Data-driven and multi-approach sampling scheme optimization: the Alimini Lakes aquifer case

E. Barca[1,*], M.C. Caputo[1], L. De Carlo[1], R. Masciale[1], G. Passarella[1]

[1] *Water Research Institute of the National Research Council, Department of Bari, Viale F. De Blasio, 5 70123 Bari, Italy; emanuele.barca@ba.irsa.cnr.it; maria.caputo@ba.irsa.cnr.it; lorenzo.decarlo@ba.irsa.cnr.it; rita.masciale@ba.irsa.cnr.it; giuseppe.passarella@ba.irsa.cnr.it.*

[*] *Corresponding author*

Abstract. *Due to the high wells drilling cost, monitoring sites are usually selected among existing wells; nevertheless, the resulting monitoring network must assure a good assessment of the main characteristics of the considered aquifer. Groundwater managers, need to find a good balance between two conflicting objectives: maximizing monitoring information and minimizing costs. In this paper, a couple of groundwater monitoring optimization methods are presented, related to the local shallow aquifer of the Alimini Lakes, located in Apulia (South-Eastern Italy) where a large number of existing wells have been pinpointed and the need of optimally reducing exists. The proposed methods differ each other for the required amount of prior information. The first proposed method, namely Greedy Deletion, just requires the geographical position of the available sites, while the second, the Simulated Annealing, also requires the knowledge of the spatial law of the considered phenomenon. The managerial need was to halve the number of monitoring sites minimizing the information loss.*

Keywords. *Monitoring networks; Shallow aquifers; Greedy deletion; Spatial simulated annealing.*

1 Introduction

Groundwater monitoring is generally rather expensive due to the wells drilling costs. Usually, monitoring networks (MNs) are designed selecting those most capable of representing the groundwater status among a wide number of wells. Nevertheless, the available wells are often irregularly spread and an unwise selection of them may cause a biased understanding of the monitored water body (Barca et al., 2015). However, a wise selection of wells is strongly driven by the a priori knowledge of the considered water body. Different approaches can be found in scientific literature, related to the OMNR, whose application is driven by the available information (Nunes et al., 2002; Wu, 2004). In general, the OMNR is an optimisation problem that is solvable through the quantitative formulation of one or more Objective Functions (OFs). The choice of the OF is strongly dependent on the available information. Nevertheless, this information is not always available simultaneously and constrains the choice of the optimization methodology driving the OF selection (Barca et al., 2015). In practice, when the spatial behaviour of the monitored parameter is known, a model-based method can be used and the OF will be strongly dependent on the variable. Conversely, when the a priori knowledge is poor, only a design-based approach applies, since these latter methods exploit geometric characteristics as OF. In this paper, two OMNR methods are presented belonging to model-based and design-based categories, respectively. In particular, the GD and the SSA methods are applied to the optimal downsize of the Alimini Lakes groundwater monitoring network, initially made of 85 wells, covering a planar area of about 80 km2. Theoretical and applicative issues are reported referred to halving the original network.

2 Materials and Methods

Optimization methods

The OMNR issue can be brought back to a combinatorial problem of extracting a subset of cardinality k with some specified properties (i.e., number of locations to be removed), from a set of cardinality N (i.e., the initial network size) (Barca et al., 2015). The exhaustive exploration of the whole solution space S^N is almost always unfeasible because it is too computationally demanding. Consequently, the use of an optimization heuristics capable of reducing the solution space becomes necessary in order to simplify the issue. In practice, given the general monitoring aim and the a priori available information, a quantitative criterion

$$\phi(S): S^N \to R^+ \tag{1}$$

where $\phi(S)$ is the OF, must be defined which automatically leads to a specific optimization method category (model/design-based). In general, the stages of an optimization method can be summarized as follows, independently from its category:

1. a starting reduced configuration S_0 is defined;
2. a new candidate configuration S_i is generated by means of a sequential neighbouring search;
3. a decision rule checks each generated configuration and selects the optimal transient solution S^*;
4. a stop criterion states the convergence to the optimal solution, S^O.

In this paper, the GD method and the SSA method belonging to the model-based and design-based categories, respectively, are presented and applied.

