Towards Automatic Fingerprinting of Groundwater Aquifers

Antonella Di Roma[1], Estrella Lucena-Sánchez[2,3], Guido Sciavicco[2], and Carmela Vaccaro[1]

[1] University of Ferrara, Dept. of Phys. and Earth Sci.
{antonella.diroma|carmela.vaccaro}@unife.it
[2] University of Ferrara, Dept. of Math. and Comp. Sci.
{guido.sciavicco|estrella.lucenasanchez}@unife.it
[3] University of Modena and Reggio Emilia, Dept of Phys., Inf., and Math.

Abstract. Geochemical fingerprinting is a rapidly expanding discipline in the earth and environmental sciences, based on the idea that geological processes leave behind physical and chemical patterns in the samples. In recent years, computational statistics and artificial intelligence methods have started to be used to help the process of geochemical fingerprinting. In this paper we consider data from 57 wells located in the province of Ferrara (Italy), all belonging to the same aquifer group and separated into 4 different aquifers. The aquifer from which each well extracts its water is known only in 18 of the 57 cases, while in other 39 cases it can be only hypothesized based on geological considerations. We devise and test a novel automatic technique for geochemical fingerprinting of groundwater by means of which we are able to identify the exact aquifer from which a sample is extracted. Our initial tests returned encouraging results.

Keywords: Geochemical fingerprinting; multi-objective optimization and feature selection.

1 Introduction

The increasing exploitation of water resources for human, industrial, and agricultural ends has brought in the last decades great attention toward the quality control of the groundwater. This attention and the complex reality of this sector has pushed the scientific community to take part in the study and the management of water resources, to improve the knowledge and to protect every realistic aspect of their management, to deal with the problems originated by the variation of volumes and intensity of precipitation due to climate change, overexploitation, salinization, anthropic pollution, degradation, and massive irrigation. Many studies have demonstrated that a mindful protection of the existing water resources could contribute to the preservation of the availability of fresh water [22, 34]. The distribution of the rains is one of the main climatic variables which has a great influence on the ground waters' turnover, on the surface runoff river flows, and on hydroelectricity production as well as on hydrogeological risk.

An hydro-geochemistry approach facilitates the understanding of the aquifer reborn, allowing to define the chemical composition of waters, and, through the application of specific models, to suspect and identify the presence of possible mixing between waters of different compositions. The quality and also the geochemical fingerprint of water bodies can be modify to interaction of a plume of polluted waters. A big data geochemical analysis allows to identify the geochemical markers and delimiting the areas of diffusion of the plume and/or the intensity of the comtamination in order to quantify the impact and the riscks. Geologists usually develop a monitoring network, and, based on the sampling provided, they build a picture of the baseline conceptual hydrogeological model of the studied area, providing a prototype monitoring for continuous data acquisition. Then, *by hand*, sometimes with the help of basic statistical tools, they try to obtain the modeling of multi-aquifer flow in order to increase the knowledge of their hydrogeological characteristic, as well as to find the geochemical fingerprint that represents a specific aquifer level. This process is very expensive and entails an elevated risk of mistake due to potential loss of information, manual loading of data, and prolonged analysis time. But recent innovations in the field of information management with intelligent systems of data acquisition for aquifer features and modeling the seasonal fluctuations of chemical and physical parameters, such as sensor systems and and big data algorithms have opened innovative scenarios for the application of computer technology to the environmental monitoring industry. Indeed, machine learning approaches are necessary to streamline the activities, from raw data to artificial intelligence applications; together with an adequate corroboration, and geochemical and geological interpretation of the obtained results, these approaches may allow to automatize activities such as geochemical fingerprint search.

In the recent literature, various statistical methods have been used in the last decades to aid the traditional geochemical investigation to understand pollution sources, possible correlation among elements, and, in some cases, the nature of the contamination [31, 4, 16]. The recent work focused on protection of groundwater against pollution, deterioration, and for input pollution identification include applying geographical information systems and decision analysis [27, 25], logistic regression model learning [21], univariate and multivariate analysis [23], and multiple regression models [10]. More in general, machine learning is emerging as an effective empirical approach for both regression and/or classification of nonlinear systems, ranging from few to thousands of variables, and they are ideal for addressing those problems where our theoretical knowledge is still incomplete but for which we do have a significant number of observations. In the past decades, it has proven useful for a very large number of applications, and among the techniques most commonly used we may mention artificial neural networks [33, 2, 18, 30], support vector machines [3], but also self-organizing map, decision trees, ensemble methods such as random forests, case-based reasoning, neuro-fuzzy networks, and evolutionary algorithms [17]. In this paper we considered 57 water wells located in the province of Ferrara, all belonging to the aquifer group A (the most superficial one), which, in turn, is separated into 4 different

