On Intelligent Fingerprinting of Antique Buildings from Clay Composition

E. Marrocchino[3], G. Sciavicco[1], E. Lucena-Sánchez[1,2], and C. Vaccaro[3]

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

Abstract. Materials research in archaeological ceramic artifacts is a consolidated practice that helps architectural heritage preservation. Ancient buildings located within the historic centres, in particular, mark the image and history of each city at different periods, and when damaged historical masonry needs restoration actions, a good characterization of both new and old material is crucial to forecast the behaviour of the system. In this paper we consider 9 antique buildings and constructions of the Medieval city of Ferrara (NE Italy), and the geochemical characterization of samples from their bricks and *terracottas*. We apply an intelligent fingerprinting technique to major and trace elements, along with some of their ratios, with the purpose of identifying the smallest and more accurate subsets that allow to uniquely identify one building from all others. We obtain very encouraging results, with accuracies over 80% even with very small fingerprints.

Keywords: Feature selection · Fingerprinting · Material research.

1 Introduction

Materials research applied to the study of archaeological ceramic artifacts is now a consolidated practice, and during the last decades the interest of scientists, architects, engineers, and archaeologists towards architectural heritage preservation has risen. The ancient buildings within the historic centres mark the image and history of each city at different periods, and when damaged historical masonry needs restoration actions, a good characterization of both new and old material is crucial to forecast the behaviour of the system. Historical understanding is not only useful to analyse and preserve objects but also to investigate the knowledge and skills used to produce and use them. In this light, the main goals of the building materials characterization are preservation and restoration, which include [26]: *(i)* origin of historical raw materials, *(ii)* processes and changes in archaeological artifacts, *(iii)* determination of original firing temperature, and *(iv)* reconstruction of firing techniques and manufacturing technologies. Bricks and ceramics can be considered artificial rocks fired in kilns and, under this

point of view, mineralogical, petrological and geochemical study approaches can be useful tools for the study of archaeological ceramic materials. Bricks in ancient buildings preserve the trace of the claystone geological formation used to create them and geology influences raw material availability and thus building methods.

The Medieval city of Ferrara (NE Italy) is located in the eastern part of the Po alluvial plain and its geographical position, situated on a major waterway at a natural crossroads between the Adriatic Sea and the Po alluvial plain, contributed to the flourishing of the city, which peaked during the Renaissance period. Ferrara reached the top of Renaissance architecture and prestige with the dominion of the Estense family [10, 29, 30, 36]. The Estense family considered the urban layout inside the first defensive walls inappropriate to represent their greatness and their dominance over the territory. This was the reason they decided to radically modify the city's layout, changing its structure and appearance with a further fortification of the wall ring and embellish of the city centre with numerous sumptuous palaces. These buildings utilized bricks (*cotto ferrarese*) and mortars as the dominant building materials. The use of these construction materials is related to the geographical position and the geomorphological framework of the Ferrara area, which is characterized by the widespread presence of silico-clastic sediments [1]. Historical bricks and terracotta decorations from some important Medieval and Renaissance buildings of Ferrara were collected and studied [7, 28]. The chemical characterization of their bricks and *terracotta* elements were used to evaluate the nature of the original raw material (i.e. clay-rich sediment of local provenance). As a result of that study it was shown that characterizing the old construction materials used, as well as establishing the causes of the decay (when possible), could be useful for the planning of suitable restoration treatment and to identify and characterize the materials that were used to make these architectural elements [38, 5]. This was motivated by the need of giving the restorer adequate information for choosing suitable new materials when replacement is necessary and to avoid incorrect restoration materials. The data obtained defined some technological information regarding the manufacture and some information regarding the different provenance of the raw materials. Such results were compared with the chemical-mineralogical data available for the sediments of the area [6, 8] in order to define the nature and provenance of the original raw materials. In this paper, instead, we use them to establish the *geochemical fingerprint* of these building, effectively proving that it is possible to distinguish one building from another by the composition of its clay. Fingerprinting is a well-known problem in geology [18], and geochemical fingerprints for the classification of natural environments are being used more and more often, as (semi-)automatic, artificial intelligence based techniques improve in accuracy and diffusion. Chemical and geological fingerprinting is applied across many fields, but it is particularly prominent in forensic studies, mineral exploration, and in tracking the source of environmental pollutants. Other applications include petroleum science, geology, botany, food and agricultural science. Geochemical fingerprints are used in different ways: to differentiate among differ-

ent mantle reservoirs, to characterize magmatic rocks from different geotectonic settings, to distinguish different water types, among others. In this particular case, we fingerprinting antique buildings helps in their restoration, conservation, and historical studying and placement.

