Jul 1st, 12:00 AM

# Optimal Sampling for Parameters Estimation

Federico Catania

Marco Massabò

R. Minciardi

Ombretta Paladino

# Optimal Sampling for Parameters Estimation

**Federico Catania**[a,c], **Marco Massabò**[a,c]**, Riccardo Minciardi**[a,b]**, Ombretta Paladino**[a,c]**, Michela Robba**[a,b]

[a]*CIMA - Centro di Ricerca in Monitoraggio Ambientale*
[b]*DIST – Department of Systems and Computer Sciences*
[c]*DIAM – Dipartimento di Ingegneria Ambientale*

*email: federico.catania@cima.unige.it, m.marco@cima.unige.it, riccardo.minciardi@unige.it, paladino@unige.it, michela.robba@unige.it*

**Abstract**    In the problems concerning prediction and modeling, parameters estimation constitutes one of the main uncertain items that must be taken into account. The easiest way to minimize this uncertainty is to collect great amounts of data. The aim of this work is to build a decision model able to choose the optimal position of the sample point used for the parameters estimation, minimizing the parameters uncertainty. The decision model is applied to the estimation of the dispersivity coefficients, longitudinal and transversal, from soil column experiment. The classical design of experiments techniques are based on the optimization of the amount of information obtained from experimental data with the hypothesis that the sample domain is defined on a continuous space over time and position. Since this assumption does not reflect the real experimental situation, especially when field campaigns are to be performed and the position of the piezometric wells is fixed, an approach based on discrete optimization over a fixed grid of possible sampling is proposed.

The soil column representation is discretized in the 2D domain, while the concentration experimental data are generated using a rigorous analytical solution of the advection dispersion model and a Monte Carlo simulator to generate the experimental error at given variance. In order to define the optimal sampling points in the soil column, binary decision variables are introduced: they assume value one when the concentration is measured at a specific point and time, zero otherwise. The objective function to be finally minimized is proportional to the calculated covariance of the estimated parameters and to the decision variables.. The formalized constraints regard the possible number of measures, according to the available funds. Finally, the results of the optimisation problem are discussed.

*Keywords:* Optimal experimental design; Parameter estimation; Column outflow experiments; Solute transport.

## 1. INTRODUCTION

Simulation of the environmental fate of solutes is becoming a theme of great interest in research studies. Several analytical solutions (two or three dimensional) of the mathematical models describing pollutant transport are usually available for non-reactive contaminants. In order to use these models to predict substance concentrations, the values of all parameters present in the mathematical model have to be disposable. Typically, parameters of dispersion describing the flow of water in the solute transport equation are unknown and have to be identified. Specifically, the dispersion coefficient is an essential parameter for the control of water pollution. Several investigations have been done in modeling the contaminants dispersion and in predicting the

distribution of pollutant downstream from its point of discharge (e.g. Fischer [1967]).

In this work the attention is focused on the inverse modeling approach. It takes to the estimate of unknown model parameters from measurement data, e.g. concentration data, by mathematical optimization. An objective function that contains quadratic deviations between computed and observed data is minimized. Some author that investigated this kind of inverse problem encountered troubles with ill posedness of the parameter estimation problem (e.g. Toorman et al. [1992], van Dam et al. [1992]), like insensitivity of the parameters to observed data. The question that is central in this paper is how high-quality data could be obtained. For this reason, methods to optimize experimental designs with regards to parameter estimation are considered. In optimal experimental design theory, design criteria are defined on the variance-covariance matrix, which summarizes the statistical properties of parameter estimates. The smaller the entries in the diagonal of the variance-covariance matrix, the more consistent are the parameter estimates. Our objective is to improve estimation results by identifying sampling schemes that are most likely to yield parameter estimates with low variances.

Few studies consider optimal experimental design problems in mathematical optimizations using statistical design criteria (e.g. Hsu and Yeh [1989], Wagner [1995], Altmann-Dieses et al. [2002]). Here an optimal experimental design problem for a typical column outflow experiment is considered with an approach that enables to optimize the sampling design, i.e. the allocation of measurement points in different time and space. The classical design of experiments techniques are based on the optimization of the amount of information obtained from experimental data with the hypothesis that the sample domain is defined on a continuous space over time and position. Since this assumption does not reflect the real experimental situation, especially when field campaigns are to be performed and the position of the piezometric wells is fixed, an approach based on discrete optimization over a fixed grid of possible sampling is proposed.