aquifers, named from $A1$ to $A4$ [1]. The aquifer from which each well extracts its water is known only in 18 of the 57 cases, while in other 39 cases it can be only hypothesized based on geological considerations; the ultimate purpose of the present study it to devise an automatic, machine learning based method to identify the geochemical fingerprint of each aquifer, so that each unknown well can be assigned an aquifer, and the control network can be improved. The number of possible combinations is exponential in the number of variables, giving rise to a feature selection problem combined with a clusterization problem, which we express as an optimization problem and solve using an evolutionary algorithm. The result can be considered as an approximation to the geochemical fingerprint of each of the four aquifer, expressed in terms of *centroid*, that is, in terms of an ideal, hypothetical set of values for each aquifer of a selection of the indicators, that represents the aquifer itself. By using such a fingerprint, we were able to assign the correct aquifer to each of unknown wells, with a reasonable expected accuracy.

This paper is organized as follows. In the next section, we give the necessary background on fingerprinting, feature selection, and clustering. In Section 3 we present our data and give a very simple exploratory analysis. In Section 3 we give a short account of our data, and in Section 4 we present the mathematical formulation of our technique. Then, in Section 5 we present and discuss our results, before concluding.

2 Background

Feature selection is a machine learning technique for data preprocessing, defined as eliminating features from the data base that are irrelevant to the task to be performed [12]. In its original formulation and meaning, feature selection facilitates data understanding, reduces the storage requirements, and lowers the processing time, so that model learning becomes an easier process. Feature selection methods that do not incorporate dependencies between attributes are called *univariate* methods, and they consist in applying some criterion to each pair feature-response, and measuring the individual power of a given feature with respect to the response independently from the other features, so that each feature can be ranked accordingly. In *multivariate* methods, on the other hand, the assessment is performed for subsets of features rather than single features. There are several different approaches to feature selection in the literature. Among them, the most versatile ones are those that define the selection problem as an optimization problem. A *multi-objective optimization problem* (see, e.g. [5]) can be formally defined as the optimization problem of simultaneously minimizing (or maximizing) a set of k arbitrary functions:

$$\begin{cases} \min / \max \ f_1(\bar{x}) \\ \min / \max \ f_2(\bar{x}) \\ \dots \\ \min / \max \ f_k(\bar{x}), \end{cases} \tag{1}$$

where $\bar{x}$ is a vector of decision variables. A multi-objective optimization problem can be *continuous*, in which we look for real values, or *combinatorial*, we look for objects from a countably (in)finite set, typically integers, permutations, or graphs. Maximization and minimization problems can be reduced to each other, so that it is sufficient to consider one type only. A set $\mathcal{F}$ of solutions for a multi-objective problem is *non dominated* (or *Pareto optimal*) if and only if for each $\bar{x} \in \mathcal{F}$, there exists no $\bar{y} \in \mathcal{F}$ such that *(i)* there exists i $(1 \leq i \leq k)$ that $f_i(\bar{y})$ improves $f_i(\bar{x})$, and *(ii)* for every j, $(1 \leq j \leq k,\ j \neq i)$, $f_j(\bar{x})$ does not improve $f_i(\bar{y})$. In other words, a solution $\bar{x}$ *dominates* a solution $\bar{y}$ if and only if $\bar{x}$ is better than $\bar{y}$ in at least one objective, and it is not worse than $\bar{y}$ in the remaining objectives. We say that $\bar{x}$ is *non-dominated* if and only if there is not other solution that dominates it. The set of non dominated solutions from $\mathcal{F}$ is called *Pareto front*. Optimization problems can be approached in several ways; among them, *multi-objective evolutionary algorithms* are a popular choice (see, e.g. [14, 9, 24]). Feature selection can be seen as a multi-objective optimization problem, in which the solution encodes the selected features, and the objective(s) are designed to evaluate the performances of some model-extraction algorithm; this may entail, for example, instantiating (1) as:

$$\begin{cases} \max\ Performance(\bar{x}) \\ \min\ Cardinality(\bar{x}), \end{cases} \tag{2}$$

where $\bar{x}$ represents the chosen features; Equation 2 can be referred to as a *wrapper*.

