Quality of Groundwater, Proceedings of an International Symposium, Noordwijkerhout, The Netherlands, 23—27 March 1981, W. van Duijvenbooden, P. Glasbergen and H. van Lelyveld (Eds.), *Studies in Environmental Science*, Volume 17

OPTIMIZATION OF A GROUNDWATER QUALITY SAMPLING PROGRAM

Y. BACHMAT and M. BEN-ZVI
Hydrological Service, Jerusalem, Israel

ABSTRACT

The paper considers information pertinent to groundwater quality as an input to groundwater management decisions. Accordingly, the design of a groundwater quality sampling program is treated as a problem of maximizing the net benefit from the collected data. It is assumed that the loss because of a suboptimal management decision is partly due to the level of information supplied by the data collection program, which is determined by the frequency, the spatial density and configuration of the observations. An algorithm for deriving the optimal values of the latter under conditions of risk is presented. The algorithm is currently implemented in planning the program of groundwater quality observations in the coastal aquifer of Israel.

INTRODUCTION

Rational management of a groundwater system requires that decisions regarding operation and development of the system be based on information about the present state of the system, its response to control inputs and the utility of the response.

As a rule the information available at the time of decision making is neither complete nor exact. Moreover, the information may be utilized by the decision making algorithms only partly or incorrectly. The resulting effects are losses of two kinds: (a) loss because of suboptimal decisions at the given level of information; (b) opportunity loss because of having made (apparently optimal) decisions at a suboptimal level of information. One of the ways to reduce losses of the second kind is to design an adequate program of data collection. The present paper handles this task with respect to groundwater quality data.

Consider the concentration of a solute, C, as a state variable of groundwater quality such that the utility of groundwater management operations (e.g. distribution of pumpage and recharge, location of waste disposal sites) depends on the resulting distribution of C in both space and time.

The variation of C in the domain of interest is described by the differential

equation (ref. 1):

$$\frac{\partial(nC)}{\partial t} = - \text{div}(\underline{q}C - n\underline{\underline{D}}.\text{grad}C) + \Gamma \qquad (1)$$

where the left-hand side represents the local rate of change of concentration per unit volume of the aquifer, the first term of the right-hand side represents net influx of the solute by convection and hydrodynamic dispersion and the last term represents net rate of supply of the mass of the solute from various sources. It is well known that there is no general method for deriving an analytic solution of equation (1) for C as a function of position and time which could be used to evaluate the consequences of groundwater management decisions. Moreover, some of the necessary input data, such as the initial (i.e. present) distribution of C, hydrogeologic and hydrochemical parameters, can be known only at a finite number of sites at nonsimultaneous time points with randomly fluctuating errors, while some other data pertinent to boundary conditions or source functions may be lacking at all. Under these circumstances, and in order to reduce the error of prediction, it seems more feasible to attempt a prediction of average values of concentration over finite spatial subdomains, referred to as aquifer cells, and over finite intervals of time, referred to as time steps rather than trying to predict local values at individual points. A similar conclusion was reached by recent studies on fixed water quality monitoring stations (ref. 2).

The task of predicting the average concentration of the solute in a cell at the end of a time step then becomes one of estimating the average concentration at the beginning of the time step as well as predicting the average fluxes of the solute through the cell boundaries and the average rate of supply of the solute within the cell during the time step.

The present paper deals with only one of the above items, namely the estimation of the average concentration in an aquifer cell, $\tilde{C}_t$, at a given time point t, on the basis of a set of point samples within the cell and within a given time interval $t \pm \delta t/2$.

STATEMENT OF THE PROBLEM

Let L be a set of values of parameters which characterize the groundwater quality sampling program (e.g. the number and spatial configuration of the sampling points and the frequency of sampling). Also let $\hat{C}_L$ denote the estimate of $\tilde{C}$ at the level of information supplied by L. The deviation of $\hat{C}_L$ from $\tilde{C}$ carries a loss $\mathcal{L}(\tilde{C}-\hat{C}_L)$. It is a random variable whose distribution is determined by the distribution of the observations, by the relationship between the observations and the estimate $\hat{C}_L$ as well as by the level of L. Owing to the randomness of $\mathcal{L}$(and taking into account the

repetitive nature of groundwater management decisions) one can adopt as an objective the minimization of its expected value, referred to as the risk function

$$R(L) = E_{\hat{C}_L} \mathcal{L}(\tilde{C}-\hat{C}_L) \tag{2}$$

which depends on the level of L. However, one should take into account the cost of the sampling program which is also a function of L. Denoting the latter by CS(L), the total loss associated with the level of the sampling program L is

$$F(L) = R(L) + CS(L) \tag{3}$$

Considering L as a set of decision variables the present paper addresses itself to the following problem:

