Research Article Environmetrics

Received: 16 May 2012, Revised: 07 March 2013, Accepted: 03 July 2013, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.2222

The case of arsenic contamination in the Sardinian Geopark, Italy, analyzed using symbolic machine learning

Germana Manca[a]* and Guido Cervone[b]

This paper analyzes the relationship among different chemical pollutants retrieved from *in situ* measurements of underground and surface water in a region of possible development. The area of study is the former mine of the Iglesiente district, which is now a United Nations Educational, Scientific, and Cultural Organization (UNESCO) protected region in the island of Sardinia (Italy). A full chemical analysis of water/soil samples were collected at the site in 2004. The data show the presence of several toxic contaminants above the national legal threshold. A symbolic machine learning classifier is employed to learn strong patterns associated with a high level of arsenic (As) in the soil samples. The patterns discovered show complex relationships that include both high and low concentrations of different chemicals. The strongest patterns are found between As and the chemicals which are usually found in the soil. This implies that when As is dissolved in the water table, it is expected these other chemicals are also present. It emerges that a specific relationship of As–phosphates is outlined and is clearly shown by applying the symbolic machine learning classifier. This leads to an understanding of the behavior of these elements in the soil, potential impacts on ecosystems, as well as the pollution of groundwater. Finally, an assertion of the advantages of the algorithm quasi-optimal learning method is clarified in term of applicability in such circumstances. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: AQ; GIS; As; machine learning; pollution

1. INTRODUCTION

1.1. General problems, associated with former mines (pollution, cleaning, resettling, and reconversion), on the Sardinia island

Since ancient history, Sardinia has been a region of intense mining activities, which have profoundly modified the island's environment. Mining in Sardinia goes back six millennia BC, when obsidian mining was the prevalent activity. The hard and precious volcanic glass, used to shape weapons and tools, was traded along the Mediterranean coasts (Contu, 2000). The following centuries progressively saw unfolding civilizations, from Phoenician to Roman, interested in the mining resources. From the 18th century until recently, mining activities both sustained and flourished in the island's economy, as shown in Figure 1 (Manca and Pireddu, 2005). The resulting heritage is represented by degraded areas, including excavations and dumping grounds, polluted groundwater, and a devastated hydrological system.

Initially, the disruption of pumping out the mines caused the galleries to flood, and as a result, the water table rose to the surface. Consequently, the interaction between water and minerals strongly determines the water quality, which is characterized by a high level of salinity due to sulfate and acid pH caused by the oxidation of sulfide and high levels of pollutant metals. Acid mine drainage phenomena are also observed in the area. The main sources of contamination are the tailings stored in mine tunnels and abandoned along fluvial banks (Concas *et al*., 2006). Furthermore, the landfills are also degraded and polluted because of the toxic materials released by the mining process and progressively stocked in those locations (Progemisa-Geoparco, 2003). The landfills' water runoff leaks and dissolves those materials causing them to flow into the river network, leading to a high concentration of pollution. Surface water mixes with the soil and infiltrates into the groundwater yielding contaminated water at both the surface and subsurface. Consequently, in this abandoned mining district, because of poor management of environmental issues, significant heavy metal contamination has been observed (Fanfani *et al*., 2000). Comparison between new and old data indicates significant heavy metal contamination of superficial waters in the investigated area (Concas *et al*., 2006). Moreover, the heavy metal contamination clearly affects the morphological diversity of the pedons, the taxonic, the functional

* *Correspondence to: Germana Manca, Department of Environmental Science and Policy, George Mason University, 4400 University Dr, Fairfax, VA 22030, U.S.A. E-mail: gmanca@gmu.edu*

a *Department of Agriculture, Regional Government of Sardinia, Via Pessagno 4,09126, Cagliari, Italy*

b *Department of Geography and Institute for CyberScience, The Pennsylvania State University, 302 Walker Building, University Park, PA, 16802, USA*

