Science of the Total Environment 603–604 (2017) 167–177

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Developing a new Bayesian Risk Index for risk evaluation of soil contamination

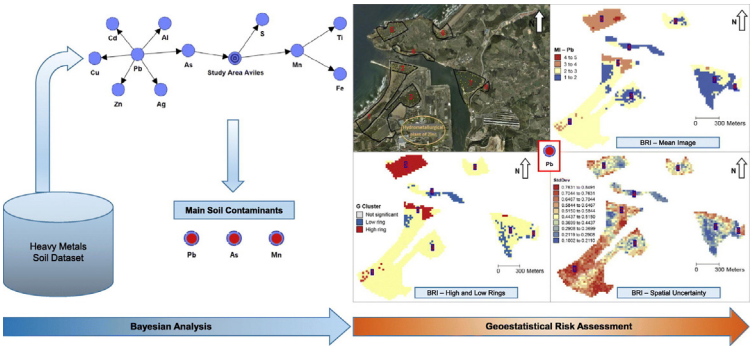
M.T.D. Albuquerque [a,b,*], S. Gerassis [c], C. Sierra [d], J. Taboada [c], J.E. Martín [c], I.M.H.R. Antunes [e,b], J.R. Gallego [f]

[a] Instituto Politécnico de Castelo Branco, 6001-909 Castelo Branco, Portugal
[b] CERENA/FEUP Research Center, Portugal
[c] Department of Natural Resources and Environmental Engineering, Univ. of Vigo, Lagoas Marcosende, 36310 Vigo, Spain
[d] Departamento de Transportes, Tecnología de Procesos y Proyectos, Universidad de Cantabria, Campus de Torrelavega, Spain
[e] ICT/University of Minho, Braga, Portugal
[f] INDUROT and Environmental Technology, Biotechnology, and Geochemistry Group, Universidad de Oviedo, Campus de Mieres, Asturias, Spain

HIGHLIGHTS

- Soils geochemical data was assessed with different Bayesian network approaches.
- A Bayesian Risk Index (BRI) for assessing soil contamination was constructed.
- BRI's spatial patterns were constructed throughout geostatistical modeling.
- Clusters of high PTEs cm, concentration are overlapping an agriculture zone.
- Metallurgical plant's emissions are the main source of soils pollutants.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 12 April 2017
Received in revised form 8 June 2017
Accepted 9 June 2017
Available online xxxx

Editor: D. Barcelo

Keywords:
Potentially toxic elements
Bayesian networks
Sequential Gaussian simulation
Local G clustering

ABSTRACT

Industrial and agricultural activities heavily constrain soil quality. Potentially Toxic Elements (PTEs) are a threat to public health and the environment alike. In this regard, the identification of areas that require remediation is crucial. In the herein research a geochemical dataset (230 samples) comprising 14 elements (Cu, Pb, Zn, Ag, Ni, Mn, Fe, As, Cd, V, Cr, Ti, Al and S) was gathered throughout eight different zones distinguished by their main activity, namely, recreational, agriculture/livestock and heavy industry in the Avilés Estuary (North of Spain). Then a stratified systematic sampling method was used at short, medium, and long distances from each zone to obtain a representative picture of the total variability of the selected attributes. The information was then combined in four risk classes (Low, Moderate, High, Remediation) following reference values from several sediment quality guidelines (SQGs). A Bayesian analysis, inferred for each zone, allowed the characterization of PTEs correlations, the unsupervised learning network technique proving to be the best fit. Based on the Bayesian network structure obtained, Pb, As and Mn were selected as key contamination parameters. For these 3 elements, the conditional probability obtained was allocated to each observed point, and a simple, direct index (Bayesian Risk Index-BRI) was constructed as a linear rating of the pre-defined risk classes weighted by the previously obtained probability. Finally, the BRI underwent geostatistical modeling. One hundred Sequential Gaussian Simulations (SGS) were computed. The Mean Image and the Standard Deviation maps were obtained, allowing the definition of High/Low risk clusters (Local G clustering) and the computation of spatial uncertainty. High-risk clusters are

* Corresponding author at: IPCB, Av. Pedro Álvares Cabral, n° 12, 6000-084 Castelo Branco, Portugal.
E-mail address: teresal@ipcb.pt (M.T.D. Albuquerque).

http://dx.doi.org/10.1016/j.scitotenv.2017.06.068

mainly distributed within the area with the highest altitude (agriculture/livestock) showing an associated low spatial uncertainty, clearly indicating the need for remediation. Atmospheric emissions, mainly derived from the metallurgical industry, contribute to soil contamination by PTEs.

1. Introduction

A map is always a simplification of reality (Lahr and Kooistra, 2010; Woodbury, 2003). A two-dimensional map can only gather and display the values of a limited number of variables or attributes up to three. Thus, when considering complex scenarios, such as environmental characterization, a reduction to a single dimension is mandatory (Moen and Ale, 1998). Risk maps, broadly mentioned in literature are keen for spatial pattern visualization of e.g. pollutant concentration distribution, exposure and its effects, vulnerability assessment; therefore, they constitute a very powerful tool to support policy-making in complex environmental risk assessment framework (Lahr and Kooistra, 2010; Li et al., 2015). Thus, Bién et al. (2005) applied their own Health Index/Risk Evaluation Tool to map the spatio-temporal probability of cancer cases in an area of soils contaminated with benzene and Zuo et al. (2017) used an Environmental Performance Index to evaluate the environmental performance of provinces in China between 2006 and 2011. Moreover, Entropy indexes have recently been used to provide pivotal information for mitigation strategies against desertification (Zambon et al., 2017) and Moreno-Jiménez et al. (2011), proposed a new methodology for assessing the site-specific environmental impact of contaminants at a local scale.

In the herein study, a set of 14 chemical elements, gathered in eight different zones (Fig. 1), was used to compute a Bayesian Network to analyze how high concentrations of PTEs linked up and to study how the presence of these elements can be mutually determined. A simple, direct index (Bayesian Risk Index - BRI) for land contamination assessment was produced as a linear rating of pre-defined risk classes weighted by the previously obtained Bayesian probabilities.

The practical development of Bayesian networks (BNs) has advanced greatly in the last two decades. Data integration allows efficient drawing solutions providing visual simplicity that is not attainable by other common statistical methods. The ability of BNs to make inferences and reduce uncertainty has caught the attention of a wide range of research fields. The oil and gas sector (Davies and Hope, 2015; Elsheikh et al., 2012) and a wide range of process industries (Gerstenberger et al., 2015; Wu et al., 2016; Zhang et al., 2013) have already used their benefits to enhance production, reduce occupational hazards, and evaluate potential risks.

Nevertheless, from an environmental perspective, these techniques have been so far practiced with a highly theoretical approach. The contribution of BNs to ecological risk assessment or natural resources management has been reported (McDonald et al., 2015; Nolan et al., 2015; Jiang et al., 2013; Phan et al., 2016). Recently, Taalab et al. (2015) proposed BNs as a suitable modeling approach for digital soil mapping taking a step forward on an issue still relatively unexploited (Aguilera et al., 2011).

Geostatistical techniques are based on the theory of regionalized variables (Matheron, 1971) which states that variables within an area show both random and spatially structured properties (Journel and Huijbregts, 1978). Experimental variograms must be estimated and modeled to quantify the spatial variability of random variables as a function of their separation lag (Antunes and Albuquerque, 2013). Geostatistics concerns to a broad methodological approach and it is more than the simple development of mathematical (probabilistic) models and methods and their application. In fact, analyzing the practical problems to be solved and formalizing them in terms of concepts is a key issue. When predicting the risk of contamination (e.g. months ahead), it is mandatory to stress the relevance of the chances for the future estimated values exceeding maximum admissible values. The delineation of zones of high and low risk requires the interpolation of risk values to the nodes of a regular grid making possible proper risk assessments, and a prediction model working as guidance to a more sustainable environmental management.

The main goal of this research is a straightforward procedure combining BNs and geostatistical techniques to evaluate the risk of soil contamination by PTEs. An industrialized site in the region of Avilés (Asturias, Spain) was used as pilot example. To this end, an innovative Bayesian Risk Index (BRI) was constructed. The subsequent geostatistical approach allowed the definition of the spatial distribution patterns of the BRI, focusing on the visualization and delineation of potential zones for future monitoring and remediation.