Cluster analysis or *clustering* is the task of grouping a set of objects so that those in the same group, or *cluster* are more similar to each other than to those in other groups. The literature on cluster analysis is very wide, and includes *hierarchical* clustering, *centroid-based* models, *distribution-based* models, *density* models, among many others. Centroid-based models are of particular interest for us, because they are especially useful for numerical, many dimensional objects such as groundwater samples. The concept of centroid is essential in the most well-known centroid-based clustering algorithm, that is, *k-means* [20]: given a group of objects and a notion of distance, its *centroid* is the set of values that describes an object C (which may or may not be a concrete object of the group) such that the geometric mean of the distances between C and every other element of the group is minimal. In the k-means algorithm the groups (and even their number) is not known beforehand (this type of cluster analysis is called *exploratory*), and the algorithm is based on an initial random guessing of the centroid that eventually converges to a local optimum. KNN [6] is a distance-based *classification* algorithm, whose main idea is that close-by objects can be classified in a similar way. In this paper we use both ideas of centroid and distance-based classification in order to systematically extract geochemical fingerprints.

Finally, geochemical fingerprinting is based on the idea that geological processes leave behind physical, chemical and sometimes also isotopic patterns in

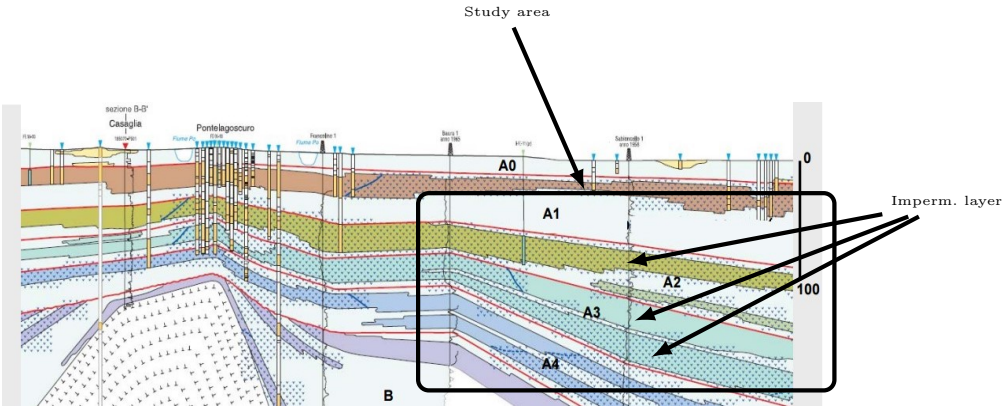

Fig. 1. The area under study, cross-section.

the samples. Many of these patterns, informally referred to as *geochemical fingerprints*, may differ only in fine details from each other. For this reason, approaching automatic fingerprinting requires highly precise and accurate data analysis [15]. Applications of geochemical fingerprinting range on a wide set of contexts, from studies on ancient artifacts such as glass or ceramics [19], to mineral identification and discovery of Jurassic-age kimberlite [29], to dust transport monitoring [11], to groundwater resources identification and study [28], to the study of the quality of food and beverages, especially wine, as shown in [26]. Because of the statistical nature of geochemical fingerprints, statistical methods are suitable for their identification. In the most recent literature, statistical methods are being progressively integrated and paired with machine learning and artificial intelligence based technology. In this paper, we develop a novel method for groundwater fingerprint identification, based on feature selection, solved as an optimization problem, and implemented via a evolutionary algorithm.

3 Data

The data used for this study consist of 910 samples extracted from 57 wells located in the province of Ferrara, all belonging to the aquifer group A (the most superficial one), from 2010 to 2017. This aquifer group is separated into 4 different aquifers, as it is shown in Figure 1, named from $A1$ to $A4$ from the most to the least superficial, which present a (probably impermeable) layer in between each pair [1]. The exact aquifer from which each well extracts its water is known only in 18 of the 57 cases, while in other 39 cases it can be only hypothesized based on geological and stratigraphic considerations. Out of the total samples, we selected those which were extracted using *single-filter* pumps (that is, that give the guarantee that the groundwater comes from one aquifer only) and of which the precise aquifer was known, reducing our data set to 229 samples. Each sample contains 13 chemical-physical indicators: η (hardness), T, $E.C.$, Na^+,

