} – current solution,
x_{new} – new tested solution,
P – probability of acceptance of the new solution.

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 from worse to better solutions and vice versa. 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: (1) analysis of the entire area around the site and all types of water quality variables, (2) analysis of the entire area around the site and separately each of variables, (3) analysis of the eastern forefield area of the site and all types of water quality variables, (4) analysis of the western forefield of the site and all types of water quality variables.

We also 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

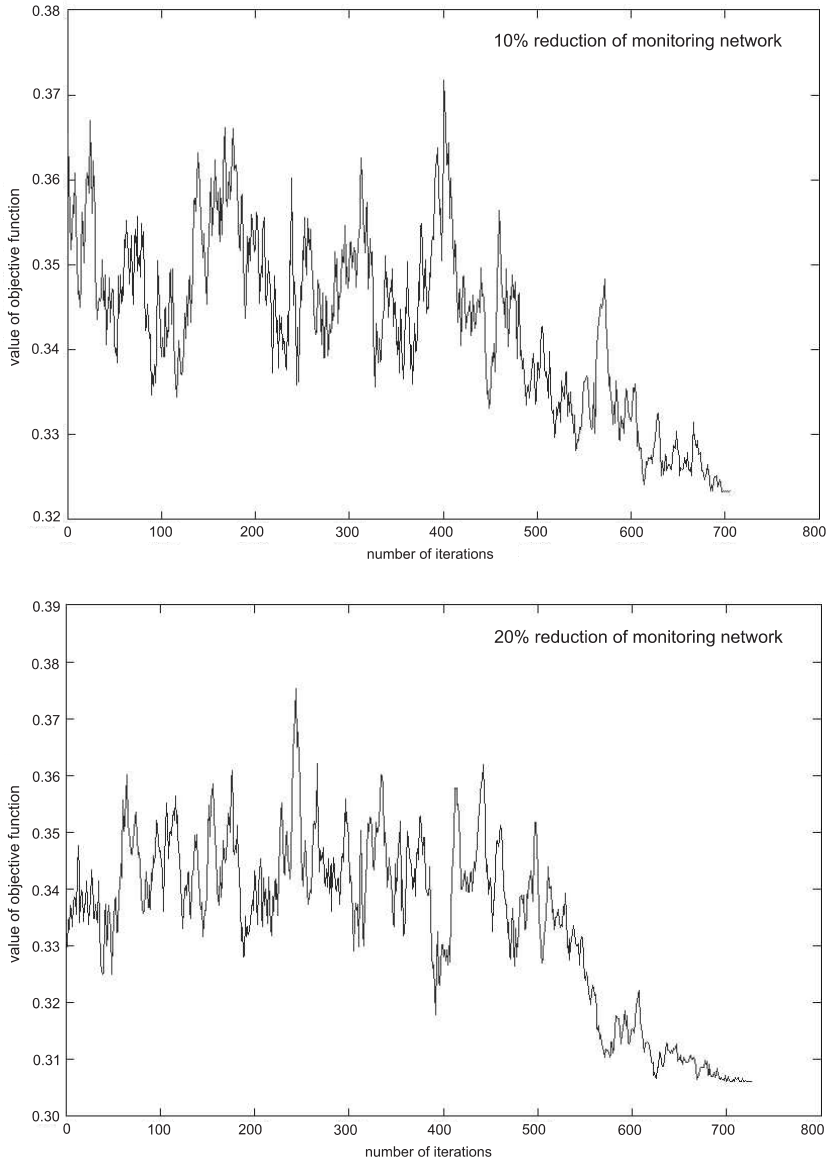

Fig. 1. Plot of the objective function in optimization of monitoring network for the whole network calculated taking into account the concentrations of Na^+, Cl^- and Cu^{2+}

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 plots of the objective function for the variants of reduction of the network by 10 and 20%. On the figure we can observe how does simulated annealing algorithm reach the optimal solution. Minimizing the objective function is not monotonic, but the algorithm copes well with local minima encountered during a global minimum search.

The optimal configuration of the network designated in the optimization process for these two exemplary variants is shown in Figure 2. Please note that the criterion of transinformation for evaluating the amount of redundant information in the network, resulting from the stochastic interaction between the different measuring points, ensures the stability of the sequential solutions. 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%. This means that the monitoring network reduction procedure can be performed sequentially.