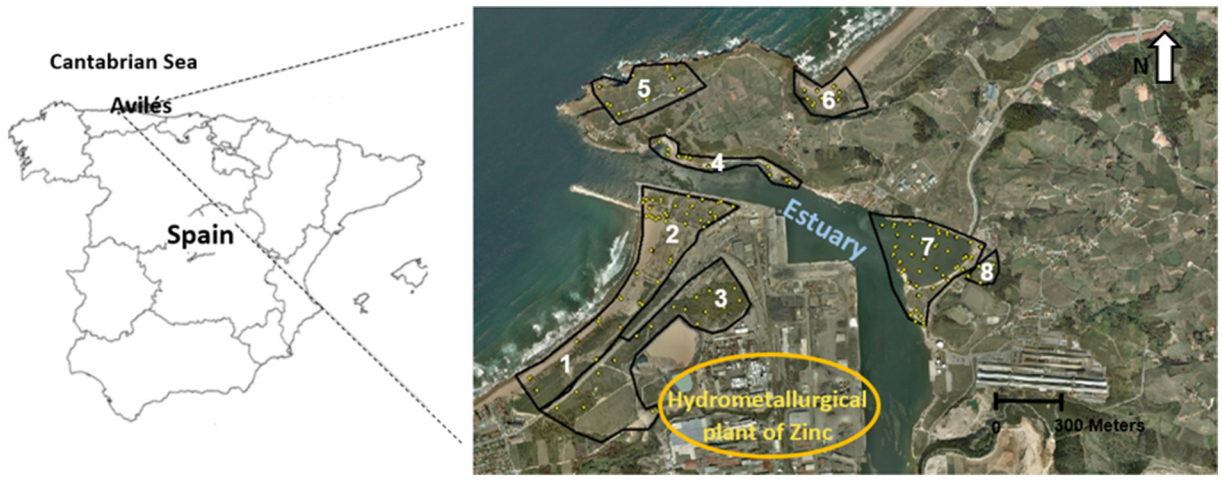


Fig. 1. Geographic location of the study area (left). Spatial distribution of the samples collected (yellow dots) and respective influence zones: 1. Beach; 2. Landfill; 3. Dunes (Landfill); 4. River beach; 5. Livestock/agriculture zone; 6. Background beach; 7. Llodero Cove pond; 8. Zeluán protected pond. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Standardized composition ranges in multifunctional soils. A - Canadian Council of Ministers of the Environment (2007); B - FAO and ISRIC guidelines for soil degradation in Central and Eastern Europe (1997); C – FOREGS; D – Others.

Heavy Metal	Low	Moderate	High	Remediation	A	B	C	D
Cu (mg/kg)	<50	50–100	100–500	>500	✓	✓		✓
Pb (mg/kg)	<70	70–140	140–600	>600	✓	✓		
Zn (mg/kg)	<200	200–300	300–500	>500	✓	✓		✓
Ag (mg/kg)	<0.2	0.2–0.3	0.3–0.5	>0.5			✓	
Ni (mg/kg)	<50	50–100	100–500	>500		✓	✓	✓
Mn (mg/kg)	<10	10–450	450–1000	>1000			✓	✓
Fe (%)	>0.1	0.1–5	5–10	>10			✓	
As (mg/kg)	>12	12–30	30–50	>50	✓	✓		
Cd (mg/kg)	<1	1–10	10–20	>20	✓	✓		
V (mg/kg)	<60	60–120	120–500	>500	✓		✓	
Cr (mg/kg)	<60	60–90	90–250	>250	✓	✓		
Ti (%)	0.02	0.02–1	1–5	>5			✓	✓
Al (%)	<0.5	0.5–10	10–25	>25			✓	✓
S (%)	<0.4	0.4–0.5	0.5–1	>1	✓		✓	

2. Materials and methods

2.1. Study area

Avilés and its surrounding area is one of the most important industrialized zones in north-west Spain. Various industries (mainly metallurgical) have been operating in this region for decades and their activities have severely impacted air, water, and soil quality (Gallego et al., 2002; Ordóñez et al., 2013). The study area is situated north of the town of Avilés, which sits on Cantabrian Sea coastline (Fig. 1). Industrial waste was commonly discharged into the Avilés estuary for many years and consequently considerable amounts of PTEs still remain. The main industrial activity in this area is metallurgical plants, namely: two Zn hydrometallurgical plants (the main one very close to the study areas), one Al production plant and one steel production plant.

Quantities and features of PTEs are dependent on a wide range of factors (Khalil et al., 2013; Kipp et al., 2009; Neiva et al., 2014), such as ore mineralogy and the distribution of trace and minor minerals in bulk minerals, among others (Gallego et al., 2002; Harvey et al., 2017; Luo et al., 2012). Several open-pit quarries related to industrial and mining activities are likewise placed in the survey region. Consequently, emissions of Pb, Ba and other PTEs are very common (Monaci and Bargagli, 1997), markedly affecting the quality of the surrounding soil/sediment.

2.2. Data collection and chemical analyses

Eight strategic polygons were defined for data gathering purposes (Fig. 1). The areas were chosen based on their geographical characteristics and predominant human activities: 1. Beach – leisure area, in direct contact with the Cantabrian Sea; 2. Landfill (soil of unknown origin); 3. Landfill Dunes (soils of unknown origin); 4. River beach; 5. Agriculture/Livestock zone (area with the highest altitude); 6. Background beach (reference area, as it is protected from atmospheric emissions and river discharge); 7. Llodero Cove; and 8. Zeluán protected ponds.

A total of 230 bed-deposited sediment samples were collected from the upper 0–20 cm beneath a water depth of nearly 20 cm. Samples were collected using a stratified systematic sampling method at short, medium, and long distances to produce a representative set showing the total variability of the chosen attributes. The samples were packed and sealed in pre-washed polyethylene bags and dried at room temperature. Inductively Coupled Plasma (Optical Emission Spectroscopy) was used for Chemical analyses after sample leaching by means of an Aqua Regia digestion (HCl + HNO_3). Fourteen elements were analyzed, namely Cu; Pb; Zn; Ag; Ni; Mn; Fe; As; Cd; V; Cr; Ti; Al and S. The content measured for these elements was after classified into four risk classes (1. Low; 2. Moderate; 3. High; 4. Remediation). For the threshold value definition, an extensive revision was carried out using different international sediment quality guidelines (SGGs) (Table 1). In this regard, the information used included the Canadian Council of Ministers of the Environment (2007) guidelines, the FAO and ISRIC guidelines for soil degradation in Central and Eastern Europe (Van Lynden, 2000), and the FOREGS European topsoil geochemistry database (http://www.gsf.fi/publ/foregsatlas/; Salminen et al., 2005). However, several authors who made relevant contributions to PTEs quantification were also considered, such as Rodríguez et al. (2008) and Moen et al. (1985), who assessed and standardized normal concentration ranges of PTEs, in soils affected by industrial activities. This approach leaded to a general framework for risk level classification.

2.3. Bayesian data analysis

Bayesian networks are directed acyclic graphs (DAG) where nodes and arcs typify the cause and effect relationships between variables (Pearl, 1986). The topological structure of a Bayesian model reflects the dependency of the variables and describes the probability distribution of certain events occurring in specific conditions. If $X = \{X_1, X_2, \ldots, X_n\}$ is a set of m-dimensional variables, then a BN is formally defined as a couplet $X = <G, P>$ where G is a DAG in which each node represents one of the variables $X_1, X_2, \ldots, X_n$ and each arc represents a direct dependency relationship between these variables; and P is a set of parameters that quantifies the network, containing the probabilities for each possible value x_i for each variable X_i.

From the decomposition theorem, the joint probability P, under the hypothesis that each node is independent of its non-descendants, can be calculated. Therefore, the Bayesian network has a single joint probability distribution is given by:

$$P(X) = P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i / X_{j(i)}\right) \quad (1)$$

where $X_{j(i)}$ is the set of parent variables of X_i for direct acyclic graph G. Consequently, application of Bayes' theorem enables to determine the posterior probability of the variable of interest through inference. In

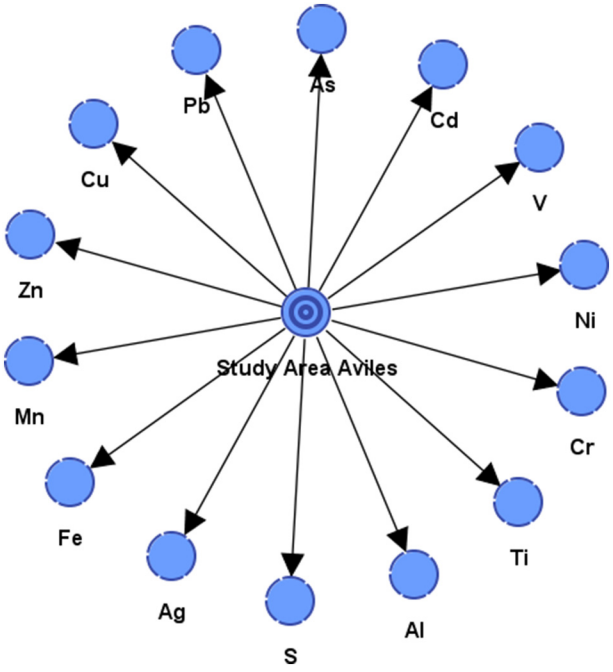


Fig. 2. Supervised learning network with Naive Bayes algorithm. The study area of Avilés is a variable (formed by the 8 defined strategic zones) that represents the network target node.