Given a rectangular aquifer cell centered at the point (x_o, y_o) of a cartesian coordinate system and bounded by the coordinate planes $x_o \mp \Delta x/2$, $y_o \mp \Delta y/2$, and given a time interval $t \mp \delta t/2$, find an optimal observation program $L = L_{opt.}$ such that

$$F(L_{opt}) \equiv R(L_{opt}) + CS(L_{opt}) = \min \tag{4}$$

within the spatial domain of the cell and the given time interval.

METHOD OF ATTACK

Estimation of the average concentration, $\tilde{C}$, necessitates the adoption of a statistical model which describes the distribution of the concentration within a cell and within a time interval. This model can then be used to derive the estimate $\hat{C}_L$ from point data of the sampling program and to compute its variance $\mathrm{Var}\,\hat{C}_L$ as a measure of the deviation of $\hat{C}_L$ from the true average, $\tilde{C}$. Having expressed $\mathrm{Var}\,\hat{C}_L$ as a function of L, R(L) which is a function of $\mathrm{Var}\,\hat{C}_L$, can also be expressed as a function of L. Together with the cost function it yields the objective function F(L) which has to be solved by one of the available optimization techniques.

STATISTICAL MODEL OF POINT OBSERVATIONS IN A CELL

The value of an observation, C, at any point within a cell and at any time in the neighbourhood of t can be presented by the linear model

$$C(x,y,\tau) = \alpha_t + \beta_t(x-x_o) + \gamma_t(y-y_o) + \varepsilon(x,y) + \eta(x,y,\tau) \tag{5}$$

$$|x-x_o| \leq \Delta x/2\ ,\quad |y-y_o) \leq \Delta y/2\ ,\quad |\tau-t| \leq \delta t/2$$

where α_t, β_t and γ_t are the paremeters of a plane common to all points of the cell and is fixed during the time interval $(t - \frac{\delta t}{2}, t + \frac{\delta t}{2})$.

ε is a random deviation which depends on the location of the point only whereas η depends on time too. The statistical characteristics of ε and η are common to all points within the cell and within the time range δt, and given by:

$$E\varepsilon = 0\ , \quad \mathrm{Var}\ \varepsilon = \sigma_1^2\ , \quad \mathrm{Cov}[\varepsilon(x,y),\ \varepsilon'(x',y')] = 0$$

$$E\,\eta = 0\ , \quad \mathrm{Var}\ \eta = \sigma_2^2\ , \quad \mathrm{Cov}[\eta(\tau),\ \eta'(\tau')] = \sigma_2^2 \rho^{|\tau-\tau'|}\ , \quad \mathrm{Cov}(\varepsilon,\eta) = 0 \qquad \text{NOTE 1} \qquad (6)$$

By (5) and (6) and adopting the ergodic hypothesis

$$\tilde{C}_t = \frac{1}{\Delta U \cdot \delta t} \int_{t-\delta t/2}^{t+\delta t/2} dt' \int_{(\Delta U)} C(x,y,t')dU = EC(x_o,y_o,t) \qquad (7)$$

ESTIMATION OF $\tilde{C}$ FROM POINT SAMPLES IN A CELL

Consider a set of k sampling points in a cell with n samples taken at each of these points during the time interval δt centered at t. According to the model presented by (5), the concentration, C_{tij} obtained from the j-the sample $(j = 1,2,\ldots,n)$ at the i-th point $(i = 1,2,\ldots k)$ is given by:

$$C_{tij} = \alpha_t^* + \beta_t X_i + \gamma_t Y_i + \varepsilon_{ti} + \eta_{tij} \qquad (8)$$

where

$$\alpha_t^* = \alpha_t + \beta_t(\bar{x}-x_o) + \gamma_t(\bar{y}-y_o), \quad X_i = x_i - \bar{x}, \quad Y_i = y_i - \bar{y}$$

$\bar{x},\bar{y}$ are the coordinates of the centroid of the set of sampling points. The translation from x_o,y_o to $\bar{x},\bar{y}$ was made for the sake of computational convenience only. Replacement of an index by a dot will indicate an average of the observations over that index. Thus, refering to a given time t, the average of the concentrations obtained from the n samples at a given sampling point, i, is:

$$C_i \equiv C_{ti.} = \alpha^* + \beta X_i + \gamma Y_i + \varepsilon_i + \eta_i. \qquad (9)$$