feature	*mean*	*p-value*	*kurtosis*	*skewness*
η	469.50	$7.01 * 10^{-20}$	7.18	2.15
T	15.81	$1.16 * 10^{-18}$	5.37	0.55
$E.C.$	1574.00	$1.36 * 10^{-18}$	2.45	1.08
HCO_3^-	606.50	$3.54 * 10^{-11}$	4.44	1.18
Cl^-	56.00	$2.43 * 10^{-21}$	2.81	1.25
SO_4^{2-}	18.05	$1.61 * 10^{-19}$	7.59	2.03
Ca^{2+}	462.70	$1.91 * 10^{-29}$	64.02	7.03
Mg^{2+}	381.33	$2.89 * 10^{-32}$	226.95	15.03
Na^+	183.52	$1.79 * 10^{-18}$	4.02	1.51
K^+	24.39	$1.27 * 10^{-29}$	42.85	6.13
NH_4	2135.61	$2.01 * 10^{-27}$	43.06	5.51
Fe	1501.52	$1.40 * 10^{-27}$	39.20	5.47
As	$3.23 * 10^{-3}$	$1.38 * 10^{-22}$	18.84	3.23

Table 1. Descriptive statistical analysis of the data.

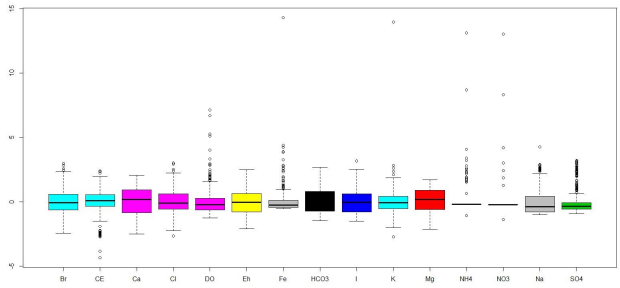

Fig. 2. Distribution statistical analysis of the data.

K^+, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, HCO_3^-, NH_4, Fe, As. Some relevant statistical measures of the different chemical elements are showed in Table 1, correlations between variables are showed in Table 2, and

As it can be observed, none of the variables follows a normal distribution (their p-values are well below 0.05), and they all present very high levels of kurtosis and skewness, being Mg^{2+} and Ca^{2+} the most evident examples. Moreover, Figure 2 shows as the data contain outliers and anomalous values. As a consequence, applying standard statistical tools such as principal component analysis [32] to identify physical-chemical fingerprints does not make much sense, as such tools usually require a certain statistical well-behaviour, for the results to be reliable [13]. This is a common problem in the analysis of real data, especially of chemical-physical origin. Approaching problems such as fingerprint identification with machine learning tools, instead, tend to bypass this obstacles, at the expenses of a more complex design.

	η	T	$E.C.$	HCO_3^-	Cl^-	SO_4^{2-}	Ca^{2+}	Mg^{2+}	Na^+	K^+	NH_4	Fe	As
η	1.00	0.13	0.76	0.78	0.62	-0.23	0.14	0.02	0.49	0.14	0.23	0.40	-0.19
T	0.13	1.00	0.05	0.06	0.02	0.20	0.01	0.03	-0.03	0.01	-0.01	0.04	-0.01
$E.C.$	0.76	0.05	1.00	0.47	0.95	-0.32	0.19	0.10	0.84	0.24	0.27	0.39	-0.20
HCO_3^-	0.78	0.06	0.47	1.00	0.24	-0.34	0.11	-0.01	0.26	0.04	0.21	0.35	-0.07
Cl^-	0.62	0.02	0.95	0.25	1.00	-0.29	0.16	0.13	0.86	0.25	0.24	0.34	-0.21
SO_4^{2-}	-0.23	0.20	-0.33	-0.34	-0.29	1.00	-0.08	-0.04	-0.36	-0.06	-0.15	-0.15	0.03
Ca^{2+}	0.14	0.01	0.19	0.11	0.16	-0.08	1.00	-0.01	-0.09	0.76	-0.06	-0.05	-0.05
Mg^{2+}	0.02	0.04	0.10	-0.02	0.13	-0.04	-0.01	1.00	0.13	-0.01	-0.02	0.03	-0.02
Na^+	0.49	-0.03	0.84	0.26	0.86	-0.36	-0.09	0.13	1.00	-0.09	0.25	0.31	-0.15
K^+	0.14	0.01	0.25	0.05	0.25	-0.06	0.77	-0.01	-0.09	1.00	-0.07	-0.06	-0.05
NH_4	0.23	-0.01	0.28	0.22	0.24	-0.15	-0.06	-0.02	0.25	-0.07	1.00	0.30	-0.01
Fe	0.41	0.05	0.40	0.35	0.34	-0.16	-0.05	0.03	0.31	-0.06	0.30	1.00	-0.02
As	-0.19	-0.01	-0.21	-0.07	-0.21	0.03	-0.05	-0.02	-0.15	-0.04	-0.01	-0.02	1.00