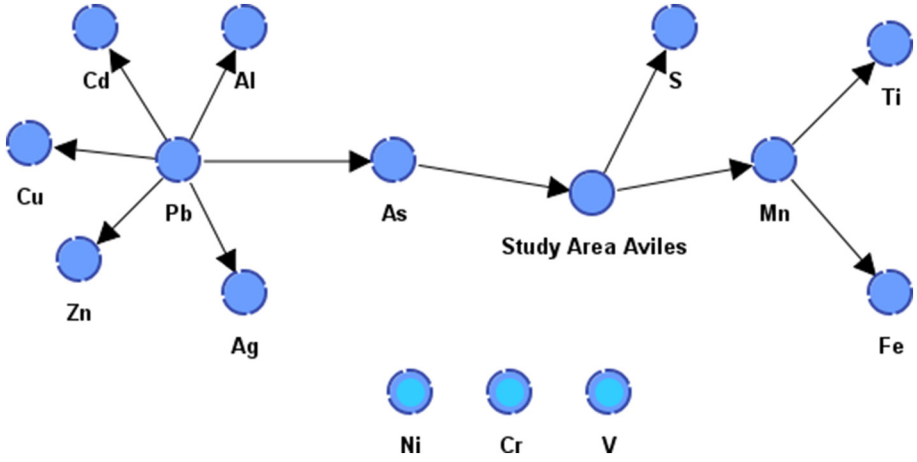


Fig. 3. Unsupervised learning network with Maximum Weight Spanning Tree algorithm. The light blue nodes (Ni, Cr and V) are uncorrelated with the main structure of the network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

this case, the variable of interest is the Study Area of Aviles. From the data collected, the aim is to infer the specific zones where PTEs exceed the concentrations established by regulations. Moreover, once trained, BNs can be used for intercausal reasoning (Druzdzel and Henrion, 1993). This opens the possibility to study whether the high concentrations of certain PTEs are linked up and how the presence of these elements can be mutually determined.

Aside from their popularity in solving complex and large-scale problems, several issues concerning the composition of the BNs that need to be stressed during the modeling process. When building a Bayesian network, it is necessary to explore distinct structures. A coherent analysis is crucial to ensure accurate characterization of phenomena (Kuhnert and Hayes, 2009). The computation of two different network learning procedures allowed to achieve the informational content maximization of the dataset.

2.3.1. Supervised network learning

In a first step, a supervised learning approach was used to independently generate a model to predict the target variable. In this approach, the only guidance provided is the node of interest, namely the study area, which is the target variable for the machine learning process. The zone, and hence the associated anthropogenic activities (agriculture, livestock, industrial and leisure), works as key factor for clarifying contamination sources.

For computation, Artificial Intelligence, and analytical software BayesiaLab v6.0.7 were used. A set of supervised learning algorithms was used to search for the optimal model.

The aim is to increase the probabilities of getting the optimal network for this environmental purpose. Given that the number of possible networks grows exponentially with the number of nodes (Friedman and Koller, 2003), this is a major challenge. To ensure that such a challenge does not become an intractable problem, it is necessary to use heuristic search algorithms to explore the search space to obtain a local optimum. Nevertheless, a single heuristic search algorithm does not assure to recover the global optimal. Therefore, BayesiaLab software algorithms were implemented progressively, from low space search strategies to high structural searches with more complex restrictions (Marcot, 2012), beginning with a Naive Bayes (e.g. Webb et al., 2005) straightforward network structure (Fig. 2) to evaluate up to what point the model supported higher complexity rates, e.g. Sons & Spouses or Markov Blanket (Conrady and Jouffe, 2015). This staggered procedure increases the probability of finding a solution closer to the global optimum, considering the adequate time and resources for learning

Given the importance of this previous step, the resulting network went through an efficiency test, in which the uncertainty reduction, provided by each variable, was analyzed. Shannon Entropy (Shannon, 1948) was used to compute the information exchanged between the target variable and any contaminant. The definition of Shannon Entropy of a discrete variable X is:

$$H(X) = -\sum_{x \in X} p(x)\log_2 p(x) \tag{2}$$

The difference between the marginal entropy of the target variable and the conditional entropy of a given target (predicted variable) is formally known as Mutual Information (Shannon, 1948) and denoted by *I*. For the study at hand, the Mutual Information between the study zone (target variable) and each element (heavy metal) is the respective marginal entropies. More generally, the Mutual Information between variables X and Y is defined by (Conrady and Jouffe, 2015):

$$I(X,Y) = H(X) - H(X|Y) \tag{3}$$

which is equivalent to:

$$I(X,Y) = \sum_{x \in X}\sum_{y \in X} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

The computation of the Mutual Information between the study zone and each predictor, in the form of contaminant, is represented by the Bayesian probability allocated to each class of the target variable. Thus, the predictors providing the maximum information can be properly identified, thereby highlighting their predictive importance and repercussion as zone contaminants.

Table 2
Dataset for BRI construction – Pb example for Zones 1, 2 and 3.

Zone	Geographical coordinates		Pb (mg/kg)	W_i	Risk Level	BRI_{Pb}
	X	Y				
1	264,646.8	4,830,389.9	22	0.726	1	1.726
1	264,567.3	4,830,342.4	135	0.110	2	2.110
1	264,550.5	4,830,335.5	141	0.137	3	3.137
1	264,320.9	4,830,163.4	38	0.726	1	1.726
2	264,264.3	4,830,242.7	24	0.726	1	1.726
2	264,257.1	4,830,242.5	21	0.726	1	1.726
2	264,220.0	4,830,309.4	18	0.726	1	1.726
2	264,602.7	4,830,392.4	170	0.137	3	3.137
3	264,354.5	4,829,971.1	41	0.726	1	1.726
3	264,356.3	4,829,986.9	37	0.726	1	1.726
3	264,266.6	4,830,618.5	715	0.027	4	4.027
3	264,444.0	4,830,571.8	160	0.137	3	3.137

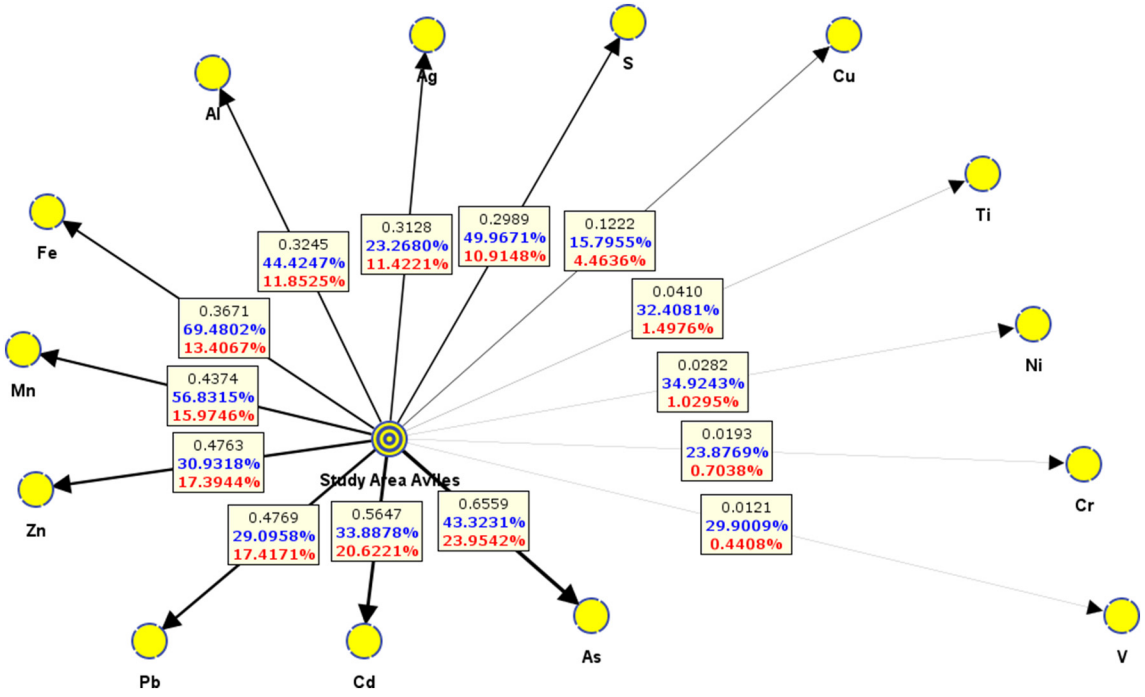


Fig. 4. PTEs Mutual information: progressive radial layout from the strongest to the weakest information node. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