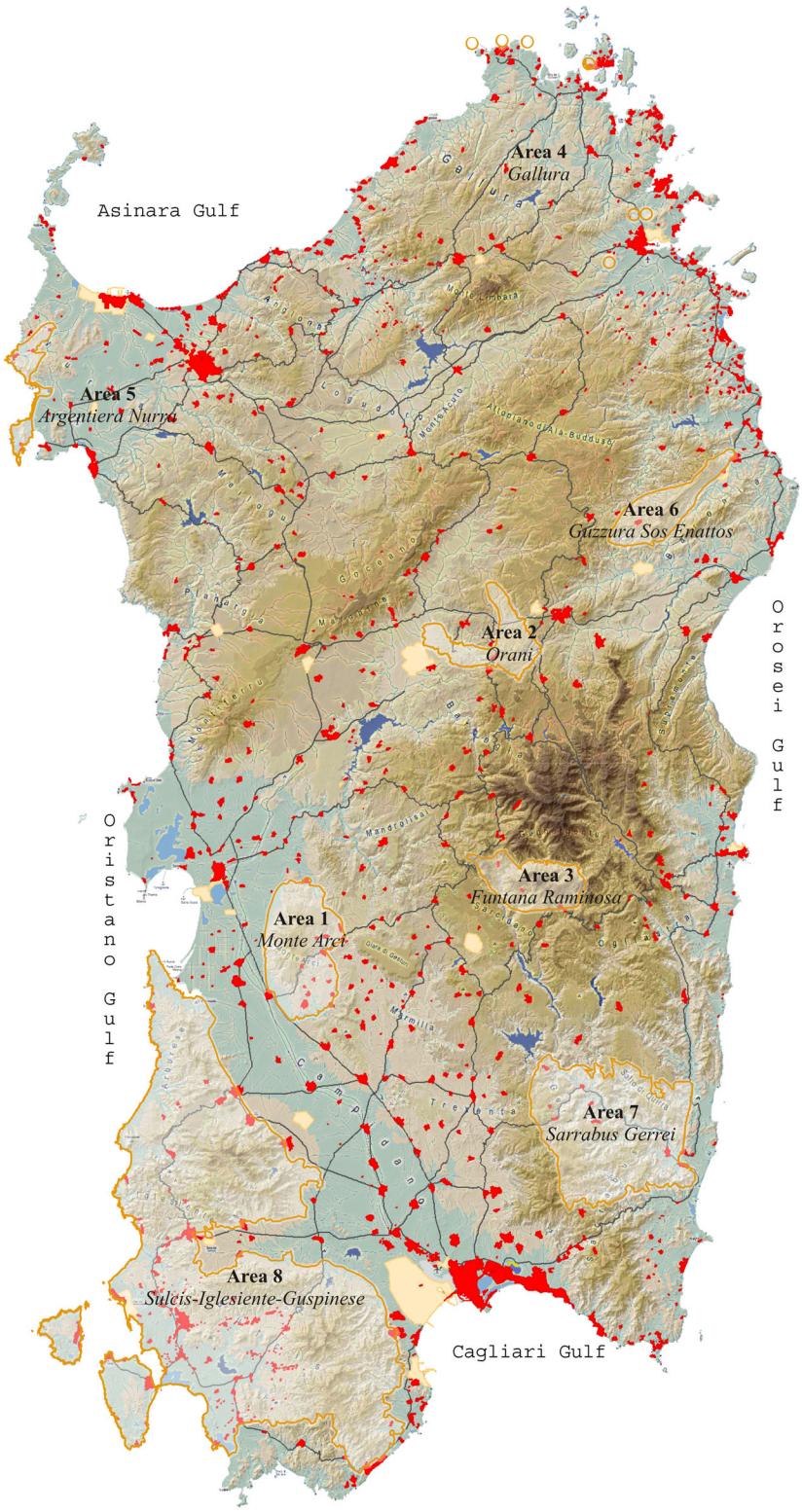

Figure 1. The Geopark's Area in the island of Sardinia. This figure is available in colour online at wileyonlinelibrary.com/journal/environmetrics

pedodiversity, and the soilscape pattern. Consequently, the total arsenic (As) content is very high in most of the soil profiles considered (Vacca *et al*., 2012). Furthermore, Costantini *et al*. (2004) highlighted the high concentration of As and its critical limit compared with other minerals in the soil and water and found it to be well over the Italian national mandatory limits (Ministero dell'Ambiente e della Tutela del Territorio, e del Mare, 2006). These concentrations limit the use of the land for public recreation areas, for industrial and commercial areas, and even for residential areas.

1.2. An international problem

Arsenic is a natural component of some raw minerals, including lead (Pb), zinc (Zn), and copper (Cu). During mineral mining, As is released by these minerals and is dispersed into the environment. Moreover, As is responsible for groundwater pollution in many countries worldwide (MIT, 2002; Ravenscroft *et al*., 2009). For instance, in Europe, the presence of several main contaminants affecting soil has been reported by the European Environmental Agency, (2010). They determined that the heavy minerals account figuring that the heavy metals count for 37.3% of the overall amount of contaminants, of which, As is one. On the other side of the shore, since 1997, the US Environmental Protection Agency has been using almost 41% of its available funds to classify As sites as polluted, the second level of attention after Pb, in the National Priority List. (US EPA, 1997, 2004). The consequences are that affected people cope with skin cancer and persistent body poisoning.