In this paper, we define the geochemical fingerprinting problem as a *feature selection for classification* problem, and we solve it via an optimization problem, similarly to [12]. In particular, we: *(i)* study the basic statistical characterization of major and trace elements in bricks and *terracottas* of the buildings; *(ii)* define an optimization problem that allows us to establish which of them can be used to uniquely identify one building from another; and *(iii)* solve such an optimization problem via an evolutionary algorithm.

2 Background

Feature selection. *Feature selection* is a machine learning technique for data preprocessing, usually employed to eliminate those features, or attributes, that are considered irrelevant or noisy for the problem at hand [17]. On the one side, feature selection is meant to be used in order to improve the interpretability of the data, or sometimes to reduce the space requirement to store them. But feature selection is itself a learning process: the selected features provide a valuable information on the data themselves, as the fact that they have been selected implies their importance in the process being analyzed. This is particularly true in fingerprinting, that is, in essence, a feature selection process in which we want to understand which features do form a fingerprint, and how good such fingerprint performs. In some cases, feature selection models consider the various features *individually* during the selection process; these 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. In *multivariate* methods, on the other hand, the choice is performed *by subset* of features rather than single features. Among the several different approaches to feature selection, the most versatile ones are those that define the selection problem as an *optimization* problem. A *multi-objective optimization problem* (see, e.g. [9]) 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*, in which 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. [20, 14, 32]). 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*.

Fingerprinting. Physical and chemical *fingerprinting* is a well-known problem in geology and other affine disciplines. The most accepted definition refers to a geo-chemical fingerprint as a *signal that provides information about the origin, the formation and/or the environment of a geological sample* [18]. Geochemical fingerprints for the classification of natural environments are also characterized by the fact that their original composition is usually preserved, or in case of small changes during later geological history, must preserve its main geochemical characteristics to an extent that the original chemical signature is still recognizable. The use of geochemical fingerprints has a long tradition in Earth sciences, and a necessary prerequisite for applying geochemical fingerprints is that suitable analytical methods exist which allow the detection of fingerprints. The typical *by hand* process for fingerprinting is very expensive and entails an elevated risk of mistake due to potential loss of information, manual loading of data, and prolonged analysis time. In recent times, several statistical methods have been applied to aid the traditional geochemical investigation to understand, for example, pollution sources via fingerprinting, possible correlation among elements, and, in some cases, the nature of the contamination [4, 21, 40]. Examples of fingerprinting include recent works focused on protection of groundwater against pollution, deterioration, and for input pollution identification include applying geographical information systems and decision analysis [33, 34], logistic regression model learning [27], univariate and multivariate analysis [31], and multiple regression models [15]. More in general, machine learning is emerging as an effective, less complicated and less expensive [19], 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, such as in fingerprinting analysis. 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 [2, 24, 39, 42], 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 [23].

3 Intelligent Fingerprinting

Classification. In this paper, we approach the fingerprinting problem of antique buildings by using an intelligent, machine learning-based novel methodology, based on defining the fingerprinting problem as a *feature selection for classification* problem. In data mining *classification* is an essential task for assigning the class values of new instances. Fingerprinting can be seen as a classification problem because we can interpret the existence of a model for classification as a proof that the identification is possible. As we shall see, each building will be represented by a set of samples, each of which, in turn, is described by a set of values of chemical elements and compounds. We interpret every sample as an instance, classified with the (unique) identification of the building from which it was taken. A classification model of such a dataset is an implicit fingerprinting method that, as we have explained, proves that those values contain enough information to distinguish one building from the others.

Feature selection and evaluation. Three problems must be addressed as this point. The first one concerns the evaluation method. Classification can be evaluated using different classical *metrics* (such as, for example, *accuracy*, *accuracy by class*, *area under the ROC curve*, and so on) combined with different *evaluation methods*, whose statistical value depends on the numerosity of dataset and the type of pre-processing. In our case, as we have explained, obtaining clay samples from antique buildings is an expensive and time-consuming process, resulting in having relatively small datasets. Small datasets are difficult to evaluate, being the most suitable method the so-called *leave-one-out cross validation* ($LOOCV$), which is a variant of the more general *cross validation* (CV). In cross validation, we randomly divide a dataset into k disjoint folds with approximately equal size, and each fold is in turn used to test the model induced from the other $k-1$ folds by a classification algorithm. The performance of the classification algorithm is evaluated by averaging the performances of the k models, and hence the level of averaging is assumed to be at fold. The leave-one-out version is used to estimate the performance of a classification algorithm by setting k to be the number of examples in the dataset, making it the best choice in the case of small datasets, and is able to produce robust enough estimates of model performance as each instance is given an opportunity to represent the entirety of the test dataset. The second problem concerns the dimension of the fingerprints. Classification per se extracts models that use all attributes, that is, the entire chemical spectrum of samples. On the one side, good performances of a classification model in our problem proves that fingerprinting is possible, which is not obvious. On

the other hand, useful fingerprints should also be small, so that they can be graphically visualized, discussed, and evaluated within the current theory. For this reason, intelligent fingerprinting of building requires minimizing the number of attributes. Finally, it so happens that the chemical interactions that describe a particular clay may not be all linear. In other words, good fingerprint may depend on *ratios* of elements instead of elements themselves; this should be taken into account in data preparation.