The variance of C_i follows from (6) and (9), yielding

$$\mathrm{Var}C_i \equiv \sigma^2 = \sigma_1^2 + \sigma_2^2 \cdot \frac{f(\rho)}{n}\ , \quad f(\rho) = [1+\rho-2\rho(1-\rho^n)/(n(1-\rho))]/(1-\rho) \qquad (10)$$

Our next task is to obtain estimates of the parameters α^*,β,γ from the

n x k obsefvations as a step towards deriving the variance of the estimate of $\tilde{C}$.

Denoting

$$X = \begin{bmatrix} 1 & X_1 & Y_1 \\ \cdot & \cdot & \cdot \\ 1 & X_k & Y_k \end{bmatrix} \qquad \underline{\beta} = \begin{bmatrix} \alpha^* \\ \beta \\ \gamma \end{bmatrix} \qquad \underline{C} = \begin{bmatrix} C_1 \\ \cdot \\ C_k \end{bmatrix} \tag{11}$$

the least square estimator of $\underline{\beta}$ is (ref. 3)

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{C} = \begin{bmatrix} (S_y^2 S_{cx} - S_{xy} S_{cy})/\Delta \\ (S_y^2 S_{cx} - S_{xy} S_{cy})/\Delta \\ (S_x^2 S_{cy} - S_{xy} S_{cx})/\Delta \end{bmatrix} \tag{12}$$

where: $S_{uz} = \sum_{i=1}^{k} (u_i - \bar{u})(z_i - \bar{z})/k$, $S_u^2 = S_{uu}$, $r_{uz} = S_{uz}/(S_{ux}S_z)$. (13)

$$\Delta = S_y^2(1-r_{xy}^2)$$

The least square estimator of EC at any point X,Y is

$$\hat{EC}(X,Y) = [1, X, Y]\,\hat{\underline{\beta}} \tag{14}$$

Whence (ref. 3)

$$\text{Var } \hat{EC}(X,Y) = \sigma^2 [1,X,Y](X'X)^{-1} \begin{bmatrix} 1 \\ X \\ Y \end{bmatrix} \tag{15}$$

or, upon substitution of (10)

$$\text{Var}\hat{EC}(X,Y) = \frac{1}{k}(\sigma_1^2 + \sigma_2^2 \frac{f(\rho)}{n}) \cdot [1 + \frac{X^2}{S_x^2(1-r_{xy}^2)} + \frac{Y^2}{S_y^2(1-r_{xy}^2)} - \frac{2XYr_{xy}}{S_x S_y (1-r_{xy}^2)}] \tag{16}$$

By (7): $\text{Var}(\tilde{C}-\hat{C}_L) = \text{Var}(\tilde{C}-\hat{EC}(X_o,Y_o)) = \text{Var } \hat{EC}(X_o,Y_o)$

Hence, by (16):

$$\text{Var}(\tilde{C}-\hat{C}_L) \equiv \sigma_L^2 = \frac{1}{k}(\sigma_1^2 + \sigma_2^2 \frac{f(\rho)}{n}) [1 + \frac{(X_o^*)^2 + (Y_o^*)^2 - 2X_o^* Y_o^* r_{xy}}{1-r_{xy}^2}] \tag{17}$$

where $X_o^* = \frac{x_o - \bar{x}}{S_x}$, $Y_o^* = \frac{y_o - \bar{y}}{S_y}$

OPTIMAL SAMPLING PROGRAM

By (17) $L = \{k, n, \phi(X_o^*, Y_o^*, r_{xy})\}$, $\phi = \frac{(X_o^*)^2 + (Y_o^*)^2 - 2X_o^* Y_o^* r_{xy}}{1-r_{xy}^2} + 1$ (18)

where k is the number of sampling points in the cell, n is the number of samples taken at each point within a prescribed interval of time centered at the time t, for which $\tilde{C}$ is evaluated and ϕ is a configuration function of the network.

The optimal sampling program is obtained by the following procedure:

(a) Formulate the loss function $\mathcal{L}(L)$

(b) Derive the risk function $E\mathcal{L}(L)$

(c) Formulate the cost function $CS(L)$

(d) Find $\text{Min}[E\mathcal{L}(L) + CS(L)]$ subject to $|x_o-\bar{x}| < \frac{\Delta x}{2}$, $|y_o-\bar{y}| < \frac{\Delta y}{2}$

Sometimes it may be difficult to formulate the expected loss stemming from the "information gap" caused by the level of L.