A soil column representation is discretized in the 2D domain, in cylindrical geometry, while the concentration experimental data are generated using the rigorous analytical solution of the advection dispersion model and a Monte Carlo simulator to generate the experimental error at given variance. In order to define the optimal sampling points in the soil column, binary decision variables are introduced: they assume value one when the

concentration is measured at a specific point and time, zero otherwise. The objective function to be finally minimized is proportional to the calculated variance-covariance matrix of the estimated parameters and to the decision variables. The formalized constraints regard the possible number of measures, according to the available funds. Finally, the results of the optimization problem, calculated through two different integer programming algorithms, a branch-and-bound algorithm (lingo software, lindo systems) and a genetic algorithm (implemented with the MATLAB Genetic Algorithm Toolbox), show the best places to measure concentration in column over this discretized grid, in order to minimize the parameter uncertainty.

## 2.    THE PROBLEM STATEMENT

In this work, the model considers fluid flow in a saturated porous media composed by a column in which there is an aqueous liquid phase and a solid phase assembled in a matrix of porosity $n$ and density $\rho$ (see Figure 1). A soluble pollutant can be transported in it by groundwater flow through the void space and dispersed mainly by two processes: molecular diffusion and hydrodynamic dispersion.
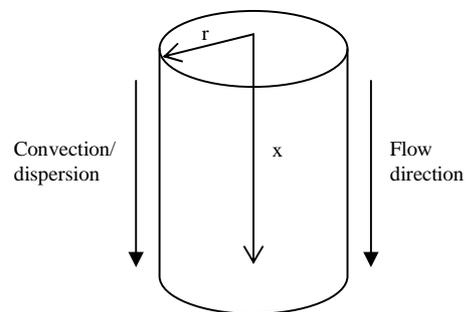


**Figure 1**: Scheme of the solute transport in the considered system.

### 2.1    Solute transport

The solute transport equation in cylindrical geometry for a non-reacting solute, taking into account radial and axial dispersion, is represented by a mathematical model that is one of the most adopted experimental device to investigate pollutant transport phenomena. Under the geometry just discussed, we furtherly assume that the initial conditions do not depend on the angular variable; so, the process preserves symmetry around the longitudinal axis.

The convection-dispersion PDE expressing the mass balance of a generic solute in terms of dimensional concentration $C(r,x,t)$, can be written as follows:

$$\frac{\partial C^*}{\partial t^*} + u^* \frac{\partial C^*}{\partial x^*} = D_R \left( \frac{\partial^2 C^*}{\partial r^{*2}} + \frac{1}{r^*} \frac{\partial C^*}{\partial r^*} \right) + D_L \frac{\partial^2 C^*}{\partial x^{*2}} \quad (1)$$

Here the constant advective term is represented by the average pore water velocity $u^*$ and anisotropic dispersion is described by means of the two mechanical dispersion coefficients $D_R$ and $D_L$. They represent different dispersion mechanisms such as molecular diffusion, hydrodinamic dispersion, eddy diffusion or mixing. Using typical dimensionless variables, the equation becomes:

$$\frac{\partial C}{\partial t} + u \frac{\partial C}{\partial x} = \frac{\eta}{Pe_R} \left( \frac{\partial^2 C}{\partial r^2} + \frac{1}{r} \frac{\partial C}{\partial r} \right) + \frac{1}{Pe_L} \frac{\partial^2 C}{\partial x^2} \quad (2)$$

where:

$$Pe_R = \frac{U_= R}{D_R}; \quad Pe_L = \frac{U_= L}{D_L}; \quad \eta = \frac{L}{R}; \quad (3)$$

with $R$ and $L$, respectively the radius and the length of the column and $U_0$ and $C_0$ the scales of velocity and concentration.
Boundaries and initial conditions are necessary to have a unique solution. We assume:

$$\left. \frac{\partial C(r,x,t)}{\partial r} \right|_{r=1} = 0; \quad (4)$$

$$\lim_{x \to +\infty} C(r,x,t) = 0; \quad (5)$$

$$\lim_{x \to +\infty} \frac{\partial C}{\partial z} = 0; \quad (6)$$

where the first condition represents the impermeability of the column walls.