A model for building fingerprinting. Summing up, our strategy for building fingerprinting is to use an evolutionary algorithm to solve the following optimization problem:

$$\begin{cases} \max \ Performance - LOOCV(\bar{x}) \\ \min \ Cardinality(\bar{x}), \end{cases} \tag{3}$$

where $\bar{x}$ represents the set of possible chemical values or ratios between chemical values that describe the instances.

4 Data

Origin. The data used for this study consist of 113 samples of clays and bricks from 9 different antique palaces located in the province of Ferrara. These included the Monastery of Sant'Antonio in Polesine (built in several phases during the 12th -16th centuries), the Church of Santa Maria in Vado (founded in the 10th century and extensively modified in the 15th-16th centuries), the Church of Santo Stefano (founded in the 10th century, but rebuilt in the 15th -16th centuries), the Cathedral of Ferrara (apse, 15th-16th centuries), the Schifanoia Palace ('Hall of Stuccoes', 15th-16th centuries), the surrounding city walls (15th century), the Palazzo Roverella (built in the 16th century), the Monastery of Santa Maria delle Grazie (built in several phases starting from the 14th century) and the church of San Andrea (built around the year 1000, in the east part of the Byzantine Castrum, but nowadays of the ancient splendor only the aisle with small semicircular apses and a small portion of the apse remain). Geochemical characterisation of their *terracottas* was carried out through X-ray fluorescence (XRF). The preparation of all the samples and the following analyses were carried out in the laboratories of the Department of Physics and Earth Sciences of the University of Ferrara. All the samples were pulverized using an agate pestle until a powder with a particle size less than 2 μm was obtained; part of the powder of each sample was used for loss on ignition (LOI) calculation where the powders were first dried in an oven at 110 $^{\circ}C$, and then placed in the ceramic crucibles and subjected to a temperature of 1000 $^{\circ}C$ for one day. After that, 0.5 g of each powder was prepared by pressing a tablet on boric acid support for obtaining the chemical composition, determined by XRF with a wavelength dispersion spectrometer ARL Advant-XP (Thermo Fisher Scientific, Waltham, MA, USA) and consisted of an X-ray tube with a Mo target and an SSD Peltier-cooled detector (10 mm^2 active area and resolution of $<$155 eV at 10 $kcps$)

according to [16, 35]. The maximum voltage and current of 50 kV and 1500 μA, respectively, were used to excite the secondary fluorescence X-rays. A collimator with a diameter of 1 mm was used to collect the emitted secondary X-rays from a surface area of about 0.79 mm^2 in air. This allowed the major elements to be determined, expressed as a percentage by oxide weight: SiO_2, TiO_2, Al_2O_3, Fe_2O_3, MnO, MgO, CaO, Na_2O, K_2O, P_2O_5. Moreover, it allowed to extract trace elements, reported in *ppm*: V, Cr, Ni, Cu, Zn, Ga, Rb, Sr, Zr, Ba, La, Ce, Pb, Sc, Co, Th, Y, Nd, S. The accuracy of the instrument, estimated on the basis of the results obtained on international standards of geological samples, and the precision, expressed as standard deviation of replicated analyses, were between 2% and 5% for the major elements, and between 5% and 10% for the trace elements. The detection limit (0.01% for major oxides) was estimated to be close to 10 *ppm* for most trace elements, except for S for which 50 *ppm* was considered. The processing of the acquired intensities and the correction of the matrix effect were performed according to the model proposed in [22, 37, 25].

Basic statistical analysis. In the context of fingerprinting via feature selection for classification, the relevant elements and the relevant ratios among elements can be considered as variables. Some relevant statistical measures of our variables are shown in Table 1. They can be classified by their behaviour in three major groups according to a Shapiro normality test: *(i) normal* ones (that follow a normal distribution), that is, SiO_2, P_2O_5, LOI, MgO, Zn, Ni, La, MgO/Cr, SiO_2/TiO_2, SiO_2/Al_2O_3, SiO_2/CaO, SiO_2/Na_2O, SiO_2/K_2O and Cr/Ni; *(ii) quasi normal* ones (that do not follow a normal distribution, but with low values of skewness and kurtosis), that is: Cr, $Y$and V; *(iii) non-normal* ones: K_2O, TiO_2, Al_2O_3, Fe_2O_3, MnO, CaO, Na_2O, K_2O, Pb, Co, Th, Zr, Rb, Sr, Ba, and Nb. As always, such an analysis highlights that more often than not real data are difficult to treat with standard statistical tools, such as PCA, justifying the application of frequency-based approaches such as the one used here.