In this case one may replace the risk function by a constraint on the accuracy and/or reliability of the estimate of $\tilde{C}$ and seek under this constraint either a sampling program at minimum cost or a program of minimum observations which just satisfies the constraint. This approach is applicable when the penalty for violating the constraint is so high that the optimal program must stay within the limits prescribed by the constraint.

Selection of an optimal sampling program by any of the methods outlined above requires that the variances of random fluctuations of observations in space, σ_1 , and in time, σ_2 , be known.

Actually these parameters have to be estimated from samples of field data.

The residual sum of squares of the linear regression model (9) is given by (ref. 3)

$$R_o^2 = \underline{C}'\underline{C} - \underline{C}'X\hat{\underline{\beta}} = \hat{\sigma}^2 \cdot (d.f.) \tag{19}$$

where (d.f.) is the number of degrees of freedom.

By (9) and (12)

$$R_o^2 = \sum_{i=1}^{k} (C_i - C_\cdot)^2 - k\, S_c^2 (r_{cx}^2 + r_{cy}^2 - 2r_{cx}r_{xy}r_{cy})/(1-r_{xy}^2) \tag{20}$$

Assuming that the variance σ^2 is independent of the time t around which the observations are made, one may evaluate σ^2 on the basis of the entire sample of historic observations around the time points $t = 1,2,\ldots,T$.

Hence, by (20):

$$\hat{\sigma}^2 = \Big[\sum_{t=1}^{T} \sum_{i=1}^{k} (C_{ti.} - C_{t..})^2 - k \sum_{t=1}^{T} S_{c(t)}^2 (r_{c(t)x}^2 + r_{c(t)y}^2 - 2r_{c(t)x}r_{c(t)y}r_{xy})/(1-r_{xy}^2) \Big]$$

$$/[T.(k-3)] \tag{21}$$

The estimate of $\sigma_2^2 = \text{Var}\ \eta$ (see (6)) is obtained by considering all historic deviations of local observations at each sampling point from their average, for each t separately:

$$\hat{\sigma}_2^2 = \sum_{t=1}^{T} \sum_{i=1}^{k} \sum_{j=1}^{n} (C_{tij} - C_{ti.})^2 / [Tk(n-1)] \qquad (22)$$

Hence, by (10):

$$\hat{\sigma}_1^2 = \hat{\sigma}^2 - \hat{\sigma}_2^2 \frac{f(\rho)}{n} \qquad (23)$$

APPLICATION TO THE COASTAL AQUIFER OF ISRAEL

The methodology outlined in this paper is presently implemented in updating the annual groundwater quality sampling program in the coastal aquifer of Israel. The aquifer is subdivided into rectangular cells. Observation wells in each cell are sampled for chlorides on a semiannual basis, a routine prescribed by budgetary and time constraints. Attempting at a more rational approach which aims at maximizing the utility of the program to groundwater management, while circumventing the difficulty of specifying the loss function, we split the problem in two parts:

(a) Given $\sigma^2_{L(1)}$, $CS(L) = CS(n,k)$

Find $L_{opt} = (n_{opt}, k_{opt})$ such that $CS(L_{opt}) = \min$ (24)

subject to $(\sigma_1^2 + \sigma_2^2/n)/k = \sigma^2_{L(1)}$

(b) Given k_{opt}

Find (x_i, y_i) $i = 1,2,\ldots,k_{opt}$ such that $(\sigma_1^2+\sigma_2^2/n_{opt})\phi/k_{opt} \to \sigma^2_{L(1)}$

The first part seeks a least cost program in terms of the number of wells and frequency of observations in a cell subject to an exogenous constraint on the accuracy of the estimated concentration.

The second part takes into account the fact that the accuracy of estimating $\hat{C}$ also depends on the configuration of the wells and seeks a configuration which satisfies the constraint as close as possible.

In our problem the cost function is formulated on an annual basis and consists of the following components:

$$CS(L) = C_o k^{a_o} + C_1 k^{a_1} + C_2 (n \cdot k)^{a_2} + C_3 (n \cdot k)^{a_3} \quad \text{NOTE 2} \qquad (25)$$

The parameters a_i $(i = 0,1,2,3)$ are evaluated on the basis of professional expertise. As a rule these are positive parameters smaller than one, reflecting the fact that the marginal cost of each component is decreasing with the size of the program.