Relationships between chemical compounds and As have been described by Costantini *et al*. (2004). in Sardinia's mine district and have been shown through correlation matrices and principal component analysis. Although the correlation matrices simply described the most significant dependencies, in particular for Pb/As, Zn/As, Cd/As, either in the soil or in the water, a principal component analysis has been carried out to better explain the variability of the chemical compounds. As shown, As is considered as one of the three components able to explain the observed pollution. Although the Principal Component Analysis reveals the intrinsic connections of the pollutants, Shlens (2009) pointed out its limitations based on the multivariate data sets reduction.

In this paper, a relationship is identified to explain the relationship of As with other chemicals observed in the soil/water samples, through the application of machine learning algorithms. Machine learning algorithms are used in different disciplines to identify non-trivial patterns in large amounts of data (Mitchell, 1997). They have been used to successfully characterize the source of non-point pollution of groundwater (Trauth and Xanthopoulos, 1997). Recently, Cordier *et al*. (2005) used machine learning classifiers to assess the environmental impact from contaminated water supplies, showing that it is possible to establish a strong link between land use and groundwater contamination.

Here, a symbolic machine learning classifier based on the AQ (algorithm quasi-optimal) methodology (Mitchell, 1997; Cervone *et al*., 2010) is used to study the relationship between high levels of As and other chemicals observed in the water/soil samples. The goal is to find an attribute value relationship between the consequent As (high concentration of As) and the conditions (the other chemicals).

2. METHODOLOGY

A machine learning classifier is used to generalize sets of examples associated with high levels of As against counterexamples associated with low levels of As. The input data consist of labeled data in a tabular form. The rows are the different soil samples taken at different locations, and the columns are the concentrations of chemicals measured. Each row of data is assigned a label indicating a level of As below or above the safety threshold. Unlike clustering, a form of unsupervised learning whose goal is dividing unlabeled data into distinct classes, classification is a form of supervised learning, where classified data are generalized to identify the characteristics of the entire class (Cervone *et al*., 2010).

In its simplest form, given two sets of multivariate descriptions, or events, $p_1 \ldots p_{n,}$ and $N_1 \ldots\ldots N_m$, a symbolic classifier finds rules that cover all p examples (a.k.a. positive events) and does not cover any N examples (a.k.a. negative events). More generally, each multivariate description is a classified event of type $\{x_1, \ldots x_k\}$and c, where each x is an attribute value, and c is the class to which it belongs.

In this paper, we used a form of AQ learning (e.g., Cervone and Panait, 2001), a proven machine learning methodology aimed at learning symbolic decision rules from a set of examples and counterexamples. It was first proposed in the late 1960s to solve the Boolean function satisfiability problem and was further refined over the following decade to solve the general covering problem (Mitchell, 1997). In its newest implementations, it is a powerful but yet little explored methodology for symbolic machine learning classification. It has been applied to solve several problems from different domains, including generating individuals within an evolutionary computation framework. The algorithm learns from examples (positives) and counterexamples (negatives) and from patterns (a.k.a. rules) of attribute values that discriminate the characteristics of the positive events with respect to the negative events. Such patterns are generalizations of the individual positive events and depending on the AQ's mode of operation may vary from being totally complete (covering all positives) and consistent (not covering any negatives) to accepting a trade-off of coverage to gain simplicity of patterns. The AQ is a general purpose symbolic machine learning classifier that learns induction rules from a set of examples and counterexamples. A detailed description of the AQ algorithm and its implementation used for this study is provided by Cervone *et al*. (2010). The presented approach is not limited to AQ, and other classifiers could be used to discriminate between the chemical samples. However, AQ provides a few crucial features that are well suited for this study. Among those, the ability to work both with numerical and categorical values is important because of the nature of the observations. In addition, AQ generates rules that are easy to inspect and to validate, thus allowing to interpret the attribute value relationships of the different chemicals. A neural network, for example, could be well suited to discriminate between the chemicals but is not very useful to understand their relationships. Finally, AQ is well suited to work with data that are incomplete and contain noise through its pattern discovery mode of operation, which was used in the current study.