5 Experiments and Results

Parametrization. To instantiate (2), we used three different classifiers, in order to compare which type of classification shows the best performances for this particular problem. In particular we used a *single decision tree* (J48), a *random forest* (RF), and a *logistic regression algorithm* (LR), all from the implementation of the Weka package [41], and all with their standard parametrization. Because our purpose is fingerprinting, the choice of the particular learning algorithm used in the wrapper is not very important. Decision trees are one example of interpretable learning method; random forest can be considered quasi-interpretable (as they are set of trees, but the decision are taken functionally); logistic regression is an example of functional, non-interpretable method. Our purpose is to prove that in this case fingerprinting can be solved regardless the characteristics of the underlying learning algorithm; using their standard parameters, as provided in the package, allows the experiments to be reproduced. The performances

Feature	Mean	P-value	Kurtosis	Skewness
SiO_2	52.27	0.023	$4.08 * 10^{+14}$	-0.32
TiO_2	0.68	0.576	$2.68 * 10^{+14}$	-0.11
Al_2O_3	14.34	0.749	$2.62 * 10^{+13}$	-0.09
Fe_2O_3	5.93	0.636	$2.71 * 10^{+14}$	0.17
MnO	0.12	0.105	$2.71 * 10^{+14}$	0.04
MgO	4.45	$0.921 * 10^{-3}$	$3.81 * 10^{+14}$	-0.73
CaO	10.80	0.555	$3.70 * 10^{+13}$	-0.02
Na_2O	1.47	1.349	16.46	$3.14 * 10^{+14}$
K_2O	2.51	9.433	24.34	$3.51 * 10^{+14}$
P_2O_5	0.24	0.003	3.33	0.71
LOI	6.96	0.013	2.25	0.29
Pb	43.47	0.729	17.60	$3.36 * 10^{+12}$
Zn	114.45	$1.273 * 10^{-7}$	75.39	$8.02 * 10^{+14}$
Ni	146.47	$0.564 * 10^{-3}$	4.21	-0.70
Co	19.49	0.212	3.00	-0.25
Cr	205.36	0.0428	3.34	-0.55
V	102.53	0.001	5.89	0.88
Th	9.84	0.322	3.15	-0.23
Nb	8.49	1.237	-	-
Zr	134.68	0.023	2.79	-0.46
Rb	118.15	0.576	52.23	$5.62 * 10^{+14}$
Sr	297.80	0.749	3.87	-0.18
Ba	417.82	0.636	25.51	$3.10 * 10^{+14}$
Y	25.34	0.105	-	-
La	31.22	$0.921 * 10^{-3}$	3.18	0.63
Ce	52.26	0.555	4.33	0.58

Table 1. Descriptive statistical analysis of the data.

of each classification model have been measured in LOOCV mode (as explained above), and models are all multi-classes (one class per building). In terms of optimization problem, we used *accuracy* to instantiate $Performance(\bar{x})$. To solve the optimization problem we used the evolutionary algorithm NSGA-II (Non-dominated Sorted Genetic Algorithm) [11], which is available as open-source from the suite *jMetal* [13]. 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 used a simple binary representation (zero means that the feature is discarded and one that the feature is selected) of each solution with standard mutation and crossover. We used an initial population of 10 individuals, each evolving 1000 times; each experiment has been run 10 times and the best absolute solution has been considered, in terms of accuracy of the classification model.

Classifier	Name	Fingerprint	Acc	ROC
J48	F1	Fe_2O_3, P_2O_5	0.75	0.75
	F2	MgO, Na_2O, P_2O_5, Nb, Zr	0.80	0.79
	F3	TiO_2, Na_2O, P_2O_5, Zr, La	0.78	0.81
	F4	Na_2O, K_2O, Nb, Ce	0.78	0.82
	F5	Na_2O, Nb, Rb	0.78	0.85
RF	F6	P_2O_5, Cr, Nb	0.81	0.91
	F7	Pb, Nb	0.67	0.80
	F8	Na_2O, Nb	0.72	0.84
	F9	Na_2O, P_2O_5	0.71	0.85
	F10	P_2O_5, Nb	0.74	0.90
LR	F11	Al_2O_3, Nb, Ba	0.78	0.89
	F12	P_2O_5, Zr	0.74	0.85
	F13	Ni, Nb, La, Ce	0.81	0.91
	F14	Al_2O_3, Co, Nb	0.73	0.86
	F15	P_2O_5, Ce	0.65	0.80

Table 2. Results of the shortest fingerprinting.