Table 2. Correlation matrix for our variables.

4 Method

Each instance in our data set can be seen as a vector in $\mathbb{R}^d$ (in our case, $d = 13$):

$$D = \begin{bmatrix} a_{11} & \dots & a_{1d} & A1 \\ \dots & \dots & \dots & \\ a_{m_1 1} & \dots & a_{m_1 d} & A1 \\ a_{(m_1+1)1} & \dots & a_{(m_1+1)d} & A2 \\ \dots & \dots & \dots & \\ a_{m_2 1} & \dots & a_{m_2 d} & A2 \\ a_{(m_2+1)1} & \dots & a_{(m_2+1)d} & A3 \\ \dots & \dots & \dots & \\ a_{m_3 1} & \dots & a_{m_1 d} & A3 \\ a_{(m_3+1)1} & \dots & a_{(m_3+1)d} & A4 \\ \dots & \dots & \dots & \\ a_{m_4 1} & \dots & a_{m_4 d} & A4 \end{bmatrix} \tag{3}$$

In order to evaluate the distance between two instances $I = (a_1, \dots, a_d)$ and $J = (a'_1, \dots, a'_d)$, we use the well-known notion of *Euclidean distance*:

$$dist(I, J) = \sqrt{\Sigma_{i=1}^{d}(|a_i - a'_i|^2)} \tag{4}$$

In this way we can compute the distance between any two samples of groundwater. Such a value is strongly influenced by the parameters (the specific subset of the d dimensions) that are taken into consideration. If we choose to represent the instances with a specific subset of parameters, instead of using all of them, the relative distances among different pairs of instances can vary very much. Consequently, the fingerprint extraction problem can be seen as a feature selection problem, that is, the problem of establishing the *best* subset of chemical-physical parameter. However, unlike the classical feature selection problem, selecting the

correct classification algorithm (i.e., the correct inference model) is not immediate. We choose to model the fingerprint of an aquifer as the set of values that best represent an (ideal) sample of groundwater from that aquifer, that is, its centroid. Thus, we have a *feature selection for centroid identification* problem, as it is a clusterization problem in which the clusters are already set.

Now, let $\bar{x} = (x_1, \ldots, x_d)$ a vector of solution variables, each taking values in the domain $\{0, 1\}$; as in a classical feature selection problem, each 1 means that the corresponding feature is selected, while 0 means that it is discarded; we denote by $C_j(\bar{x})$ the centroid of the j-th aquifer ($1 \leq j \leq 4$, in our case) computed using precisely the attributes that correspond to $\bar{x}$. In order to adapt (2) to our problem, we need to define how we evaluate the performances of the solution, which, in our case, means defining what classification problem we want to solve. To this end, indicating by $A(I)$ the (true) aquifer to which the instance I correspond, we compute the number of correct predictions as:

$$\#Correct(\bar{x}) = \Sigma_{I \in D} \begin{cases} 1 \text{ if } A(I) = argmin_{A_j} d(I, C_j(\bar{x})) \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

and, consequently define the *accuracy* of $\bar{x}$ over D as:

$$Acc(\bar{x}) = \frac{\#Correct(\bar{x})}{|D|} \tag{6}$$

The accuracy of a fingerprint selection can be used to reformulate our problem as an optimization problem, as it can be seen as suitable performance indicator. Minimizing the cardinality of the selected features is also correct in fingerprinting selection, as smaller fingerprints are more interpretable. In order to take into account the fact that some geochemical processes are not necessarily linear, we can slightly complicate our formulation by introducing a third objective. As a matter of fact, we can expand the domain of each solution variable x_i to take value in $\mathbb{N}$, instead of $\{0, 1\}$. While we still interpret 0 as discarding the corresponding parameter, we now interpret a positive value as the power to which the corresponding parameter is raised; we simulate, in this way, a sort of dynamic normalization of our data. It is possible to optimize the complexity of the resulting fingerprint in terms of non-linear behaviour, that is, by minimizing the maximum exponent:

$$MaxExp(\bar{x}) = \max_{i=1}^{d} x_i \tag{7}$$

So, in each execution, a vector of solutions variables $\bar{x}$ entails a transformation of the original data set (3) into:

$$D = \begin{bmatrix} {a_{11}}^{x_1} & \dots & {a_{1d}}^{x_d} & A1 \\ \dots & \dots & \dots & \\ {a_{m_1 1}}^{x_1} & \dots & {a_{m_1 d}}^{x_d} & A1 \\ a_{11} & \dots & a_{1d} & A2 \\ \dots & \dots & \dots & \\ {a_{m_2 1}}^{x_1} & \dots & {a_{m_2 d}}^{x_d} & A2 \\ a_{11} & \dots & {a_{1d}}^{x_d} & A3 \\ \dots & \dots & \dots & \\ {a_{m_3 1}}^{x_1} & \dots & {a_{m_1 d}}^{x_d} & A3 \\ {a_{11}}^{x_1} & \dots & {a_{1d}}^{x_d} & A4 \\ \dots & \dots & \dots & \\ {a_{m_4 1}}^{x_1} & \dots & {a_{m_2 d}}^{x_d} & A4 \end{bmatrix} \tag{8}$$

where, for simplicity of notation, we have not shown the case of discarded attributes.

Summing up, we can reformulate (2) for fingerprint extraction as the following optimization problem:

$$\begin{cases} \max \ Acc(\bar{x}) \\ \min \ MaxExp(\bar{x}) \\ \min \ Cardinality(\bar{x}) \end{cases} \tag{9}$$

5 Implementation and Results

Multi-objective evolutionary algorithms are known to be particularly suitable to perform multi-objective optimization, as they search for multiple optimal solutions in parallel. In this experiment we have chosen the well-known NSGA-II (Non-dominated Sorted Genetic Algorithm) [7] algorithm, which is available as open-source from the suite *jMetal* [8]. NGSA-II is an elitist Pareto-based multi-objective evolutionary algorithm that employs a strategy with a binary tournament selection and a rank-crowding better function, where the rank of an individual in a population is the non-domination level of the individual in the whole population. We used the standard parameters in each experiment, and implemented elementary variants of mutation and crossover to make them specific to our solution format. To cope with the intrinsic unbalancing of our data (over 70% of the samples belong to $A1$), we operated a re-sampling, to obtain a training set with 10 samples per each aquifer ($D_{training}$), and left every other sample for test (D_{test}). Test was performed in the natural way, that is, by applying the accuracy function(s) to D_{test} using the centroid and the selected attributes extracted from the chosen solution. We have executed 10 runs of the model (9), each with a different seed; the population side was 100 in each experiment, and we set each experiment for 100 generations each. A multi-objective optimization problem gives rise to a Pareto front, that is, to a *last* population of (non-dominated) individuals from which one or more individuals can be selected via a decision-making process. The standard approach to decision making

fingerprint	*acc*	*recall*			
		A1	*A2*	*A3*	*A4*
$(\eta)^2$,HCO_3^-, $(NH_4)^3$	0.59	0.55	0.71	1.00	1.00
$(\eta)^3$, HCO_3^-, $(Fe)^3$	0.60	0.54	0.81	0.90	1.00
$(\eta)^3$, $(HCO_3^-)^3$, $(Fe)^3$	0.60	0.53	0.81	1.00	1.00
$(\eta)^2$, $(HCO_3^-)^3$, $(NH_4)^2$, Fe	0.60	0.55	0.71	1.00	1.00
$(T)^2$, $(HCO_3^-)^2$, SO_4^{2-}, NH_4	0.54	0.47	0.81	1.00	1.00
$(\eta)^3$, $(HCO_3^-)^3$, $(Fe)^3$	0.60	0.47	0.81	1.00	1.00
$(\eta)^3$, $(HCO_3^-)^3$, $(Fe)^3$	0.60	0.53	0.81	1.00	1.00
$(T)^2$, η, $(HCO_3^-)^2$	0.55	0.47	0.81	1.00	1.00
$(\eta)^3$, HCO_3^-, $(Fe)^3$	0.60	0.55	0.71	1.00	1.00
$(T)^2$, η, HCO_3^-	0.55	0.47	0.81	1.00	1.00

Table 3. Test results.

is selecting the solution with the best value in the most important among the objectives; in our case that would be the accuracy. Unfortunately, this strategy gives rise to fingerprints with too many characteristics, which would be too difficult to interpret. Therefore, our decision-making strategy is to select the most accurate solution with strictly less than six chosen characteristics.