2.3.2. Unsupervised network learning

In a second step, unsupervised structural learning was used to develop a model. This approach represents the purest form of knowledge as there are no constraints for the exploration of potential relationships between variables (Erhan et al., 2010). The Maximum Weight Spanning Tree algorithm (MWST) (e.g. Bazlamaçci and Hindi, 2001) was implemented (Fig. 3) and the Minimum Description Length (MDL) was used to determine the rate of attribute association. The MDL is a two-component score that must be minimized to obtain the best solution. It can be written as (Conrady and Jouffe, 2015):

$$\mathrm{MDL}(B, D) = \alpha \mathrm{DL}(B) + \mathrm{DL}(D|B) \quad (5)$$

where α represents the structural network coefficient, DL(B) is the number of bits to represent the Bayesian network B (graph and probabilities) and DL(B) is the number of bits to represent the dataset D given the Bayesian network B. This score quantifies the best trade-off between the two conditions. This is obtained by finding a point between the simplest structure, in which the network is fully unconnected, and the fully connected network, in which no structural independences are stated (Wong et al., 1999).

2.4. Bayesian Risk Index (BRI)

The BRI (Bayesian Risk Index) is fixed as a rating that reflects the soil risk level of contamination by PTEs. First, each chemical parameter was assigned a different weight (w_i-Bayesian Inference Weight). These weights were obtained through the inference process based on a Bayesian network procedure and, therefore, ranging between 0 and 1, where 0 represents the lowest probability of belonging to a determined risk class and 1 the highest probability (Fig. 2, Table 2). The BRI was calculated for each observed point as follows:

$$\mathrm{BRI} = \text{Risk class} + w_i \quad (6)$$

where w_i is the risk (class) conditional probability to a target zone (Table 1).

The new Bayesian Risk Index (BRI) is defined as a 'Regionalized Variable' (Matheron, 1971) and consequently additive by construction, since the mean value within a given observed support is equal to the arithmetic average of sample values, regardless of the statistical distribution of the values. This ensures that two samples with given profiles in the variable can be replaced by a new individual (Albuquerque et al., 2010; Rivoirard, 2005). The resulting scores correspond to the final index values, which range between 1 and 5. A subsequent geostatistical approach, aiming to construct the pattern of spatial risk, was used to define the areas in future need of monitoring and remediation.

2.5. Spatial modeling – geostatistical approach

The spatial probability patterns of the BRI were constructed following a three-step geostatistical modeling methodology:

1) Selected attributes went through a structural analysis and experimental variograms were computed. The variogram is a vector function used to compute the spatial variation structure of regionalized variables (Matheron, 1971; Journel and Huijbregts, 1978).

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \quad (7)$$

Its argument is h (distance) where $Z(x_i)$ and $Z(x_i + h)$ are the numerical values of the observed variable at points x_i, and $x_i + h$. The number of forming pairs for a h distance is $N(h)$. Therefore, it is the average value of the square of the differences between all couples of points existing in the geometric field spaced at a h distance (Journel and Huijbregts, 1978). The graphic behavior study of the variogram provides an overview of the spatial structure of the variable. One of the parameters that provide such information is the nugget effect (C_o), which shows the behavior at the origin. The other two parameters are the sill (C_1) and the amplitude (a) which define correspondently the inertia

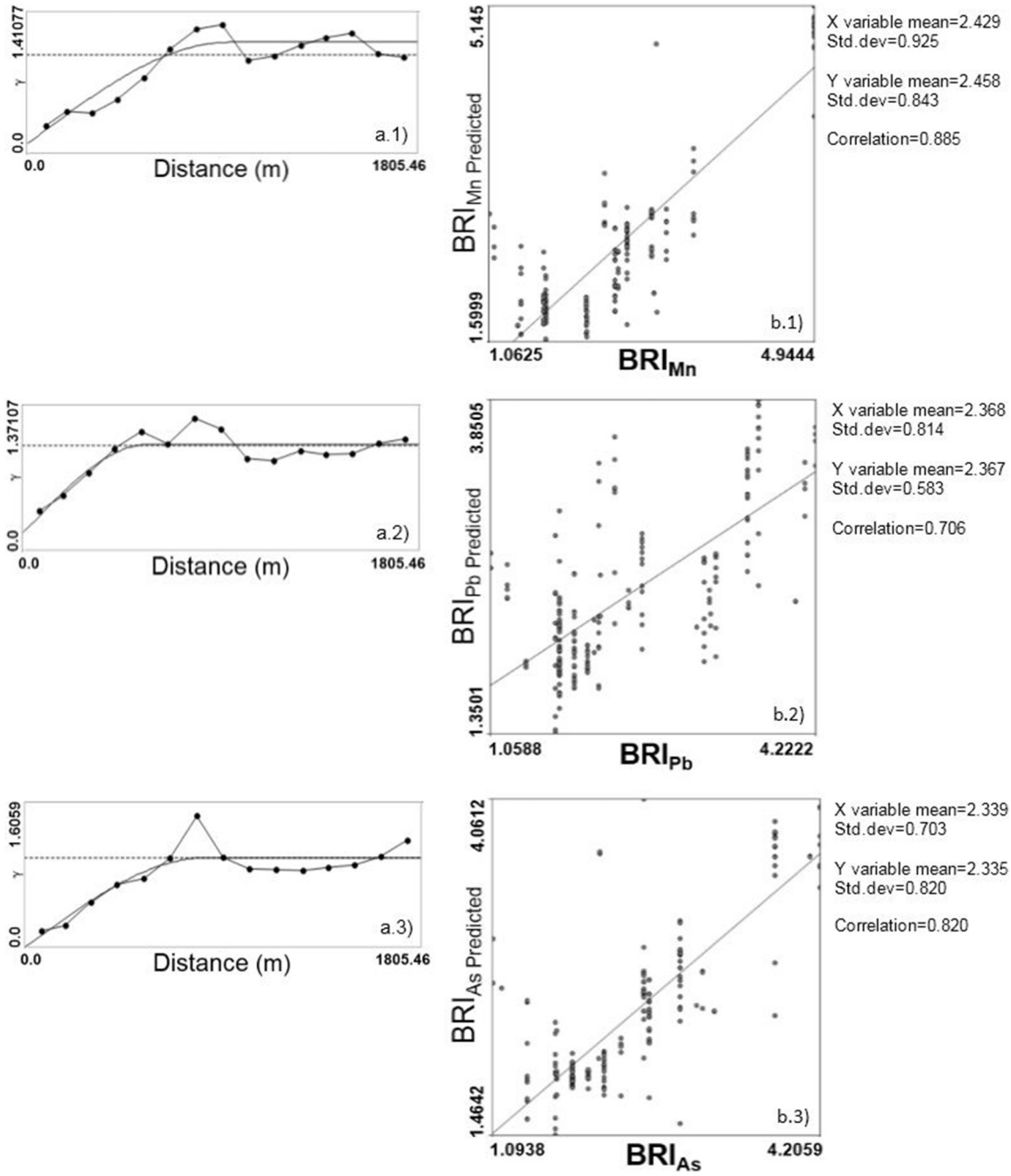


Fig. 5. (a) Isotropic variograms and fitted models for: a.1) BRIMn; a.2) BRIPb and a.3) BRIAs; (b) Scatterplots of measured BRIs versus predicted ones computed by linear regression: b.1) BRIMn; b.2) BRIPb and b.3) BRIAs.

used in the interpolation process and the influence radius of the variable.