Results. In Table 2 we show the results of the experiments. In particular, we show five selected elements from the Pareto fronts of each group of experiments; we selected them as the most accurate fingerprints per each group with less than six features each. This additional constraint allows us to deal with fingerprints that can be interpreted, yet being still accurate enough. From Table 2 we can initially conclude that the fingerprinting problem *per se* can be solved, as the average accuracies in leave one out cross-validation mode range from 65% to 81%. More specifically, though, we can also observe that interpretable and quasi-interpretable models seem to work slightly better than non-interpretable ones, in this case, considering that, across all selected fingerprints, their average accuracy is slightly higher. Moreover, we can analyze the behaviour of fingerprints for specific buildings in Table 3. We find that some buildings are more recognizable than others. This is the case for example of Palazzo Roverella and Santa Maria delle Graze, which can be identified with accuracies over 90%. In the same way, some buildings such as Santa Maria in Vado present a more complex situation, and only the classifier based random forest is able to reach (an unimpressive) 66%. This proves that while our technique is able to solve the problem, there are situations that deserve more attention; possible strategies include the use of ratios of elements, instead of simple elements. Finally, our approach can be completed with a graphical account of selected solutions. We show two examples in Fig. 1 and 2, in which we have displayed only the buildings that separate best from each other. As it can be observed, separating lines are highly non-linear: this can be taken as an empirical proof that non-automatized approaches to this problem are unfeasible, as fingerprints cannot be easily noticed by visual examination of the data.