The set of chosen results is shown in Table 3. As it can be seen, we reach a level of accuracy between 0.55 and 0.60; taking into account that we have a four-classes problem, we may consider it acceptable. Moreover, the recall level (i.e., the ratio of correct answers) per class, shows that, in general, our fingerprints are able to identify three out of four aquifers in a very precise way. Finally, it can be observed how while different executions have produced different fingerprints, they share many elements, indicating that our approach is stable.

6 Conclusions

In this paper we have considered the results of the geochemical analysis of groundwater samples from 57 water wells located in the province of Ferrara, all belonging to the same aquifer group. We considered the problem of identifying the geochemical fingerprint of each aquifer of the group, so that those wells that extract water from the same group but from an unknown aquifer can be safely assigned one without making decisions based on the depth of the well itself. We proved that our method, based on an artificial intelligence technique which we called feature selection for centroid identification, returns fingerprints with a sufficiently high level of accuracy.

Our method can be improved in many ways. Future directions include considering different optimization functions (e.g., building a *fault-tolerant* accuracy

that considers neighbour aquifers as less severe mistake, or optimizing the complexity of the resulting fingerprint in different ways), and using *ratios* among affine elements/characteristics instead of, or in conjunction with, the original variables. Moreover, if we take into account the temporal component of the data, this problem can be seen as a multivariate temporal series classification problem, so to define a fingerprint as a *temporal pattern* instead of a set of values. By doing so, one can set up a completely different set of experiments, to establish by how much, and in which way, fingerprints change over time.

References

1. Amorosi, A., Bruno, L., Rossi, V., Severi, P., Hajdas, I.: Paleosol architecture of a late quaternary basin–margin sequence and its implications for high-resolution, non-marine sequence stratigraphy. Global and Planetary Change **112**, 12–25 (2014)
2. Atkinson, P., Tatnall, A.: Introduction: neural networks in remote sensing. International Journal of Remote Sensing **4**(18), 699–709 (1997)
3. Azamathulla, H., Wu, F.: Support vector machine approach for longitudinal dispersion coefficients in natural streams. Applied Soft Computing **2**(11), 2902–2905 (2011)
4. Belkhiri, L., Mouni, L., Narany, T.S., Tiri, A.: Evaluation of potential health risk of heavy metals in groundwater using the integration of indicator kriging and multivariate statistical methods. Groundwater for Sustainable Development **4**, 12 – 22 (2017)
5. Collette, Y., Siarry, P.: Multiobjective Optimization: Principles and Case Studies. Springer Berlin Heidelberg (2004)
6. Dasarathy, B.: Nearest Neighbour (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press (1991)
7. Deb, K.: Multi-objective optimization using evolutionary algorithms. Wiley, London, UK (2001)
8. Durillo, J., Nebro, A.: Jmetal: a Java framework for multi-objective optimization. Avances in Engineering Software **42**, 760 – 771 (2011)
9. Emmanouilidis, C., Hunter, A., Macintyre, J., Cox, C.: A multi-objective genetic algorithm approach to feature selection in neural and fuzzy modeling. Evolutionary Optimization **3**(1), 1–26 (2001)
10. Farhadian, H., Katibeh, H.: New empirical model to evaluate groundwater flow into circular tunnel using multiple regression analysis. International Journal of Mining Science and Technology **27**(3), 415 – 421 (2017)
11. Galleta, S., Jahn, B., Lanoë, B.V.V., Dia, A., Rossello, E.: Loess geochemistry and its implications for particle origin and composition of the upper continental crust. Earth Planet Science Letters pp. 157–172 (1989)
12. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research (3), 1157–1182 (2003)
13. Jiang, B., Pei, J.: Outlier detection on uncertain data: Objects, instances, and inferences. In: Proc. of the 27th International Conference on Data Engineering. pp. 422–433 (2011)
14. Jiménez, F., Sánchez, G., García, J., Sciavicco, G., Miralles, L.: Multi-objective evolutionary feature selection for online sales forecasting. Neurocomputing **234**, 75–92 (2017)