Considering the following further initial and boundary conditions describing the pollutant release:

$$C(r,0,t) = C_0 H(t); \quad (7)$$

$$C(r,x,0) = 0; \quad (8)$$

in which $H(t)$ is the heavyside function, finally we obtained (Massabò et al., 2004):

$$C(r,x,t) = \sum_{k=0}^{+\infty} A_K J_0(Z_k^1 r) \frac{1}{2} \exp\left[ \frac{x u Pe_L}{2} \right] \cdot$$

$$\cdot \left\{ \exp\left[ x \sqrt{\frac{u^2 Pe_L^2}{4} + \eta [Z_k^1]^2 \frac{Pe_L}{Pe_R}} \right] \cdot \right.$$

$$\cdot erfc\left[ \frac{1}{2} x \sqrt{\frac{Pe_L}{t}} + \sqrt{\frac{u^2 Pe_L t}{4} + \eta \frac{[Z_k^1]^2}{Pe_R} t} \right] +$$

$$\cdot \left\{ \exp\left[ -x \sqrt{\frac{u^2 Pe_L^2}{4} + \eta [Z_k^1]^2 \frac{Pe_L}{Pe_R}} \right] \cdot \right.$$

$$\cdot erfc\left[ \frac{1}{2} x \sqrt{\frac{Pe_L}{t}} - \sqrt{\frac{u^2 Pe_L t}{4} + \eta \frac{[Z_k^1]^2}{Pe_R} t} \right] \quad (9)$$

And the coefficient $A_k$ are given by:

$$A_k = \frac{2 \int_0^1 \rho \, f(\rho) J_0(Z_1^k \rho) \, d\rho}{[J_0(Z_1^k)]^2}, \quad k = 1,2,...$$

$$A_0 = \int_0^1 \rho \, f(\rho) \, d\rho \quad (10)$$

## 2.2 Parameter estimation

In this context, we are in the situation that the unknown parameters $Pe_R$ and $Pe_L$ have to be estimated from measurement data $C_{sper}$ representing the concentrations of the solute, recorded at different depths $x_i$ and different times $t_i$. In the analysis here proposed, we assume the general case that measures can be simulated by adding to model outputs some experimental error belonging to a normal distribution with zero mean $\varepsilon \sim N(0, \sigma^2)$:

$$C_{sper} = C(r,x,t) \cdot (1 + \varepsilon) \quad (11)$$

Under the hypotheses of a reduced model and negligible experimental errors on measures of spatial-temporal variables, the parameter estimation procedure consists in finding the best two unknown parameters, $Pe_R$ and $Pe_L$, that solve the following least square problem:

$$\min \quad \sum_r \sum_x \sum_t (C(r,x,t) - C_{sper})^2 \quad (12)$$

Since true values $C(r,x,t)$ are the output of equation (9), this is a nonlinear optimization problem and sometimes the location of the minimum often involves an iterative search of the parameter space. Initial guess values or previous estimates of

parameters $Pe_R$ and $Pe_L$ must be supplied. The numerical method here used is based on a Marquardt's modified algorithm (Marquardt [1963], Bard [1974]) with analytical first order derivatives supplied by the authors.

## 2.3 The optimization model for the design of experiments

The main objective of the optimisation model is to design the optimal sampling by determining, in time and space, the best points $P_i(r,x,t)$ for measuring solute concentrations, in order to obtain the best estimates for $Pe_R$ and $Pe_L$. The objective function corresponds to an approximation of the parameter variance-covariance matrix that should be minimized (Bard [1974]). The decision variables, $\delta_z$ (z=1..,Z), where Z represents the total number of possible sample points to select, are binary and represent the possible measures that can be selected in order to perform analysis in the column. They assume value 1 when the sampling point is chosen, 0 otherwise. The constraints regard a maximum allowable number of sample points, according to a pre-defined budget. The first-order approximation of the parameter variance-covariance matrix is:

$$\tilde{V} = (B^T \Pi^{-1} B + V_0^{-1})^{-1} \qquad (13)$$

where $B$ is the matrix of sensitivities of simulated concentration respect to the parameters $Pe_R$ and $Pe_L$, $\Pi$ is the error covariance matrix for solute concentration measurements and $V_0$ is the error covariance matrix for parameter estimates, calculated in our case after a run of unconstrained minimization (12). Besides, this run supplies also the estimates of $Pe_R$ and $Pe_L$ in order to insert them in the calculus of $B$ and $\Pi$.

So the purpose is to identify the sampling strategy that minimizes the trace (A-optimal design, Jacquez, [1998]) of $\tilde{V}$, subject to a constraint on the total number of sampling points.

$$\min \left[ \; tr\left( \tilde{V} \right) \; \right] \qquad (14)$$

subject to

$$\sum_z \delta_z = A \qquad (15)$$

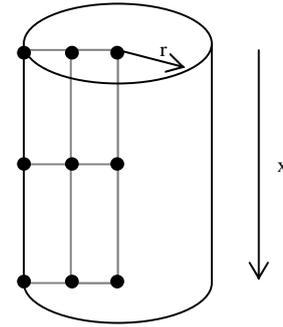where A is the maximum allowed number of sample points (Figure 2).