N°	F1				F2				F3				F4				F5				
Classifier	acc	sens	spec	roc	acc	sens	spec	roc	acc	sens	spec	roc	acc	sens	spec	roc	acc	sens	spec	roc	Class
J48	0.70	0.45	0.95	0.65	0.69	0.45	0.92	0.63	0.64	0.36	0.91	0.67	0.80	0.63	0.96	0.88	0.59	0.27	0.91	0.80	Polesine
	0.48	0.00	0.96	0.49	0.65	0.33	0.97	0.58	0.76	0.55	0.97	0.70	0.74	0.55	0.93	0.86	0.59	0.22	0.96	0.79	MariaInVado
	0.64	0.28	1.00	0.75	0.85	0.71	0.98	0.89	0.63	0.28	0.97	0.60	0.63	0.28	0.97	0.61	0.93	0.85	1.00	0.91	Stefano
	0.76	0.54	0.98	0.77	0.64	0.36	0.91	0.67	0.70	0.45	0.94	0.75	0.71	0.45	0.97	0.80	0.76	0.54	0.97	0.83	Schifanoia
	0.87	0.85	0.88	0.82	0.86	0.81	0.91	0.79	0.84	0.76	0.91	0.83	0.70	0.43	0.96	0.84	0.77	0.57	0.97	0.81	Duomo
	0.77	0.71	0.82	0.75	0.91	0.85	0.96	0.89	0.86	0.78	0.92	0.83	0.86	0.86	0.86	0.84	0.88	0.86	0.89	0.85	Pietro
	0.99	0.75	0.99	0.85	0.97	0.62	0.95	0.77	1.00	1.00	0.99	1.00	0.98	0.87	0.97	0.92	0.97	0.75	0.94	0.96	Roverella
	0.92	0.87	0.97	0.90	0.94	0.87	1.00	0.94	0.93	0.87	0.99	0.94	1.00	1.00	0.99	0.99	0.79	0.62	0.96	0.98	Grazie
	0.49	0.00	0.98	0.98	0.50	0.00	1.00	0.71	0.50	0.00	1.00	1.00	0.50	0.00	1.00	0.75	0.50	0.00	0.99	0.86	Andrea
Name	**F6**				**F7**				**F8**				**F9**				**F10**				
Classifier	acc	sens	spec	roc	acc	sens	spec	roc	acc	sens	spec	roc	acc	sens	spec	roc	acc	sens	spec	roc	**Class**
RF	0.69	0.45	0.93	0.86	0.67	0.36	0.97	0.82	0.56	0.18	0.94	0.69	0.60	0.27	0.92	0.80	0.79	0.63	0.95	0.90	Polesine
	0.66	0.33	0.98	0.75	0.60	0.22	0.98	0.74	0.63	0.33	0.93	0.85	0.63	0.33	0.93	0.67	0.61	0.22	0.99	0.70	MariaInVado
	0.79	0.57	1.00	0.98	0.71	0.42	0.99	0.97	0.77	0.57	0.97	0.87	0.63	0.28	0.97	0.88	0.85	0.71	0.98	0.99	Stefano
	0.71	0.45	0.97	0.85	0.76	0.54	0.97	0.90	0.58	0.18	0.97	0.74	0.71	0.45	0.96	0.81	0.66	0.36	0.96	0.88	Schifanoia
	0.85	0.86	0.85	0.89	0.68	0.71	0.65	0.61	0.71	0.67	0.75	0.72	0.67	0.47	0.87	0.84	0.84	0.86	0.82	0.86	Duomo
	0.92	0.89	0.95	0.98	0.78	0.64	0.92	0.91	0.82	0.68	0.95	0.95	0.74	0.60	0.87	0.97	0.83	0.71	0.94	0.95	Pietro
	0.99	0.62	0.98	0.96	0.97	0.25	0.96	0.78	0.98	0.25	0.97	0.75	0.98	0.87	0.97	0.91	0.93	0.12	0.92	0.92	Roverella
	0.87	0.75	0.98	0.89	0.55	0.12	0.97	0.84	0.80	0.62	0.97	0.96	0.80	0.62	0.97	0.93	0.61	0.25	0.96	0.88	Grazie
	0.50	0.00	0.99	0.91	0.50	0.00	0.99	0.70	0.50	0.00	0.99	0.98	1.00	1.00	1.00	0.42	0.75	0.50	0.99	0.99	Andrea
	F11				**F12**				**F13**				**F14**				**F15**				
Classifier	acc	sens	spec	roc	acc	sens	spec	roc	acc	sens	spec	roc	acc	sens	spec	roc	acc	sens	spec	roc	**Class**
LR	0.65	0.36	0.94	0.86	0.78	0.63	0.93	0.91	0.73	0.54	0.91	0.86	0.56	0.18	0.94	0.86	0.67	0.45	0.88	0.86	Polesine
	0.58	0.22	0.93	0.73	0.50	0.00	0.99	0.62	0.48	0.00	0.95	0.75	0.52	0.11	0.92	0.69	0.50	0.00	1.00	0.44	MariaInVado
	0.91	0.86	0.97	0.84	0.71	0.43	0.99	0.93	0.78	0.57	0.98	0.73	0.71	0.43	0.99	0.83	0.70	0.43	0.97	0.89	Stefano
	0.73	0.54	0.91	0.93	0.72	0.54	0.89	0.84	0.66	0.36	0.96	0.87	0.77	0.64	0.90	0.93	0.48	0.00	0.96	0.77	Schifanoia
	0.80	0.71	0.89	0.85	0.87	0.81	0.93	0.94	0.90	0.86	0.93	0.87	0.72	0.62	0.81	0.87	0.68	0.52	0.84	0.79	Duomo
	0.98	1.00	0.96	0.93	0.86	0.86	0.87	0.88	0.93	0.89	0.96	0.91	0.96	0.96	0.96	0.94	0.73	0.68	0.79	0.83	Pietro
	0.98	0.00	0.98	0.82	0.95	0.25	0.94	0.85	0.99	0.75	0.98	0.97	1.00	0.00	1.00	0.71	0.96	0.50	0.94	0.93	Roverella
	0.80	0.62	0.98	0.96	0.73	0.50	0.96	0.88	0.93	0.87	0.98	0.98	0.54	0.12	0.95	0.96	0.73	0.50	0.95	0.91	Grazie
	0.50	0.00	1.00	0.41	0.50	0.00	1.00	0.44	0.74	0.50	0.98	0.98	0.75	0.50	0.99	0.50	0.49	0.00	0.98	0.71	Andrea

Table 3. Results of the shortest fingerprinting with accuracy by class.

6 Conclusions

Materials research applied to the study of archaeological ceramic artifacts is a consolidated practice that raised the interests of scientists, architects, engineers, and archaeologists towards architectural heritage preservation during the last decades. The main goals of the building materials characterization are their preservation and restoration. In some cases, material research is conducted with the help of accurate statistical analysis of the geochemical compositions of samples. Such an analysis can be carried on by visual exploration of the data, or by semi-automated techniques. The classical statistical tools to this end, however, often require a certain numerosity of samples, as well as that the statistical variables described by such samples behave in some specific way, in order to be applied. Artificial intelligence methods, on the other hand, are frequency-based, and do not require any specific hypothesis to be applied, resulting in a more realistic range of applicable cases. In this paper, we proved that geochemical characterization of antique building by means of clay composition analysis can be seen as a feature selection for classification problem. Using only few available samples of clay composition of bricks and *terracottas* of 9 different building in the city of Ferrara (NE Italy) we were able to devise unique geochemical fingerprints of the building themselves, within a reasonable statistical reliability of the results. Not only our results are important in the context from which they are obtained, but they are also an empirical evidence that in most cases, with