As to the prescription of $\sigma^2_{L(1)}$ we took into account the fact that the recipient of the data may prefer to express his requests in terms of accuracy (Δ) and reliability $(1-\alpha)$ of estimating the average concentration in a cell.

The relationship between $\sigma^2_{L(1)}$ and these quantities is given by Chebyshev's inequality (ref. 3)

$$P(|\tilde{C}-\hat{C}_L| \geq \Delta) \leq \sigma^2_{L(1)}/\Delta^2 \qquad (26)$$

where Δ is a prescribed level of accuracy. (26) expresses the idea that the deviation $|\tilde{C}_L-C_L|$ will exceed the value of Δ in no more than $(\sigma_{L(1)}/\Delta)^2 \times 100$ per cent of cases, on the average.

Consider the constraint

$$P(|\tilde{C}-C_L| > \Delta) = \alpha$$

By (24) and (26) this constraint will be satisfied if

$$\sigma^2_{L(1)}/\Delta^2 = (\sigma_1^2+\sigma_2^2/n)/k \leq \alpha\Delta^2 \qquad (27)$$

This constraint is valid for any distribution function of the observations. In pursuit of a least cost program (27) should be replaced by an equality constraint, thus providing a minimal program, which is sufficient to satisfy the constraint. Selection of prescribed levels of Δ and α, along with the specification of (25), concludes the formulation of the part (a) problem in terms which are suitable for the recipient of the data.

An important merit of solving part (b) separately from part (a) is the possibility of taking into account constraints regarding the location of wells as prescribed by the actual conditions in the field.

SUMMARY AND CONCLUSIONS

Rational management of a groundwater system is based among other things on information about the present level of solute concentrations in the groundwater. For practical reasons this information is sought in terms of averages over spatial subdomains, referred to as aquifer cells, and over temporal subdomains around given time points.

Errors in estimating these averages on the basis of field data carry an opportunity loss which can be reduced by increasing the number of observations.

This benefit, however, is curtailed by the cost of data acquisition. Therefore the problem becomes one of designing a level of information which reduces the total loss to a minimum.

The present paper expresses the level of information in terms of three parameters of a groundwater quality sampling program, namely: the number of sampling points, the spatial configuration of these points and the frequency of sampling.

The paper proposes a methodology for deriving an optimal sampling program by minimizing the total risk (i.e. the sum of the expected opportunity loss and the cost of observations).

The solution procedure takes as a point of departure a spatially linear statistical model of point observations in a cell. Underlying this model is

the assumption that any single observation contains a random part which consists of two independent components: a spatial component and a temporal one. This model leads to the computation of the standard deviation of the estimated average concentration as a function of the parameters of the statistical model and the parameters of the sampling program. Once derived, this standard deviation can be used to evaluate the risk for any given distribution function of the loss. Together with the cost it yields an objective function which explicitly depends on the parameters of the sampling program.

In case of a difficulty in evaluating the expected loss, the latter may be replaced by a prescribed constraint on the accuracy and reliability of the estimated average concentration. Thereby the problem reduces to one of finding a least cost sampling program which satisfies the constraint.

The statistical model was formulated in this paper in two spatial dimensions and took into consideration the effect of autocorrelation. An extension of the model to three dimensions and inclusion of cross-correlation between observations in space is straightforward. The methodology developed in this paper can also be applied to any program of observations.

ACKNOWLEDGEMENT

This paper is part of an ongoing research project on methodologies for planning groundwater observation programs conducted at the Hydrological Service of Israel. The authors are indebted to Mr. M. Jacobs, Director of the Hydrological Service, for his permission to publish the paper.

REFERENCES

1 J. Bear, Dynamics of Fluids in Porous Media, American Elsevier, New York, 1972, Ch. 10.
2 R.C. Ward et al, Statistical Evaluation of Sampling Frequencies in Monitoring Networks, Journal WPCF, Vol. 51, No. 9, 1979, 2292-2300.
3 C.R. Rao, Linear Statistical Inference and Its Application, John Wiley & Sons, Inc., 1967.

(1) $|\tau-\tau'|$ denotes a number of elementary time steps between two observations.
(2) The first term on the right-hand side represents the cost of construction and maintenance of wells; the second term represents the cost of visiting the wells; the third term represents the cost of sampling and analyses, whereas the last term refers to the cost of data processing.