**Figure 2**: Grid of possible spatial points in the column.

The equations (14) and (15) enable to quantify the model-parameter uncertainty as a function of the available concentration and, besides, they can also be used to estimate the worth of alternative sampling strategies for reducing parameter uncertainty.

In this paper, two algorithms were used to solve the problem: a branch-and-bound algorithm and a genetic algorithm.

### 2.3.1 Branch-and-bound algorithm

We used a branch-and-bound algorithm similar to the ones presented by Carrera et al. [1984] and Wagner [1995]. This algorithm arranges through different sampling strategies to find the single combination of measurements that minimizes the uncertainty while keeping the total number of data constant, but inferior to the number of possible points on the grid. The algorithm is based on the fact that uncertainty cannot increase when data are added to the measurement network and cannot decrease when data are removed from this network.

### 2.3.2 Genetic algorithm

The genetic algorithm (Goldberg [1989]) uses a scheme of directed random search to sort through the alternative designs and identify efficient sampling strategies. It considers each sampling alternative to be a "string creature" consisting of zero-one indicator variables, $\delta_z$, $\delta_z$=1 if measurement is taken, zero otherwise. It begins by randomly generating $j$ string creatures, each representing an alternative sampling design, and then, keeping on generating through three operators (reproduction, cross-over and mutation), it reaches the best sampling strategy, in which the variance-covariance matrix is minimized, with the constraint of the total number of data.

## 3. CONCEPTUAL MODEL

The experimental design models presented in the previous section will be demonstrated using a hypothetical grid on the column with 18 available spatial and temporal points; the use of this grid provides a valuable mechanism for evaluating the performance of the two algorithm models. Recall that the transport problem presented here is transient, so the concentrations at any $P_i(r,x,t)$ will vary through time, as will the information content of the concentration measurements.

### 3.1 Numerical simulations

Numerical simulations were carried out in order to get to the problem solution. Equation (2) was solved analitically with the boundary and initial condition (4) - (8). Then, after generating $C_{sper}$ with an error $\varepsilon \sim N(0,0.01)$, expression (12) was minimized in order to get to the first parameters guess values (initial set of data for estimating parameters and quantifying parameter-estimate uncertainty). The values of the hydraulic and geometric properties inserted to solve (13) – (15) are summarized in Table 1.

| $L$ [m] | 100 | |
|---|---|---|
| $R$ [m] | 10 | |
| $U_0$ [m$^2$/s] | 1 | |
| $PeR\_guess$ | 99,5 | |
| $PeL\_guess$ | 100,9 | |
| $V_0$ | 1,2 | 0 |
| | 0 | 2,8 |

**Table 1**: Geometric and hydraulic characteristic of the simulated experiment.

*PeR_guess* and *PeL_guess* are mean of the parameter estimates, calculated after 30 runs of unconstrained minimization (12) (each with a different error $\varepsilon$ generated by Monte Carlo simulator) and $V_0$ is the prior covariance matrix for parameter estimates, based on these 30 runs.
The properties of the genetic algorithm used are presented in Table 2.

| Parameter | Value |
|---|---|
| Number of strings | 90 |
| Number of generations | 30 |
| Crossover probability | 0,90 |
| Mutation probability | 0,01 |

**Table 2**: Summary of the parameter values adopted for the genetic algorithm.

### 3.2 Results

The branch-and-bound algorithm has been solved using Lingo software (Lindo Systems), while the genetic algorithm was settled using the MATLAB Genetic Algorithm Toolbox. The binary variable $\delta_z$, that was introduced to define which are the most suitable experiments, has been found. Specifically, the problem has been solved for 18 possible sample points and a maximum of 10 sample points, according to economic considerations.
Figure 3 reports the 10 optimal sample points and the times in which we obtained the maximum reduction in parameter-estimate uncertainty, comparing branch-and-bound and genetic algorithm. We can notice that both experimental design models favour measurements points near the inlet section.
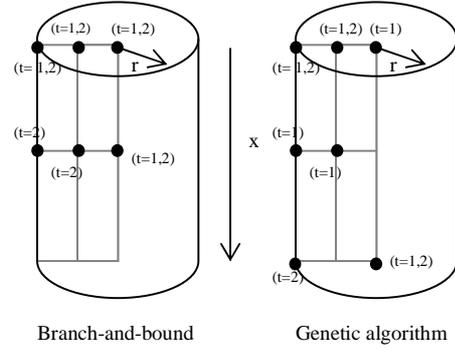