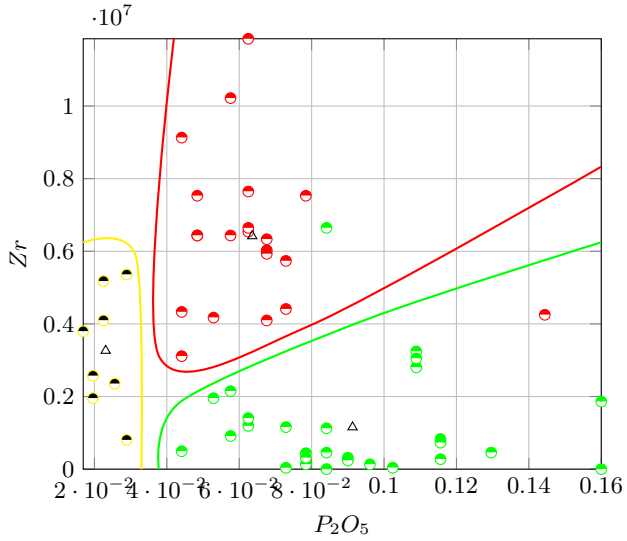

Fig. 1. Fingerprint representation of solution 12 (Duomo building in red, Pietro building in green and Roverella building in yellow; centroids are represented with a triangle)

real-world, expensive data (in this case, samples were collected in a period that spans 10 years) intelligent techniques must, and can, adapt to the situation in a better way compared to classical approaches.

As future work, we want to extend this research to include a second fingerprinting step which would allow us to establish which type of clay has been used for each building. This can be done starting from *typical clay samples* of the area surrounding Ferrara, and using an adaptation of our technique to establish which samples is the closest one to actual composition of the clay taken from the buildings. Such a fingerprinting process, if successful, can help the historians in their job, and could be replicated in other, less historically clear situations around Italy and the world.

Acknowledgments

E. Lucena-Sánchez and G. Sciavicco acknowledge the support of the project *New Mathematical and Computer Science Methods for Water and Food Resources Exploitation Optimization*, founded by the Italian region Emilia-Romagna.

References

1. Amorosi, A., Centineo, M., Dinelli, E., Lucchini, F., Tateo, F.: Geochemical and mineralogical variations as indicators of provenance changes in late quaternary deposits of se po plain. Sedimentary Geology **151**(3-4), 273–292 (2002)

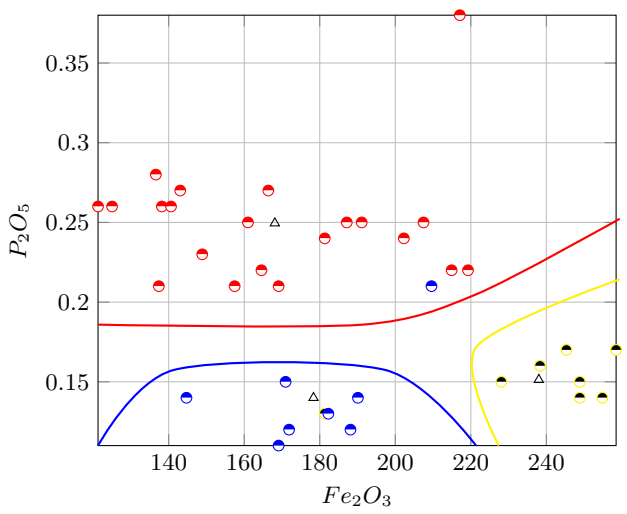

Fig. 2. Fingerprint representation of solution 1 (Grazie building in blue, Duomo building in red and Roverella building in yellow; centroids are represented with a triangle).

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. Benedetto, C.D., Graziano, S., Guarino, V., Rispoli, C., Munzi, P., Morra, V., Cappelletti, P.: Romans'established skills: Mortars from d46b mausoleum, porta mediana necropolis, cuma (naples). Mediterranean Archaeology and Archaeometry **18**, 131–146 (2018)
6. Bianchini, G., Laviano, R., Lovo, S., Vaccaro, C.: Chemical–mineralogical characterisation of clay sediments around ferrara (italy): a tool for an environmental analysis. Applied Clay Science **21**(3-4), 165–176 (2002)
7. Bianchini, G., Marrocchino, E., Moretti, A., Vaccaro, C.: Chemical-mineralogical characterization of historical bricks from ferrara: an integrated bulk and micro-analytical approach. Geological Society, London, Special Publications **257**(1), 127–140 (2006)
8. Blo, G., Conato, C., Contado, C., Fagioli, F., Vaccaro, C., Dondi, F.: Metal content in river suspended particulate matter: Data on po river. Annali di Chimica: Journal of Analytical, Environmental and Cultural Heritage Chemistry **94**(5-6), 353–364 (2004)
9. Collette, Y., Siarry, P.: Multiobjective Optimization: Principles and Case Studies. Springer Berlin Heidelberg (2004)