3. REGION OF THE STUDY AND DATASET—BARRAXIUTTA AREA

Barraxiutta is located in southwestern Sardinia (Italy), and the area of interest spreads over the entire watershed of Domusnovas (Figure 2), which extends 38.5 km^2 and includes local rivers and small ponds. Its area is also involved in the Territorial Landscape Planning (Autonomous Region of Sardinia), defined as "Bacino Metallifero 11" (Mines Basin 11) and named as "Marganai". In the 1700s and 1800s, huge accumulations of mineral waste were visible close to the old flotation area, and high concentrations of lead (around 10–14%) and silver (around 60%) were persistent. The extraction of mineral continued until 1967 and stopped definitively 3 years later in 1970 because of the mine depletion.

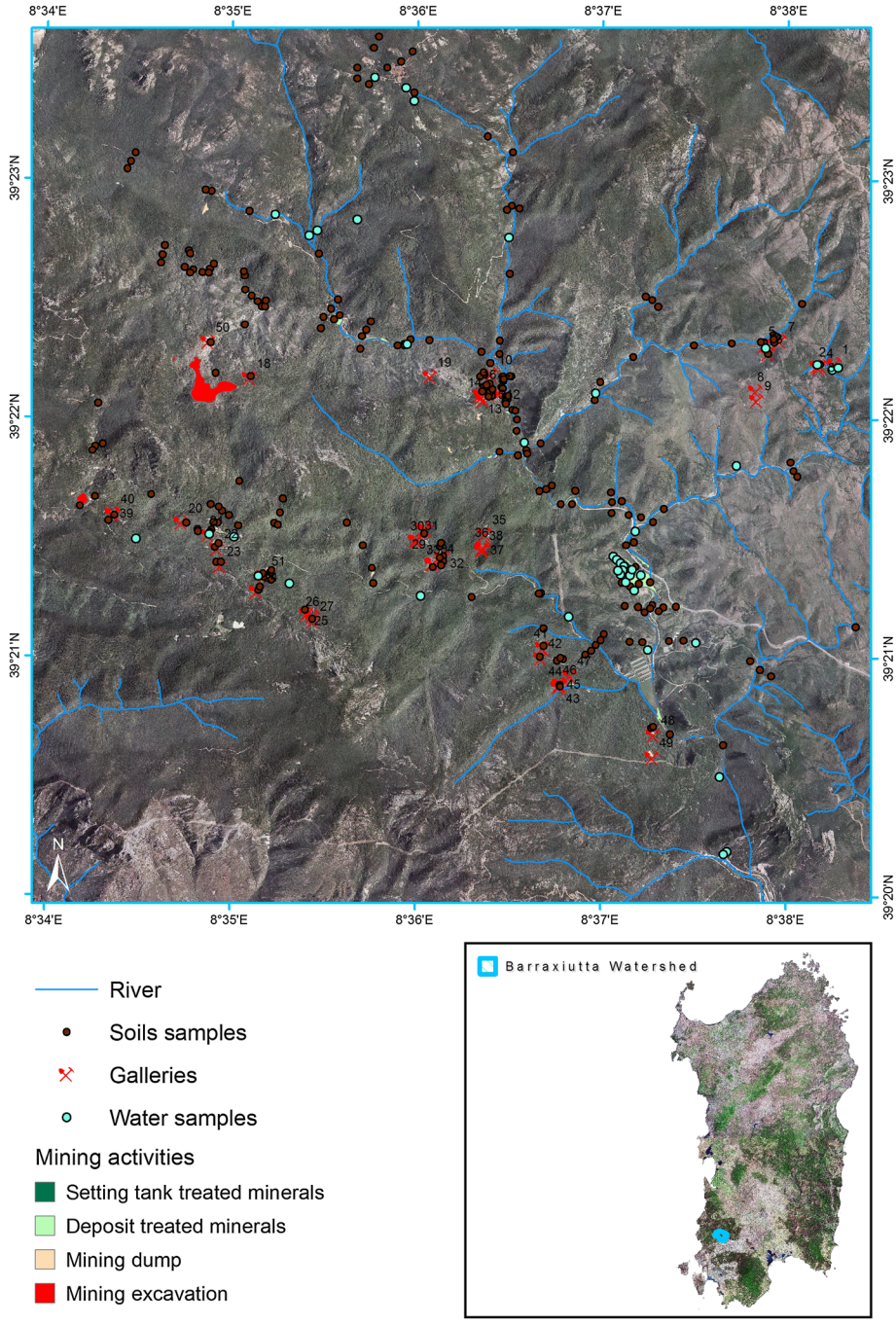

Figure 2. Barraxiutta Basin, area of interest, and geographical location of the samples. This figure is available in colour online at wileyonlinelibrary.com/journal/environmetrics