2) Sequential Gaussian Simulation (SGS) was used as stochastic simulation algorithm. SGS starts by defining the univariate distribution of values, performing a normal score transform of the original values to a standard normal distribution. Normal scores at grid node locations were simulated sequentially with simple kriging (SK) using the normal score data and a zero mean (Goovaerts, 1997). Once all normal scores had been simulated, they were back-transformed to original grade values. For the computation, the Space-Stat Software V. 4.0.18, Biomedwere, was used (Albuquerque et al., 2014). The outcome of a simulation is a twisted version of an estimation process, which reproduces the statistics of the known data, making a realistic look of the exemplar, but providing a low prediction behavior. If a multiple sequence of simulation is designed, it is possible to obtain more reliable probabilistic maps;
3) Finally, Local G clustering allowed measurement of the degree of association that results from the concentration of weighted points (or region represented by a weighted point) and all other weighted points included within a radius of distance from the original

Table 3
Variogram parameters for the fitted isotropic models.

Model		C_0	C_1	a (m)
BRI_{As}	Spherical	0	1,61 (100%)	1200
BRI_{Mn}	Spherical	0.098	1.015 (91.2%)	950
BRI_{Pb},	Spherical	0.165	1.21 (88%)	1000

weighted point. Considering a given area subdivided into n regions, $i = 1, 2, \ldots, n$, where each neighborhood is distinguished with a point whose Cartesian coordinates are known. Each i has associated with it a value x (a weight) taken from a variable X. The variable has a natural origin and is positive. The G(i) statistic developed below allows for tests of hypotheses about the spatial concentration of the sum of x values associated with the j points within d of the ith point. The following statistic is obtained:

$$G_i(d) = \frac{\sum_{j=1}^{n} W_{ij}(d)x_i}{\sum_{j}^{n} x_j} \quad (8)$$

where *Wij* is a symmetric one/zero spatial weight matrix with ones for all links defined as being within distance *d* of a given *i*; all other links are zero, including the link of point *i* to itself. The numerator is the sum of all *xj* within d of *i* but not including *xi*. The denominator is the sum of all *xj*, excluding *xi* (Getis and Ord, 1992).

3. Results and discussion

3.1. Bayesian results

Bayesian analysis has allowed to identify Naive Bayes in Fig. 2 as the best supervised learning predictor (Qi and Zhu, 2003). Given the nature of the problem, this algorithm makes probabilistic inference much faster and clearer than other proven supervised ones. However, when seeking a greater understanding of the spatial patterns of the elements, the unsupervised learning network with MWST (Fig. 3) showed better performance (Erhan et al., 2010; Marcot, 2012).

Mutual Information analysis is shown in Fig. 4. The top number in the box represents the information exchanged between each node and the target node. This is represented graphically in accordance with the thickness of the arc and the distance to the target node. This is a symmetric measure, as such, the amount of information that e.g. As provides on each zone in the study area is the same as the amount of information that each zone provides about As.

However, without context, the number might not be meaningful. Hence, two additional measures based on Shannon Entropy calculations

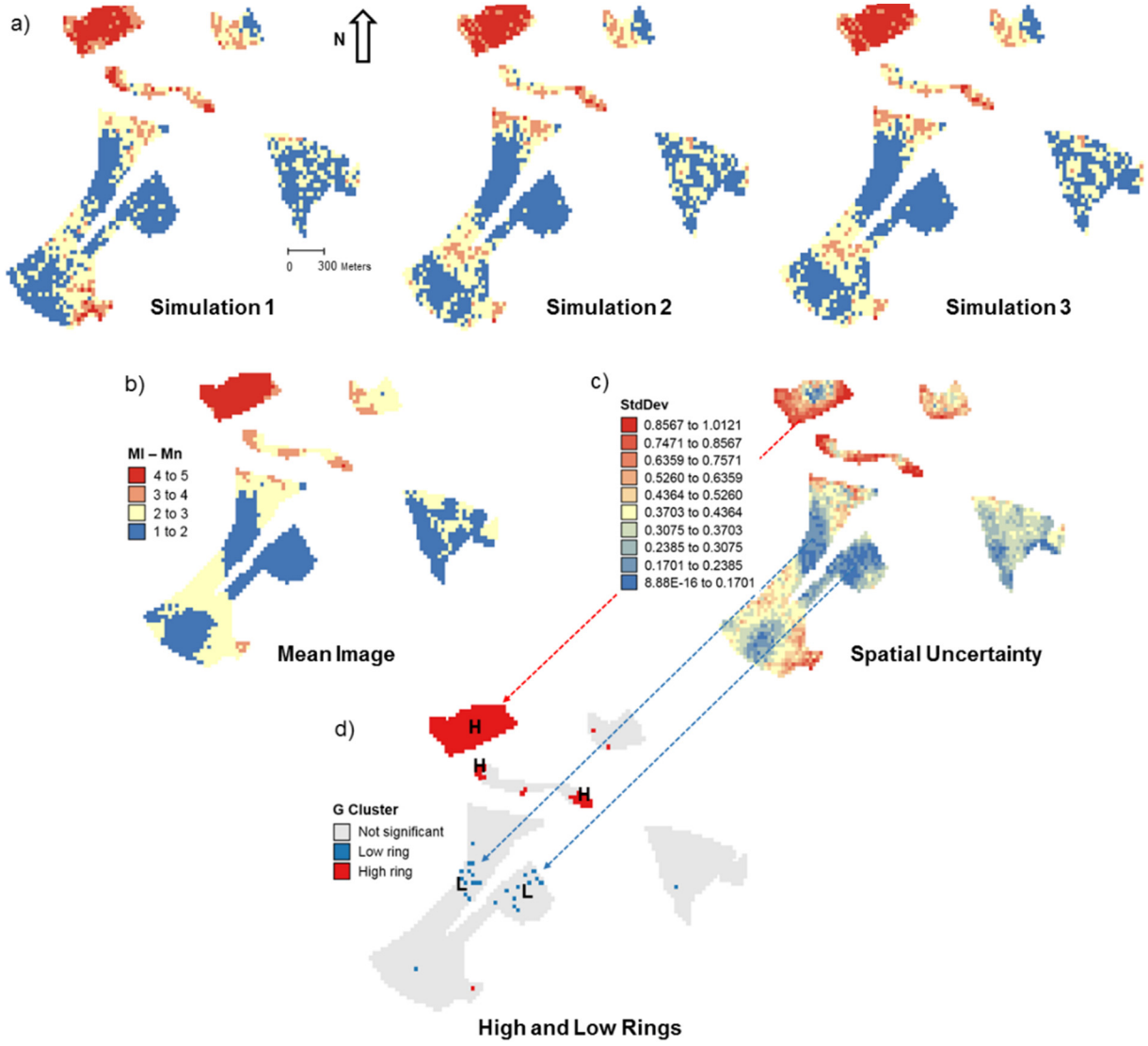


Fig. 6. BRIMn: a) 3 SGS scenario; b) Mean Image (MI); c) significant G clusters of low (L) and high (H) and d) Spatial Uncertainty (StdDev) – blue arrow indicating low spatial uncertainty associated with the significant cluster and red arrow indicating moderate to high spatial uncertainty associated with the significant cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

were computed (Conrady and Jouffe, 2015; Shannon, 1948). The blue number in the boxes (Fig. 4) shows the relative mutual information about the child node. Conversely, the red number shows the relative mutual information regarding the parent node. Thus, knowing e.g. the zone of the study area, the uncertainty regarding As is reduced by 43.32% on average. On the other hand, knowing the As threshold allows uncertainty linked to the zone to be reduced by approximately 23.95%.

The knowledge of the all nodes allows identification of the most predictive ones regarding whether a specific zone is contaminated or not. Consequently, As, Cd, Pb, Zn and Mn can be identified as the key PTEs in the study area (Fig. 4).

These results are supplemented by the unsupervised learning Bayesian network (Fig. 3) (Erhan et al., 2010). Thus, V, Ni and Cr were uncorrelated to the network structure, since these elements showed the least information exchange with the study zones (Fig. 4). Likewise, Pb, As and Mn were key attributes as they belonged to the main structure of the BN and thus had high predictive importance. Finally, Pb explained the spatial distribution of Cd, Al, Cu, Zn and Ag, while Mn explains together Ti and Fe spatial distribution (Fig. 3).

The Bayesian results allowed a greater understanding of the spatial patterns shown by the elements, offering at the same time a reduction in the dataset dimensionality from 14 to 3 attributes. In addition, the Bayesian network framework developed proved to be a suitable modeling approach to cover the informational needs and relationships required in later geostatistical techniques.