10. Dean, T.: Land and Power in Late Medieval Ferrara: The Rule of the Este, 1350-1450. No. 7, Cambridge University Press (1987)
11. Deb, K.: Multi-objective optimization using evolutionary algorithms. Wiley, London, UK (2001)
12. Di Roma, A., Lucena-Sánchez, E., Sciavicco, G., Vaccaro, C.: Towards automatic fingerprinting of groundwater aquifers. In: Proc. of the 6th International Conference on Technologies and Innovation. Communications in Computer and Information Science, vol. 1309, pp. 73–84. Springer (2020)
13. Durillo, J., Nebro, A.: Jmetal: a Java framework for multi-objective optimization. Avances in Engineering Software **42**, 760 – 771 (2011)
14. 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)
15. 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)
16. Franzini, M., et al.: Revisione di una metodologia analitica per fluorescenza-x, basata sulla correzione completa degli effetti di matrice. (1975)
17. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research (3), 1157–1182 (2003)
18. Hoefs, J.: Geochemical fingerprints: A critical appraisal. European Journal of Mineralogy (22), 3 – 15 (2009)
19. Jang, W.S., Engel, B., Yeum, C.M.: Integrated environmental modeling for efficient aquifer vulnerability assessment using machine learning. Environmental Modelling & Software **124**, 104602 (2020)
20. Jiménez, F., Sánchez, G., García, J., Sciavicco, G., Miralles, L.: Multi-objective evolutionary feature selection for online sales forecasting. Neurocomputing **234**, 75–92 (2017)
21. 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)
22. Lachance, G.: Practical solution to the matrix problem in x-ray analysis. Canadian Spectroscopy **11**, 43–48 (1966)
23. Lary, D., Alavi, A., Gandomi, A., Walker, A.: Machine learning in geosciences and remote sensing. Geoscience Frontiers **7**(1), 3 – 10 (2016)
24. Lary, D., Muller, M., Mussa, H.: Using neural networks to describe tracer correlations. Atmospheric Chemistry and Physics (4), 143–146 (2004)
25. Liu, H., Zhou, X., Zhang, X., Wu, K., Lu, C.: Experimental study and matrix effect correction of pseudobinary samples in xrf analysis. In: IOP Conference Series: Materials Science and Engineering. vol. 389. IOP Publishing (2018)
26. López-Arce, P., Garcia-Guinea, J., Gracia, M., Obis, J.: Bricks in historical buildings of toledo city: characterisation and restoration. Materials Characterization **50**(1), 59–68 (2003)
27. 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)
28. Marrocchino, E., Telloli, C., Vaccaro, C.: Geochemical and mineralogical characterization of construction materials from historical buildings of ferrara (italy). Geosciences **11**(1), 31 (2021)

29. McIver, K.A.: The este monuments and urban development in renaissance ferrara **13**, 230–233 (1999)
30. McIver, K.A., Rosenberg, C.M.: The este monuments and urban development in renaissance ferrara **29**, 121 (1998)
31. 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)
32. 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)
33. 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)
34. 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)
35. Potts, P.: X-ray fluorescence analysis: principles and practice of wavelength dispersive spectrometry. In: A Handbook of Silicate Rock Analysis, pp. 226–285. Springer (1987)
36. Rosenberg, C.M.: The este monuments and urban development in renaissance ferrara (1997)
37. Rousseau, R.M.: Corrections for matrix effects in x-ray fluorescence analysis—a tutorial. Spectrochimica Acta Part B: Atomic Spectroscopy **61**(7), 759–777 (2006)
38. Sanfilippo, G., Aquilia, E.: Multidisciplinary process aimed at the diagnosis and treatment of damages in stony cultural heritage: The balustrade of villa cerami (catania). Mediterranean Archaeology & Archaeometry **18**(5) (2018)
39. Shahin, M., Jaksa, M., Maier, H.: Artificial neural network applications in geotechnical engineering. Australian Geomechanics **1**(36), 49–62 (2001)
40. 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)
41. Witten, I., Frank, E., Hall, M.: Data mining: practical machine learning tools and techniques, 3rd Edition. Morgan Kaufmann, Elsevier (2011)
42. 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)