As a consequence of the previous mine activities, Barraxiutta is described and is identified by experts through the "Piano di Caratterizzazione di Barraxiutta" (Progemisa-Geoparco, 2003), as an area of potential pollution. An area of potential pollution is a specific site where residuals mineral materials are collected in high concentrations. These mineral materials are dispersed by means of wind and rain, and cause widespread contamination of the soil, surface and underground waters. The pollutants are composed of wastes of different sizes, such as agglomerates of residuals wrapped into silt, clay, and mixed by sand. Furthermore, the report demonstrates that the chemical and physical alteration of the soil, vegetation, and waters are still increasing. An index has been calculated to quantify the contamination and it reccomends high attention to human health and the ecosystem, further suggesting that restoration is a priority for this region.

In this study, the AQ algorithm was applied to analyze 303 soil samples and to generate patterns that characterize the relations between the chemicals. In addition, the soil samples were collected from several places surrounding the mines. For each soil sample, the description is reported, such as its elevation, the geographical location and name, watershed, land use type, contamination, lithology, description, deep, and texture; in the laboratory, analysis on pH, loss on ignition, and other chemical components (Na_2O, MgO, Al_2O_3, SiO_2, P_2O_5, K_2O,

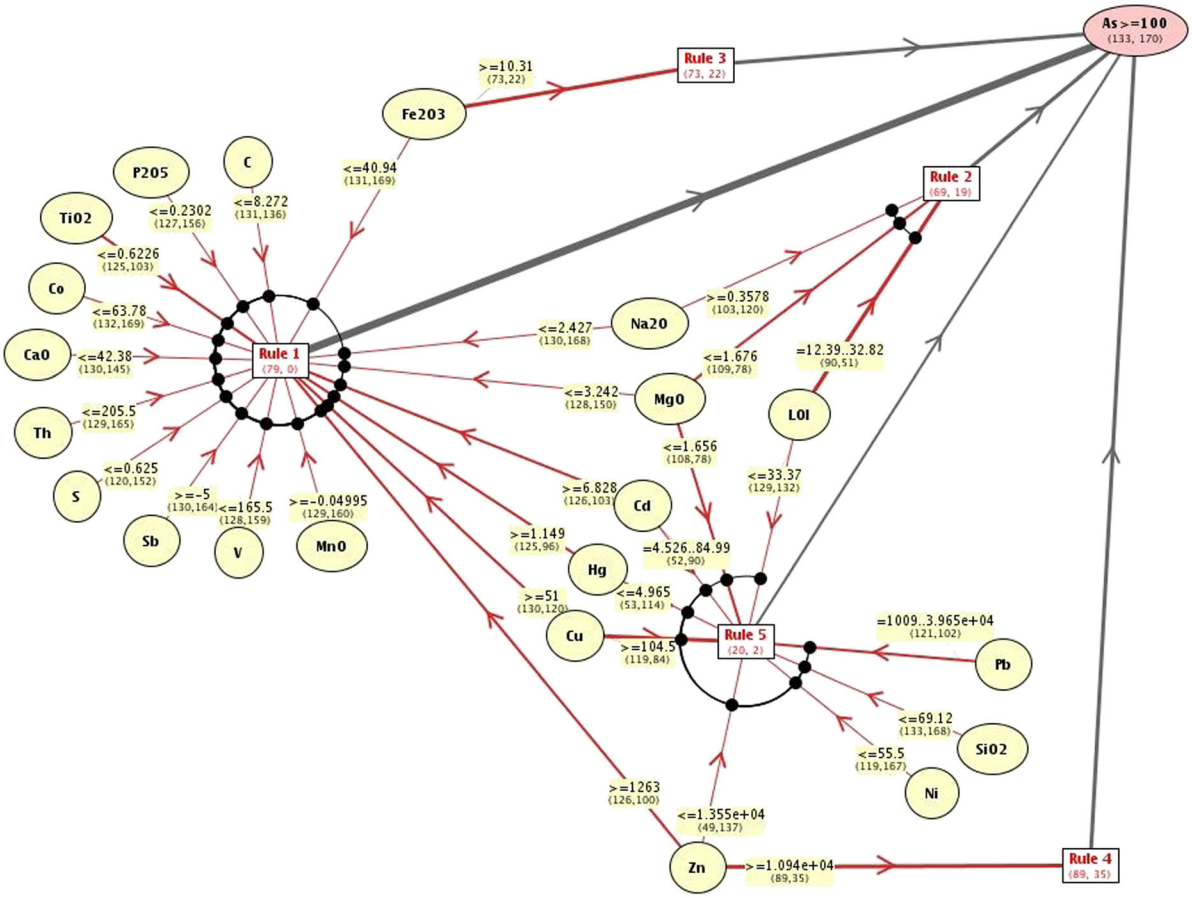

Figure 3. Algorithm quasi-optimal learning output. This figure is available in colour online at wileyonlinelibrary.com/journal/environmetrics

CaO, TiO_2, MnO, Fe_2O_3, As, Cd, Co, Cu, Hg, Ni, Pb, Sb, Th, V, Zn, C, and S) [1] are measured. Additionally, a series of groundwater samples were collected, but not used for the analysis, from wells that were already available and from newly built wells. For each well, the piezometric level is determined, and pH, temperature, and conductibility are detected. In the laboratory, analysis on the ionic balance, salinity, nitrogen, ammonium, sulfate ion, and 10 metallic elements are identified (Al, As, Cd, Cu, Fe, Hg, Mn, Ni, Pb, and Zn). The geographic location of these samples is shown in Figure 2.

4. APPROACH: ALGORITHM QUASI-OPTIMAL SYMBOLIC MACHINE LEARNING CLASSIFIER

The AQ was used to learn patterns associated with high levels of As (As > 100 mg/kg) over the threshold determined by the law D. Lgs. 152/06 (Figure 3).

The goal is to identify the relationship between high As concentrations with respect to the amounts of other chemicals. The AQ discovered five rules, visualized in Figure 3 in the form of an association graph, used to visualize attribution rules that characterize multivariate dependencies between the target class and the input attributes. The target class (As > 100 mg/kg) is associated only with unique patterns of input parameters.