3.2. Spatial patterns and spatial uncertainty

In a first step, experimental variograms for each selected BRI (BRI_{Pb}, BRI_{As} and BRI_{Mn}) were computed for variable structural characterization. No clear evidence of anisotropies was found, and isotropic variograms were computed and corresponding models were fitted. The quality of the model of uncertainty provided by simple kriging (SK) (zero mean) was assessed using the same source and destination geography approach, whereby SK results at sampled locations u_α were compared to observations. Correlation indices ranged between 0.70 and 0.88 (Fig. 5). Therefore, cross-validation results were considered satisfactory for the selected models, thereby indicating consistency between the estimated and observed values.

The graphic behavior of the variogram function provides an overview of the spatial variation structure of the variable (Chica, 2005). One of the parameters that provide such information is the nugget effect

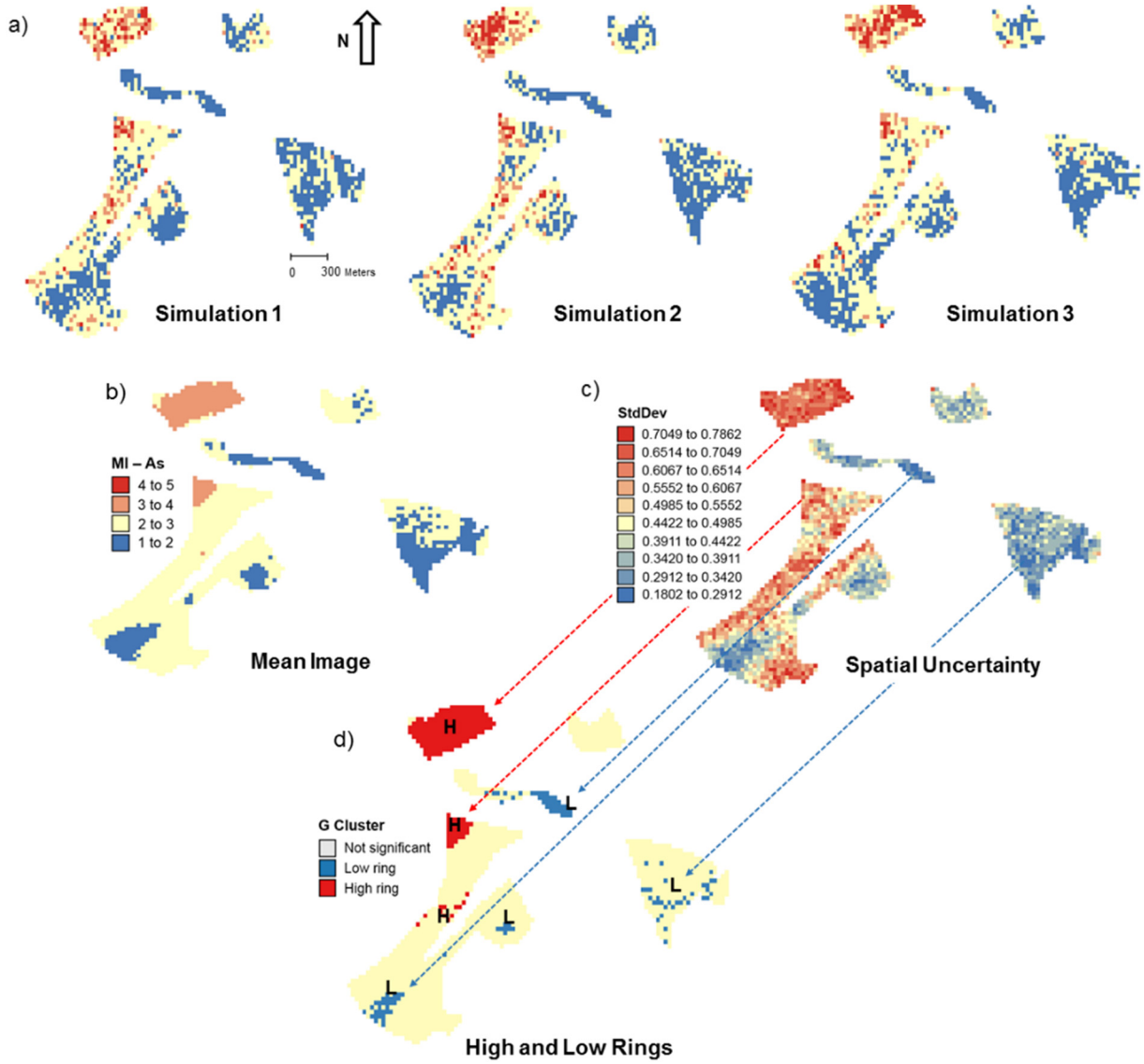


Fig. 7. BRIAs: a) 3 SGS scenario; b) Mean Image (MI); c) significant G clusters of low (L) and high (H) and d) Spatial Uncertainty (StdDev) – blue arrow indicating low spatial uncertainty associated with the significant cluster and red arrow indicating moderate to high spatial uncertainty associated with the significant cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

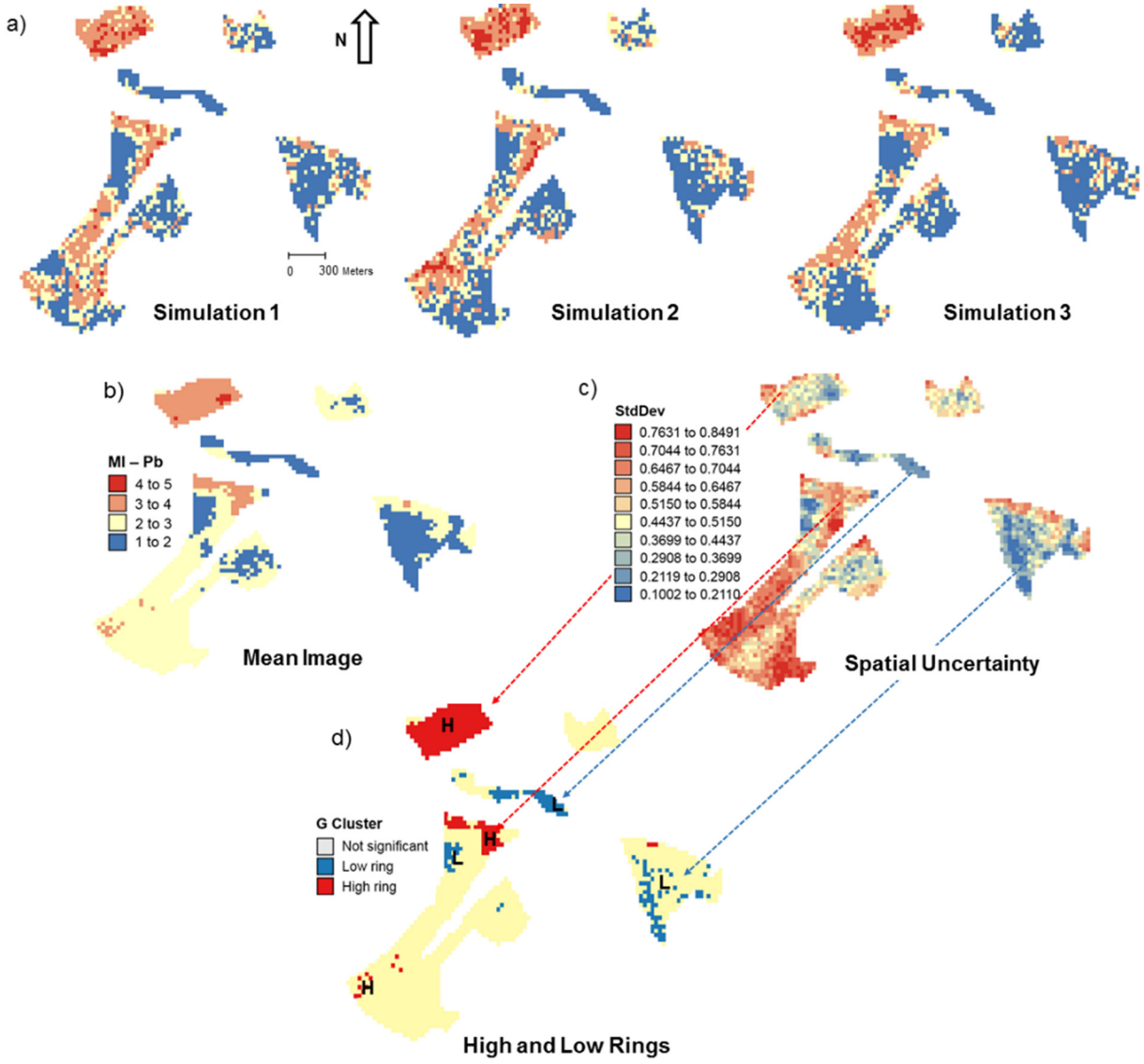


Fig. 8. BRIPb: a) 3 SGS scenario; b) Mean Image (MI); c) significant G clusters of low (L) and high (H) and d) Spatial Uncertainty (StdDev) – blue arrow indicating low spatial uncertainty associated with the significant cluster and red arrow indicating moderate to high spatial uncertainty associated with the significant cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(C_0), which shows the behavior at the origin (Pereira et al., 1993). The other two parameters are the sill (C_1) and the Range (a), which define correspondently the inertia used in the interpolation process and the variable structure influence zone (Table 3).