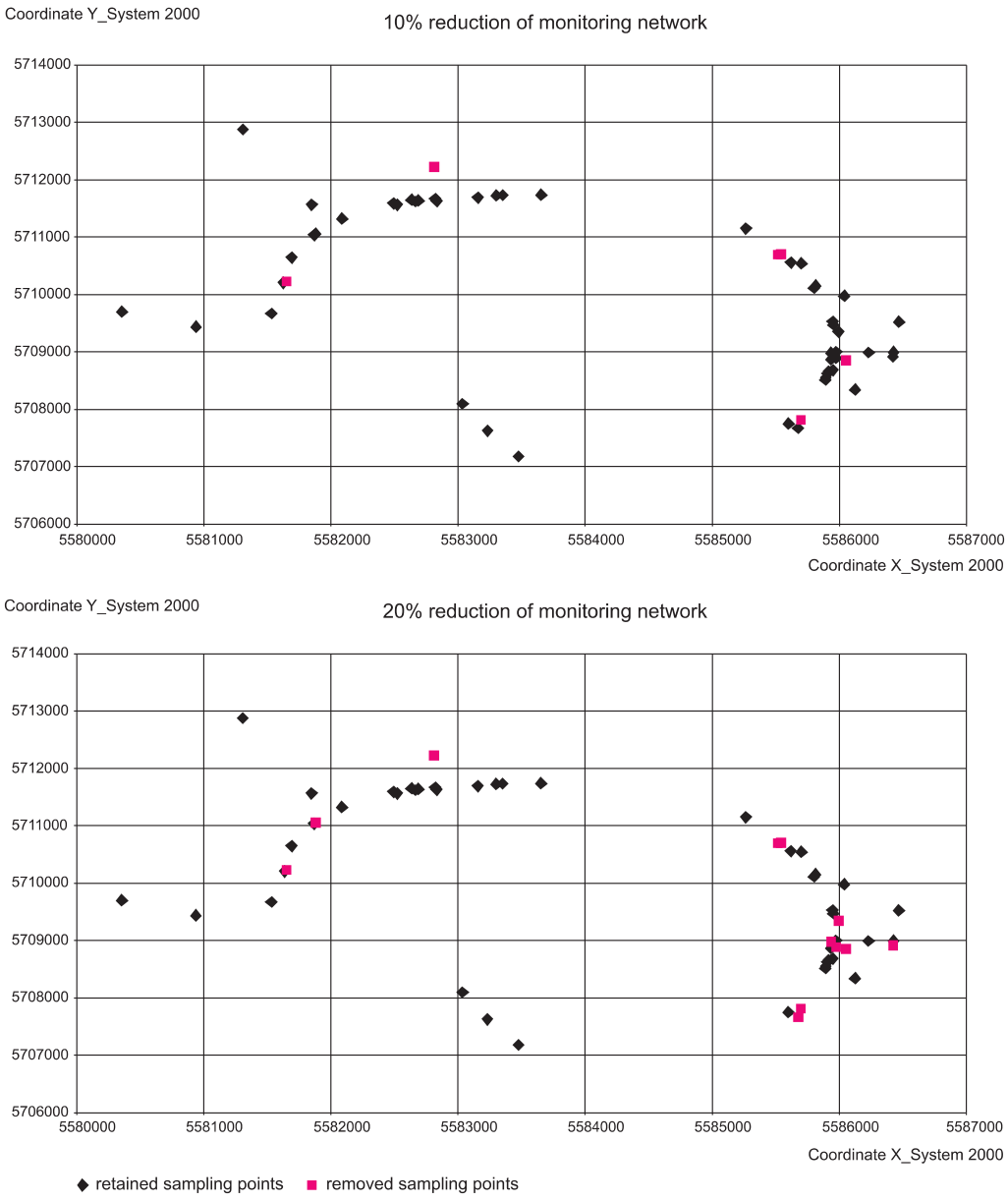

Fig. 2. Removed and retained sampling points while reducing the monitoring network by 10 and 20%

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 are 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. Of course, it must be remembered that the statistical calculations (probability distributions) are always based on historical data. When planning new monitoring points we might use the additional knowledge, for example, geological exploration on any preferred direction of migration of pollutants.

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. *Gosp. Sur. Miner*., **19**, 4: 69–88.

HARMANCIOGLU N.B., ALPASLAN N., 1992 — Water quality monitoring network design: a problem of multi-objective decision making. *Wat. Res. Bull.*, **28**, 1: 179–192.

KIRKPATRICK S., GELATT C.D., VECCHI M.P., 1983 — Optimization by simulated annealing. *Science*, **220**, 4598: 671–680.

KUCHAREK M., TREICHEL W., 2006 — Application of information entropy to assessment of groundwater quality monitoring networks. *Ochr. Środ*., **28**, 3: 45–49 [in Polish].

MASOUMI F., KERACHIAN R., 2009 — Optimal redesign of groundwater quality monitoring networks: a case study. *Env. Monit. and Assess.*, Springer.

MOGHEIR Y., SINGH V.P., 2002 — Application of information theory to groundwater quality monitoring networks. *Wat. Res. Manage.,* **16**, 1: 37–49.

MOGHEIR Y., de LIMA J.L.M.P., SINGH V.P., 2003 — Assessment of spatial structure of groundwater quality variables based on the entropy theory. *Hydrol. Earth Syst. Sc*., **7**, 5: 707–721.

SHANNON C.E., WEAVER W., 1949 — The mathematical theory of communication. The University of Illinois Press, Urbana, Illinois.