Greedy Deletion

The GD method is a greedy heuristic for the optimal reduction of the monitoring network. The initial configuration S_0 is made up of all the locations of the original network. This method substantially operates three nested loops. The outer loop is governed by the problem size, namely the k locations to be removed. The second intermediate loop identifies the two closest locations s_a^* and s_b^* of the current optimal solution S^*. This step can actually be viewed as the generation of two candidate configurations, each made up of S^* reduced by s_a^* and s_b^*, respectively. Finally, at the inner loop, a simple decision rule is applied, which accepts as the optimal transient configuration one of the two which minimizes the OF:

$$\hat{\phi}_{NPD}(s_i, S_{N-l}\setminus\{s_i\}) = \frac{1}{|S_{N-l}\setminus\{s_i\}|}\sum_{s_j \in S_{N-l}\setminus\{s_i\}} d(s_i, s_j) \tag{2}$$

where $S_{N-l}\setminus\{s_i\}$ is the monitoring network under reduction deprived of s_i and $1 < l < k$ (Ortner et al. 2007). The method stops when the required number of locations (k) has been removed from the original network.

Spatial Simulated Annealing

The SSA is basically structured as a pre-processing stage followed by two nested loops. In the pre-processing, some parameters are estimated, needed to trigger the actual optimization method, namely the initial configuration S_0 and the initial temperature T_0. The outer loop is governed by the temperature and stops when this approaches zero. The inner loop is related to the problem size, that is, if k is defined as the number of locations to be added or removed, the inner loop consists of k iterations for a given temperature value. Within the inner loop, the candidate solutions are generated and subjected to the decision rule (Barca et al. 2015). Concerning the OF, the Average Ordinary Kriging Variance (AKV) has been used. The well-known ordinary kriging variance formulation in a generic unsampled location x_i is (Isaaks and Srivastava 1989):

$$\sigma_R^2(x_i) = \sum_{j=1}^{N} \lambda_j(x_i)\,\gamma(x_j, x_i) - \mu(x_i) \tag{3}$$

where $\lambda_j(x_i)$ are the kriging estimation weights, $\gamma(x_j, x_i)$ is the variogram value for the location pair (x_j, x_i), and $\mu(x_i)$ are the Lagrange multipliers. Consequently, the OF can be written as:

$$\phi_{AKW} = \frac{1}{N}\sum_{i=1}^{N} \sigma_R^2(x_i) \tag{4}$$

It is assumed that a priori knowledge about the spatial law (variogram model) of the variable to be monitored is available. Monitoring-network optimization based on AKV tends to remove locations where the monitoring information is redundant (Barca et al. 2008).

3 Study area

The Alimini lakes are two shallow coastal lakes located in the South-Eastern part of the Apulia Region along the Adriatic Sea coast (Figure 1). Actually, the Northern Lake, named Alimini Grande, is a lagoon, since it is directly connected to the Adriatic Sea through a narrow entrance. The smaller Lake is connected to the other by a natural channel. Both of them are mainly and constantly fed by groundwater recharge, through a number of coastal springs, but also, by surface runoff collected by a network of channels as well as directly by rainfall. In this study, the shallow aquifer has been considered, which is the only one directly connected to the Lakes. From a geological standpoint it is made by Plio-Pleistocene sediments, consisting of an alternating sequence of calcarenites, sands and sandy clays (Margiotta & Negri, 2005). The geometry of aquifer is often hard to determine, since the water lies in limited intervals of permeable rock in a more general context of impermeable deposits. With the aim of investigating the qualitative and quantitative features of the shallow aquifer, 76 agricultural and domestic wells were selected among those located in the neighbourhood of the lakes (Figure 1). This monitoring network is mostly made by dug wells and seldom by drilled wells, whose main characteristics, (e.g. depth, stratigraphy, etc.) are often unknown.

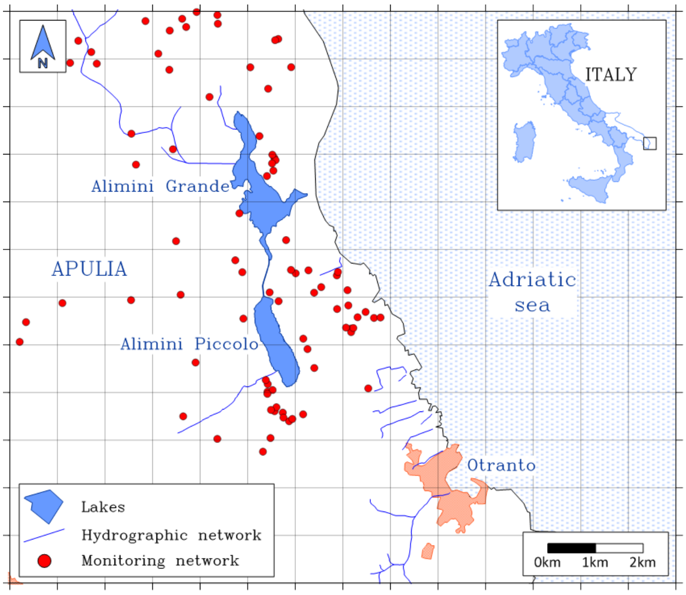

Figure 1: Study area

4 Results and Discussion

The two chosen optimization methods have been applied to the original monitoring network composed by 85 sites. Both the methods have been constrained to halve the network and after the optimization, 41 monitoring sites have been discarded. The network configuration produced by the SSA has an average kriging variance (AKV) which compared with the original network AKV has a percentage of worsening of about 0.9%. As it can be drawn from Figure 2, the two reduced configurations are very similar each other; in fact, they share about the 86% of the same sites. Consequently, it can be expected that, from the representativeness standpoint, the two reduced configurations behave in a similar fashion. In effect, if we try to re-estimate the respective discarded sites values by means of the two reduced network configurations, we obtain the following results:

	MBE	RMSE	RMAE
Greedy Deletion	0.302	1.358	23.161
SSA	0.028	1.367	21.917

Table 1: Summary statistics of network configurations performances.

Analyzing the Table I, we can see that the values estimated by means of SSA show to be significantly less biased than the GD ones. Furthermore, the percentage error (RMAE) is slightly better with respect the GD one. In summary, two out of three indices are very close each other. Consequently, we can conclude that the two reduced configurations perform as expected. A possible explanation of the similar structure and behavior of two configurations can be the extreme clustering of the complete network. Since, the two applied optimization methods tend to intervene on the geometrical configuration of the network; this can explain the obtained result.

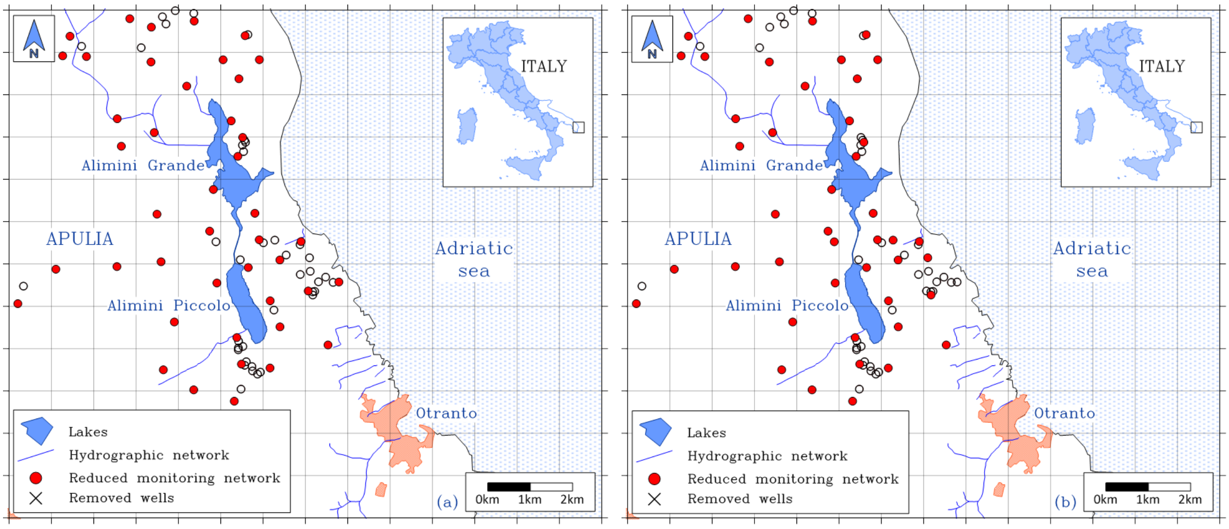

Figure 2: Downsizing results: (a) – Greedy Deletion method; (b) – Spatial Simulated Annealing method.

References

[1] Barca, E., Passarella, G., Uricchio, V., (2008). Optimal extension of the rain gauge monitoring network of the Apulian regional consortium for crop protection. Environ Monit Assess 145(1–3):375–386.

[2] Barca, E., Passarella, G., Vurro, M., & Morea, A. (2015). MSANOS: Data-driven, multi-approach software for optimal redesign of environmental monitoring networks. Water Resources Management, 29(2), 619-644.

[3] Isaaks, E.H., Srivastava, R.M. (1989). An introduction to applied geostatistics. Oxford University Press.

[4] Margiotta S., Negri S. (2005) - Geophysical and stratigraphical research into deep groundwater and intruding seawater in the mediterranean area (the Salento Peninsula, Italy). *Nat. Haz. Earth Sys. Sci.* n.5, pp 127–136.

[5] Nunes, L.M., Cunha, M.C., and Ribeiro, L. (2002). Monitoring network optimization using both spatial and temporal information. In: 3rd Int. Conf. on Decision Making in Urban & Civil Engineering, London.

[6] Wu, Y. (2004). Optimal design of a groundwater monitoring network in Daqing. Env. Geology, 45, 527–535.