A hundred simulations were performed using Sequential Gaussian Simulation (SGS) as a conditional stochastic simulation of the BRI value distribution for Pb, As and Mn (BRI_{Pb}; BRI_{As} and BRI_{Mn}). The calculation of the spatial uncertainty - the Standard Deviation of each pixel - set aside the discussion of local accuracy and, finally, Local G clustering identified the BRI high-rings for the selected elements in the subject area.

Geostatistics allows finding answers to problems with space–time indexation (Kyriakidis and Journel, 1999). A stochastic SGS model on a 100 × 100 m grid was used to generate 100 equiprobable scenarios. The realization numbers 1, 15 and 99 in BRI_{Mn}, BRI_{As} and BRI_{Pb} are shown in Figs. 6a, 7a and 8a, respectively. However, the issue is that no single realization can be taken as a better representation of reality than any other, and the mean spatial images (MI) - average maps - are afterwards used to assess the spatial pattern of each variable (Figs. 6b, 7b and 8b) while the spatial variability images (standard deviation maps) allow quantification of spatial uncertainty for each attribute (BRI_{Mn}, BRI_{Pb}, and BRI_{AS}) (Figs. 6c, 7c and 8c). The characterization of aggregates of Low and High risk, for Mn, As and Pb contamination, was achieved using the MI maps and Local G clustering (Getis and Ord, 1992).

The spatial patterns shown and the computed clusters (Figs. 6, 7 and 8) allowed classification of zones 1, 4 and 5 as "hot-spots" for Mn, As and Pb, corresponding to leisure (zones 1 and 4) and agricultural/livestock (zone 5) activities, respectively, and the "cold-spots" (clean zones) as those overlapping the protected ponds of Zeluán and Llodero Cove (zones 7 and 8).

The area selected for background reference (zone 6) shows a moderate risk of PTEs contamination, thereby indicating that the contamination is spreading towards the north-east of the region and thus affecting a larger area than what was previously assumed. Future monitoring sampling must be implemented throughout the north-east of the study area for further clarification.

The landfill areas (zones 2 and 3) show inverse patterns, which may indicate recent remediation action concerning zone 3 (Low cluster) and the need of removal for zone 2 (High cluster). Low to moderate spatial uncertainty (Figs. 6c, 7c and 8c) is generally associated with the prediction scenarios, indicating accurate representations for risk contamination with trace elements.

Finally, it is important to stress that the high altitude of zone 5 may reflect atmospheric emissions of industrial plants as responsible for PTE deposition. Future monitoring actions must be implemented in this zone, and a set of climatic soft covariates considered (e.g. wind preferential directions) for risk modeling purposes.

4. Conclusions

Avilés and its surroundings is one of the most important industrialized zones in north-west Spain. In the herein study a set of 14 chemical elements, gathered in eight different zones, to compute Bayesian Network structures to analyze how high concentrations of PTEs linked up and how their presence can be mutually influenced. From this point, a simple, direct index (Bayesian Risk Index - BRI) for soil contamination assessment was developed as a linear rating of pre-defined risk classes weighted by the previously obtained Bayesian probabilities. Within this framework, Pb, As and Mn were found to be the key attributes, as they belong to the main structure of the Bayesian network and, therefore, having high predictive importance. The spatial probability patterns of the BRI for the three key elements (BRI_{Pb}, BRI_{As} and BRI_{Mn}) were obtained through a 3-step geostatistical modeling methodology. A hundred simulations were performed using Sequential Gaussian Simulation (SGS) as conditional stochastic simulation algorithm, and Local G clustering was used to identify BRI_{Pb}, BRI_{As} and BRI_{Mn} high-rings. The high-altitude zone 5 may indicate atmospheric emissions by industrial plants as sources of PTE depositions. Future monitoring actions must be carried out and a set of climatic soft covariates, considered for modeling purposes, such as e.g. wind preferential directions.

Acknowledgments

Carlos Sierra obtained a grant from the “Severo Ochoa (BP10-112)” Programme (Ficyt, Asturias, Spain).

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version, at http://dx.doi.org/10.1016/j.scitotenv.2017.06.068. These data include the Google map of the most important areas described in this article.