Each relationship is indicated with links between the input attributes and the target class. Several links are grouped together in individual rules to show the conjunctions of the attribute value relationship associated with the target class. These conjunctions are indicated with arcs, and each rule is shown with a rectangular node. The target class is associated with a disjunction of rules, indicating that it is enough that one rule is satisfied for the target class to be implied.

The thickness of the links indicates the weight of a particular parameter value combination in the definition of the cluster. Each input attribute, shown in its own ellipsis, is associated to a specific rule with a link. On each link, there is an annotation in the form of a relation (=, ≥, ≤) and a value. Underneath the relation, there are two values in parentheses, which indicate the number of positive and negative events

[1]AL, aluminum; As, arsenic; C, carbonium; Cd, cadmium; Co, cobalt; Cu, copper; Fe, iron; Hg, mercurium; K, potassium; Mn, mangnese; Na, sodium; Ni, nickel; P, phosphorus; Pb, lead; S, sulfur; Sb, Antimony; Th, thorium; Ti, titanium; V, vanadium; Zn, zinc.

covered by the specific relation. Each rule, shown in a rectangle, is also annotated with the number of positive and negative events that it covers; the line thickness linking rule and output attribute ($As > 100$) is a function of this coverage.

The rules show the relationship between the input attributes and the output attributes. In the dataset used in the study, there are 133 cases with $As \geq 100$, called the positive events and 170 cases with $As < 100$, called negative events. Rule 1 covers 79 positive events and 8 negative events, thus establishing a strong pattern between the conjunction of the input attributes and the consequent, in this case, $As \geq 100$. This rule is paramount for characterizing new unseen events, or events for which As has not been observed. If a new event is observed with measurements that fall within the constraints set by the rules found, it is possible to immediately predict that the As content will be above 100 the acceptability threshold and thus will represent an area with a substantial pollution risk. According to the rules, $As > 100$ is related to P_20_5 through a very complex relationship. For example, Rule 1 in the figure is the following:

$$
\begin{aligned}
As > 100 \quad \leftarrow \quad & Fe_20_3 \leq 40.94, \\
& C \leq 8.72, \\
& P_2O_5 < 0.2302, \\
& Ti0_2 \leq 0.6226, \\
& C0 \leq 63.78, \\
& Ca0 \leq 42.38, \\
& Th < 205.5, \\
& S < 0.625, \\
& Sb \geq -5, \\
& V \leq 165.5, \text{ and} \\
& Mn0 \geq -0.04995
\end{aligned}
$$

To indicate that all these conditions must be held true for $As > 100$ to be found. This rule covers 79 positive and 0 negative examples (positive examples are $As > 100$ and negative examples are $As \leq 100$). This rule covers only 79 of the 133 cases, and a combination of all the rules, 1 through 5, is necessary to describe all the positive cases.