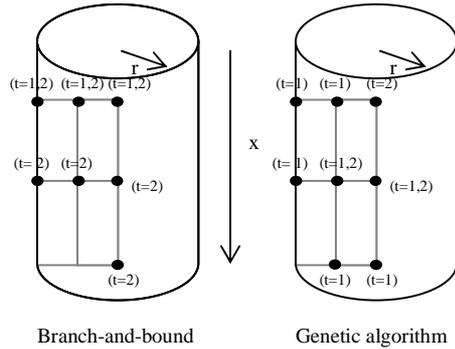**Figure 3**: Comparison between obtained results.



**Figure 4**: Comparison between obtained results, considering another grid of possible points. Also here points near the inlet section are favourite.

At the same way, to see how the model progressively works, the optimization problem has been solved, excluding the initial section (x=0) and inserting 3 new points in which x=0,2, in order to neglect, in the

5

experimental design procedure, the information coming from the boundary condition. Figure 4 reports the results of this new optimization problem; the points nearest to the inlet section are again privileged. Besides, it is possible to notice that both algorithms lead to similar results.

## 4.   CONCLUSION

In this work, we propose a decision model able to choose the optimal position of the sample point used for the parameters estimation, minimizing the parameters uncertainty. The decision model is applied to the estimation of the dispersivity coefficients, longitudinal and transversal, from soil column experiment. After discretizing the soil column representation in the 2D domain and generating the concentration experimental data using a rigorous analytical solution of the advection dispersion model and a Monte Carlo method to simulate the experimental error at given variance, binary decision variables are introduced in order to define the optimal sampling points in the soil column. The objective function that was minimized with constraints regarding the possible number of measures is proportional to the calculated covariance of the estimated parameters and to the decision variables. The constraints regard economic consideration about the allowable number of sampling measurements.

Finally, the results have been reported using two different optimisation techniques, in order to test the model and to achieve efficient results for the design of the experiments. Both algorithms show that the best sampling points are the points that are closest to the inlet section where the concentration profile is more affected by the dispersion parameters. Further developments regard the sensitivity analysis of the obtained results on the input parameters of the model (the experimental error variance, the number of possible sampling points, etc.).

## 5.   REFERENCES

Altmann-Dieses, A.E., J.P. Schlöder, H.G. Bock, and O. Richter, Optimal Experimental Design for Parameter Estimation in Column Outflow Experiments, *Water Resources Research, 38*(10), 1186, doi:10.1029/2001WR000358, 2002.

Bard, Y., *Nonlinear Parameter Estimation,* Academic, 341 pp., San Diego, Calif., 1974.

Carrera, J., E. Usunoff, and F. Szidarovszky, A Method for Optimal Observation Network Design for Groundwater Management, *Journal of Hydrology, 73*, 147-163, 1984.

Fischer, H.B., The mechanics of dispersion in streams, *Journal of the Hydraulic Division, 93*(6), 187-216, 1967.

Goldberg, D.E., *Genetic Algorithm in Search, Optimization and Machine Learning*, 412 pp., Addison-Wesley, Reading, Mass., 1989.

Hsu, N.-S., and W. W.-G. Yeh, Optimal Experimental Design for Parameter Identification in Groundwater Hydrology, *Water Resources Research, 25*(5), 1025-1040, 1989.

Jacquez, J.A., Design of Experiments, *Journal of Franklin Institute, 335*(2), 259-279, 1998.

Marquardt, D.W., An Algorithm for Least Squares Estimation of Nonlinear Parameters, *SIAM Journal of Applied Mathematics, 11,* 431-441, 1963.

Massabò, M., R. Cianci and O. Paladino, Some Analytical Solutions for the Dispersion-Convection – Reaction Equation in cylindrical geometry, submitted.

Toorman, A.F., P.J. Wierenga, and R.G. Hills, Parameter Estimation of Hydraulic Properties from One-step Outflow Data, *Water Resources Research, 28*(11), 3021-3028, 1992.

van Dam, J.C., J.N. Stricker, and P. Droogers, Inverse Method for Determining Soil Hydraulic Functions from One-step Outflow Experiments, *Soil Science Society of America Journal, 56*, 1042-1050, 1992.

Wagner, B.J., Sampling Design Methods for Groundwater Modeling under Uncertainty, *Water Resources Research, 31*(10), 2581-2591, 1995.