References

Aguilera, P.A., Fernández, A., Fernández, R., Rumí, R., Salmerón, A., 2011. Bayesian networks in environmental modelling. Environ. Model. Softw. 26, 1376–1388.
Albuquerque, M.T.D., Dias, V.H., Poellinger, N., Pinto, J.F., 2010. Construction of a quality index for granules produced by fluidized bed technology and application of the correspondence analysis as a discriminant procedure. Eur. J. Pharm. Biopharm. 75 (3). http://dx.doi.org/10.1016/j.ejpb.2010.04.002.
Albuquerque, M.T.D., Antunes, I.M.H.R., Seco, M.F.M., Roque, N., Sanz, G.L., 2014. Uranium and arsenic spatial distribution in the Águeda watershed groundwater. Procedia Earth Planet. Sci. 8, 13–17.
Antunes, I.M.H.R., Albuquerque, M.T.D., 2013. Using indicator kriging for the evaluation of arsenic potential contamination in an abandoned mining area (Portugal). Sci. Total Environ. 442:545–552. http://dx.doi.org/10.1016/j.scitotenv.2012.10.010.
Bazlamaçci, C.F., Hindi, K.S., 2001. Minimum-weight spanning tree algorithms. A survey and empirical study. Comput. Oper. Res. 28, 767–785.
Bién, J.D., Ter Meer, J., Rulkens, W.H., Rijnaarts, H.H.M., 2005. A GIS-based approach for the long-term prediction of human health risks at contaminated sites. Environ. Model. Assess. 9, 221–226.
CCME, 2007. Canadian Soil Quality Guidelines for the Protection of Environmental and Human Health: Summary Tables. Environment Canada, National Guidelines and Standards Office (No. 1299, ISBN 1-896997-34-1).
Chica, M., 2005. La Geoestadística como herramienta de análisis espacial de datos de inventario forestal. Actas de la I reunión de inventario y teledetección forestal. Cuad. Soc. Esp. Cienc. For. 19, 47–55.
Conrady, S., Jouffe, L., 2015. Bayesian Networks & BayesiaLab - A Practical Introduction for Researches. Franklin, TN, Bayesia USA (ISBN:978-0-9965333-0-0).
Davies, A.J., Hope, M.J., 2015. Bayesian inference-based environmental decision support systems for oil spill response strategy selection. Mar. Pollut. Bull. 96, 87–102.
Druzdzel, M.J., Henrion, M., 1993. Intercausal Reasoning with Uninstantiated Ancestor Nodes. Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence pp. 317–325.
Elsheikh, A.H., Jackson, M.D., Laforce, T.C., 2012. Bayesian reservoir history matching considering model and parameter uncertainties. Math. Geosci. 44, 515–543.
Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., Bengio, S., 2010. Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. 11, 625–660.
Friedman, N., Koller, D., 2003. Being bayesian about network structure. A bayesian approach to structure discovery in bayesian networks. Mach. Learn. 50, 95–125.
Gallego, J.R., Ordóñez, A., Loredo, J., 2002. Investigation of trace element sources from an industrialized area (Avilés, northern Spain) using multivariate statistical methods. Environ. Int. 27, 589–596.
Gerstenberger, M.C., Christophersen, A., Buxton, R., Nicol, A., 2015. Bi-directional risk assessment in carbon capture and storage with Bayesian networks. Int. J. Greenhouse Gas Control 35, 150–159.
Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. Geogr. Anal. 24, 189–206.
Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. University Press, New York: Oxford.
Harvey, P.J., Rouillon, M., Dong, C., Ettler, V., Handle, H.K., Taylor, M.P., Tyson, E., Tennant, P., Telfer, V., Trinh, R., 2017. Geochemical sources, forms and phases of soil contamination in an industrial city. Sci. Total Environ. 584–585, 505–514.
Jiang, Y., Song, Z., Kusiak, A., 2013. Very short-term wind speed forecasting with Bayesian structural break model. Renew. Energy 50, 637–647.
Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. Academic Press, San Diego.
Khalil, A., Hanich, L., Bannari, A., Zouhri, L., Pourret, O., Hakkou, R., 2013. Assessment of soil contamination around an abandoned mine in semi-arid environment using geochemistry and geostatistics: pre-work of geochemical process modeling with numerical models. J. Geochem. Explor. 125, 117–129.
Kipp, G.G., Stone, J.J., Stetler, L.D., 2009. Arsenic and uranium transport in sediments near abandoned uranium mines in Harding County, South Dakota. Appl. Geochem. 24, 2246–2255.
Kuhnert, P.M., Hayes, K.R., 2009. How Believable is your BBN? In: 18th World IMACS/MODISM Congress. Cairns, Australia.
Kyriakidis, P.C., Journel, A.G., 1999. Geostatistical space–time models: a review. Math. Geol. 31, 651–684.
Lahr, J., Kooistra, L., 2010. Environmental risk mapping of pollutants: state of the art and communication aspects. Sci. Total Environ. 408:3899–3907. http://dx.doi.org/10.1016/j.scitotenv.2009.10.045.
Li, P., Lin, C., Cheng, H., Duan, X., Lei, K., 2015. Contamination and health risks of soil heavy metals around a lead/zinc smelter in south western China. Ecotoxicol. Environ. Saf. 113, 391–399.
Luo, X., Yu, S., Zhu, Y., Li, X., 2012. Trace metal contamination in urban soils of China. Sci. Total Environ. 421–422, 17–30.
Marcot, B.G., 2012. Metrics for evaluating performance and uncertainty of Bayesian network models. Ecol. Model. 230, 50–62.
Matheron, G., 1971. The Theory of Regionalized Variables and Its Applications. Les Cahiers du Centre de Morphologie Mathématique, no. 5, Ecole des Mines de Paris (211 p).
McDonald, K.S., Ryder, D.S., Tighe, M., 2015. Developing best-practice Bayesian belief networks in ecological risk assessments for freshwater and estuarine ecosystems: a quantitative review. J. Environ. Manag. 154, 190–200.
Moen, J., Ale, B.J.M., 1998. Risk maps and communication. J. Hazard. Mater. 61, 271–278.
Moen, J.E.T., Cornet, J.P., Evers, C.W.A., 1985. Soil Protection and Remedial Actions: Criteria for Decision Making and Standardization of Requirements. Proc. 1st INt. TNO Conf. on Contaminated Soil, Utrecht, Netherdlands. Martinus Nijhoff, Dordrecht pp. 441–448.
Monaci, F., Bargagli, R., 1997. Barium and other trace metals as indicators of vehicle emissions. Water Air Soil Pollut. 100, 89–98.
Moreno-Jiménez, E., García-Gómez, C., Oropesa, A.L., Esteban, E., Haro, A., Ramón Carpena-Ruiz, R., Tarazona, J.V., Peñalosa, J.M., Fernández, M.D., 2011. Screening risk assessment tools for assessing the environmental impact in an abandoned pyritic mine in Spain. Sci. Total Environ. 409:692–703. http://dx.doi.org/10.1016/j.scitotenv.2010.10.056.
Neiva, A.M.R., Carvalho, P.C.S., Antunes, I.M.H.R., Silva, M.M.V.G., Santos, A.C.T., Pinto, M.M.S.C., Cunha, P.P., 2014. Contaminated water, stream sediments and soils close to the abandoned Pinhal do Souto uranium mine, central Portugal. J. Geochem. Explor. 136, 102–117.
Nolan, B.T., Fienen, M.N., Lorenz, D.L., 2015. A statistical framework for groundwater nitrate models of the Central Valley, California, USA. J. Hydrol. 531, 902–911.
Ordóñez, C., Sierra, C., Albuquerque, T., Gallego, J.R., 2013. Functional data analysis as a tool to correlate textural and geochemical data. Appl. Math. Comput. 223:476–482. http://dx.doi.org/10.1016/j.amc.2013.08.032.
Pearl, J., 1986. Fusion, propagation, and structuring in belief networks. Artif. Intell. 29, 241–288.
Pereira, H.G., Brito, M.G., Albuquerque, T., Ribeiro, J., 1993. Geostatistical Estimation of a Summary Recovery Index for Marble Quarries. Geostatistics Troia'92 5. Kluwer Academic Publishers, pp. 1029–1040.
Phan, T.D., Smart, J.C.R., Capon, S.J., Hadwen, W.L., Sahin, O., 2016. Applications of Bayesian belief networks in water resource management: a systematic review. Environ. Model. Softw. 85, 98–111.
Qi, F., Zhu, A., 2003. Knowledge discovery from soil maps using inductive learning. Int. J. Geogr. Inf. Sci. 17, 771–795.
Rivoirard, J., 2005. Concepts and methods of geostatistics. Space, structure and randomness. In: Meyer, F., Schmitt, M. (Eds.), Contributions in Honor of Georges Matheron in the Fields of Geostatistics, Random Sets and Mathematical Morphology, Bilodeau (ISBN: 978-0-387-20331-7).

Rodríguez, L.L., Hengl, T., Reuter, H.I., 2008. Heavy metals in Europe soils: a geostatistical analysis of the FOREGS geochemical database. Geoderma 148, 189–199.
Salminen, R., Batista, M.J., Bidovec, M., Demetriades, A., De Vivo, B., De Vos, W., Duris, M., Gilucis, A., Gregorauskiene, V., Halamic, J., Heitzmann, P., Lima, A., Jordan, G., Klaver, G., Klein, P., Lis, J., Locutura, J., Marsina, K., Mazreku, A., O'Connor, P.J., Olsson, S.Å., Ottesen, R.T., Petersell, V., Plant, J.A., Reeder, S., Salpeteur, I., Sandström, H., Siewers, U., Steenfelt, A., Tarvainen, T., 2005. FOREGS Geochemical Atlas of Europe. Methodology and Maps (Part 1, pp. 526 and Part 2, pp. 690).
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27: 379–423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x.
Taalab, K., Corstanje, R., Zawadzka, J., Mayr, T., Whelan, M.J., Hannam, J.A., Creamer, R., 2015. On the application of Bayesian networks in digital soil mapping. Geoderma 259-260, 134–148.
Van Lynden, G.W.J., 2000. Guidelines for the assessment of soil degradation in Central and Eastern Europe. Report 97/08b, Revised edition FAO and ISRIC, Wageningen.
Webb, G.I., Boughton, J.R., Wang, Z., 2005. Not so Naive Bayes: aggregating one-dependence estimators. Mach. Learn. 58, 5–24.
Wong, M.L., Lam, W., Leung, K.S., 1999. Using evolutionary programming and minimum description length principle for data mining of Bayesian networks. IEEE Trans. Pattern Anal. Mach. Intell. 21, 174–178.
Woodbury, P.B., 2003. DOs and DON'Ts of spatially explicit ecological risk assessment. Environ. Toxicol. Chem. 22, 977–982.
Wu, J., Xu, S., Zhou, R., Qin, Y., 2016. Scenario analysis of mine water inrush using Bayesian networks. Saf. Sci. 89, 231–239.
Zambon, I., Colantoni, A., Carlucci, M., Morrow, N., Sateriano, A., Salvati, L., 2017. Land quality, sustainable development and environmental degradation in agricultural districts: a computational approach based on entropy indexes. EIA Rev. 64:37–46. http://dx.doi.org/10.1016/j.eiar.2017.01.003.
Zhang, D., Yan, X.P., Yang, Z.L., Wall, A., Wang, J., 2013. Incorporation of formal safety assessment and Bayesian network in navigational risk estimation of the Yangtze River. Reliab. Eng. Syst. Saf. 118, 93–105.
Zuo, X., Hua, H., Dong, Z., Hao, C., 2017. Geographic environmental performance index at the provincial level for China 2006–2011. Ecol. Indic. 75:48–56. http://dx.doi.org/10.1016/j.ecolind.2016.12.016.