5. CONCLUSIONS

The soil mobility of the As depends on the type and quantity of colloids (Fe, Mn, Al, clay, and organic matter) in the soil, which determines absorption. Also, pH, redox potential, and anion, and compete with As for the same area of adsorption (Sadiq *et al*., 1996). Substantially, the As is found either in stable "inner-sphere complexes" with Fe, Mn, and aluminum oxides, or with clay minerals. Clay minerals especially capture the As, on the basis of their potential absorption surface through the "inner-sphere" link, where the positive charge of the clay material attracts the As and fixes to it. This is why Figure 3 shows a strong link with the oxides Fe, Al, and Mn, which specifically absorb heavy metals, and their impact on the ecosystem is clarified and is well known (Cornell and Schwertmann, 1996).

In contrast, the link with phosphorus was not so predictable. The image clarifies and strengthens the concept that the phosphates compete against the As to merge with the clay minerals *in situ*, and where phosphates are represented in the soil, As is released. The behavior of the As ion is similar to that of phosphates because of the specific anion absorption in the soil moisture and the soil components creating stable links on the surface of the clay minerals such as the "inner-sphere" already mentioned (Sun and Doner, 1996; Liu *et al*., 2001; O'Really *et al*., 2001). The mobility of As, then, is determined by the soil de-absorption caused by the high concentration of phosphates (Woolson *et al*., 1973; Peryea, 1991; Violante and Pigna, 2002). Furthermore, phosphates determine an increase of bioavailability and concentration of As in the soil for the root systems of the vegetation, as a consequence of the competition between As and phosphates in the absorption site of the soil (De Santis, 2010). The AQ algorithm supports the hypothesis that As is released into the environment when phosphate minerals are present, showing that the association–competition plays an important role in the water–soil balance. It is important to remember that rules generated by any classifiers through inductive reasoning could be incorrect. By nature, inductive reasoning is not truth preserving, such as deductive reasoning (e.g., Aristotelian syllogism—modus ponens), therefore, it cannot be used to validate a hypothesis but only to provide support for it.

The consequence is that where phosphates are available, As is also available. Phosphates are likely absorbed by the clay surface, to the detriment of As. Therefore, the symbolic machine learning analysis performed enhances the understanding of the "game of chemicals life" and plays a key role in deducting and synthesizing the chemical behaviors. It also shows the connections among the chemical compounds and determines the relevance of a certain compound over the others.

The present research illustrates an alternative way to investigate the cause–effect of pollutants in soil samples using a symbolic machine learning classifier. The results, relative to the area of Barraxiutta in Sardinia, Italy, confirm the presence of strong patterns between high levels of As and other chemicals, identified in the areas surrounding the disused mines. This methodology can be applied to a wide range of similar studies.

Acknowledgements

We would like to thank Laura Pireddu, IFRAS, for suggesting the database used in this analysis.

REFERENCES

Cervone G, Panait LA. 2001. The development of the AQ20 learning system and initial experiments. *Tenth International Symposium on Intelligent Information Systems*, Zakopane, Poland.

Cervone G, Franzese P, Keesee A. 2010. Algorithm quasi-optimal (AQ) learning. WIREs: Computational Statistics 2: 218–236.

Concas A, Ardau C, Cristini A, Zuddas P, Cao G. 2006. Mobility of heavy metals from tailings to stream waters in a mining activity contaminated site. Chemosphere 63: 244–253.

Contu E. 2000. The Prehistoric Altar of Monte D'Accoddi, Carlo Delfino Edition Sassari, Italy. ISBN 88-7138-206-20004

Cordier MO, Garcia F, Gascuel-Odoux C, Masson V, Salmon-Monviola J, Tortrat F, Trpos R. 2005. A machine learning approach for evaluating the impact of land use and management practices on streamwater pollution by pesticides. *Proceedings of MODSIM*.

Cornell RM, Schwertmann U. 1996. THE IRON OXIDES: STRUCTURE, PROPERTIES, REACTIONS, OCCURRENCES AND USES. VCH: Weinheim.