15. Kamber, B.: Geochemical fingerprinting: 40 years of analytical development and real world applications. Applied Geochemistry **24**(6), 1074 – 1086 (2009)
16. Kozyatnyk, I., Lövgren, L., Tysklind, M., Haglund, P.: Multivariate assessment of barriers materials for treatment of complex groundwater rich in dissolved organic matter and organic and inorganic contaminants. Journal of Environmental Chemical Engineering **5**(4), 3075 – 3082 (2017)
17. Lary, D., Alavi, A., Gandomi, A., Walker, A.: Machine learning in geosciences and remote sensing. Geoscience Frontiers **7**(1), 3 – 10 (2016)
18. Lary, D., Muller, M., Mussa, H.: Using neural networks to describe tracer correlations. Atmospheric Chemistry and Physics (4), 143–146 (2004)
19. Li, B., Greig, A., Zhao, J., Collerson, K., Quan, K., Meng, Y., Ma, Z.: Icp-ms trace element analysis of song dynasty porcelains from Ding, Jiexiu and Guantai kilns, north China. Journal of Archaeological Sciences (32), 251–259 (2005)
20. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability. pp. 281 — 297 (1967)
21. Mair, A., El-Kadi, A.: Logistic regression modeling to assess groundwater vulnerability to contamination in Hawaii, USA. Journal of Contaminant Hydrology **153**, 1 – 23 (2013)
22. Martinelli, G., Minissale, A., Verrucchi, C.: Geochemistry of heavily exploited aquifers in the Emilia-Romagna region (Po Valley, Northern Italy). Environmental Geology **4-4**(36), 195–206 (1998)
23. Menció, A., Mas-Pla, J., Otero, N., Regàs, O., Boy-Roura, M., Puig, R., Bach, J., Domènech, C., Zamorano, M., Brusi, D., Folch, A.: Nitrate pollution of groundwater; all right..., but nothing else? Science of The Total Environment **539**, 241 – 251 (2016)
24. Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., Coello, C.C.: A survey of multiobjective evolutionary algorithms for data mining: Part I. IEEE Transactions on Evolutionary Computation **18**(1), 4–19 (2014)
25. Ozdemir, A.: Gis-based groundwater spring potential mapping in the Sultan Mountains (Konya, Turkey) using frequency ratio, weights of evidence and logistic regression methods and their comparison. Journal of Hydrology **411**(3), 290 – 308 (2011)
26. Pepi, S., Vaccaro, C.: Geochemical fingerprints of Prosecco wine based on major and trace elements. Environmental Geochemestry and Health **2**(40), 833–847 (2018)
27. Pizzol, L., Zabeo, A., Critto, A., Giubilato, E., Marcomini, A.: Risk-based prioritization methodology for the classification of groundwater pollution sources. Science of The Total Environment **506**, 505 – 517 (2015)
28. Ranjbar, A., Mahjouri, N., Cherubini, C.: Development of an efficient conjunctive meta-model-based decision-making framework for saltwater intrusion management in coastal aquifers. Journal of Hydro-environment Research (2019)
29. Ross, J., Jaques, A., Perguson, J., Green, D., O'Reilly, S., Danchin, R., Janse, A.: Sodium in garnet and potassium in clinopyroxene: criteria for classifying mantle eclogites. Kimberlites and Related Rocks pp. 27–832 (1989)
30. Shahin, M., Jaksa, M., Maier, H.: Artificial neural network applications in geotechnical engineering. Australian Geomechanics **1**(36), 49–62 (2001)
31. Singh, C.K., Kumar, A., Shashtri, S., Kumar, A., Kumar, P., Mallick, J.: Multivariate statistical analysis and geochemical modeling for geochemical assessment of groundwater of Delhi, India. Journal of Geochemical Exploration **175**, 59 – 71 (2017)

32. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems **2**(1-3), 37–52 (1987)
33. Yi, J., Prybutok, V.: A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environmental Pollution **3**(92), 349–357 (1996)
34. Zuppi, G., Sacchi, E.: Hydrogeology as a climate recorder: Sahara–Sahel (North Africa) and the po plain (Northern Italy). Global and Planetary Change **40**, 79–91 (2004)