Costantini S, Bodano L, Giordano R, D'Ilio S. 2004. Contaminazione ambientale da metalli pesanti connessa con attivita' mineraria dismessa in Sardegna. Studio preliminare. *Rapporti ISTISAN* 4/28. ISSN 1123-3117.

De Santis R. 2010. Mobilita', Speciazione e fitodisonibilita' di arsenic nel sistema suolo-acqua-pianta. PhD Thesis University of Naples.

European Environmental Agency. 2010. Overview of contaminants affecting soil and groundwater in Europe. *European Topic Center Soil*. Sep 05, 2011. Web May 5, 2012. Available online at http://www.eea.europa.eu/data-and-maps/figures/overview-of-contaminants-affecting-soil-and-groundwater-in-europe

Fanfani L, Caboi R, Cidu R, Cristini A, Frau F, Lattanzi P, Zuddas P. 2000. Impatto ambientale dell'attivita' mineraria in Sardegna: studi mineralogici e geochimici. Rendiconti Seminario Facolta' Scienze Universita' Cagliari. Supplemento Vol.70.

Liu F, De Cristofaro A, Violante A. 2001. Effect of pH phosphate and oxalate on the dsorption/desorption of arsenate on/from goethite. Soil Science Society of America Journal 166: 197–208.

Ministero dell'Ambiente e della Tutela del del Territorio e del Mare. Repubblica Italiana, 2006. Decreto Legislativo 3 aprile 2006, n. 152. Norme in materia ambientale. Gazzetta Ufficiale n.88. Supplemento ordinario n.96.

Manca G, Pireddu L. 2005. Sardinian Environmental Historical Geopark: a development pportunity. 6th European Geopark Meeting. Lesvos, Greece.

MIT, Massachussets Institute of Technology. 2002. Arsenic in Bangladesh, drinking wells maybe linked to crop irrigation. *MIT news*. Nov 21, 2002. Web May 5, 2012. Available online at http://web.mit.edu/newsoffice/2002/bangladesh.html

Mitchell T. 1997. Machine Learning. Mac Graw Hill: New York.

O'Really SE, Strawn DG, Sparks DL. 2001. Residence time effects on arsenate adsorption/desorption mechanisms on goethite. Soil Science Society of America Journal 65: 67–77.

Peryea FJ. 1991. Phosphate indiced release of arsenic from soil contaminated with lead arsenate. Soil Science Society of America Journal 55: 1301–1306.

Progemisa-Geoparco. 2003. Piano della Caratterizzazione di Barraxiutta. Relazione Tecnico Descrittiva.

Ravenscroft P, Brammer H, Richards K. 2009. Arsenic Pollution. A Global Synthesis, Wiley-Blackwell Edition: Singapur ISBN978-1-4051-8602-5

Sadiq R Olszowy H, Shaw G, Biltoft R, Cornell D. 1996. Soil and water contamination by arsenic from tannery waste. Water, Air, & Soil Pollution 78: 189–198.

Shlens J. 2009. A tutorial on principal component analysis. Web April 22, 2009. Version 3.01. Available online at http://www.snl.salk.edu/~shlens/pca.pdf

Sun X Doner HE. 1996. An investigation of arsenate and arsenite bonding structures on goethite by FTIR. Soil Science 161: 865–872.

Trauth R Xanthopoulos C. 1997. Non-point pollution of groundwater in urban areas. Water Research 31(11): 2711–2718, ISSN 0043-1354, doi:10.1016/S0043-1354(97)00124-3.

U.S. Environmental Protection Agency, EPA. 1997. Office of Solid Waste and Emergency Response. Recent Development for In-Situ Treatment of Metal Contaminated Soils; U.S. Government Printing Office: Washington, DC; EPA-542-R-97-004.

U.S. Environmental Protection Agency, EPA. 2004. Monitoring arsenic in the environment: a review of science and technologies for field measurements and sensors.

Vacca A, Bianco MR, Murolo M, Violante P. 2012. Heavy metals in contaminated soils of the Rio Sitzerri floodplain (Sardinia, Italy): characterization and impact on pedodiversity. Land Degradation & Development 23: 350–364.

Violante A, Pigna M. 2002. Competitive sorption of arseniate and phosphate on different clay minerals and soil. Soil Science Society of America Journal 66: 1788–1796.

Woolson EA, Axely JH Kearney PC. 1973. The chemistry and phytotoxicity of arsenic in soils: effect of time and phosphorus. Soil Science Society of America Journal 37: 254–259.

