Journal of Geochemical Exploration 279 (2025) 107868

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

Improving spatial interpolation for anomaly analysis in presence of sparse, clustered or imprecise data sets

Stéphane Belbèze [a], Jérémy Rohmer [a], Dominique Guyonnet [a], Philippe Négrel [a,*], Timo Tarvainen [b]

[a] BRGM, 45060 Orléans, France
[b] Geological Survey of Finland, GTK, Vuorimiehentie 5, P.O. Box 96, FI-02151 Espoo, Finland

ARTICLE INFO

Keywords:
European geochemistry
Sparse, clustered
Spatial interpolation
Anomaly detection
Uncertainty

ABSTRACT

In this study, we present a new method of interpolation and anomaly detection especially designed for sparse, clustered or imprecise environmental data (SIC). Such data cannot be processed by current state of the art spatial methods and models, including the most widely used, such as kriging. Indeed, the statistics obtained on SIC data (on the order of 5–30) do not allow us to define a covariance or to calibrate the numerous hyper-parameters of sophisticated Bayesian or deep image prior models. We therefore adapted an information dissemination algorithm to handle SIC data. This probabilistic model has been enriched (anisotropy, de-clustering, auto-variography, multi-support, treatment of covariates, and censored data) in a way that fully meets the needs for environmental SIC data and can be used in conjunction with hybrid propagation of epistemic and aleatoric uncertainties and anomaly detection, whatever their mathematical form. The new interpolator for anomaly detection was applied on a very small set of 13 sparse data points characteristic of small-scale environmental studies, on digital-challenge datasets and on two real datasets, i.e., a large-scale geochemical dataset and a SIC urban soil dataset. Results highlight the added value of the proposed algorithm, that is able to pinpoint anomalies in SIC data, while avoiding in particular the smoothing effects of certain previous methods.

1. Introduction

Data collected in the context of geochemical surveys for mineral exploration (i.e., data indicative of the presence of mineralization) or for environmental studies (e.g., soil quality that may represent a risk for human health or for ecosystems) is often used to produce maps. Regarding mineral exploration, such maps can help to identify, e.g., geochemical anomalies or drilling locations for further investigations. In an environmental context, the maps can serve to highlight contamination anomalies, to prioritize remedial action, or to identify pedo-geochemical background values (Belbèze et al., 2023). Drawing maps requires the interpolation of values at locations where there have not been any measurements. But this task can be seriously complicated by the fact that the data may be SIC, i.e., Ssparse (small number of observations) or Imprecise (data subject to measurement errors) or Clustered (heterogeneously distributed). This SIC character can also be found in datasets which are not “sparse” per se but which cover larger scales than the size of the anomalies or objects of interest; e.g., in geochemical surveys over hundreds of km^2, in chemical monitoring studies of urban soils at the scale of a few km^2, or in potentially contaminated site studies (scale of a few hm^2).

Spatial interpolation requires a solution to a complex problem that estimates a value from observed values. There exists a plethora of interpolation methods, all of which have their specific advantages and drawbacks, especially in presence of SIC data. In this study, we present a new method of interpolation and anomaly detection especially designed for such data. The underlying hypotheses of previous interpolation methods, including the widely known kriging method, are often ill-suited to SIC data. Indeed, the statistics obtained on SIC data do not allow the definition of a covariance or to calibrate the numerous hyper-parameters of sophisticated Bayesian or deep image prior models. As geostatistical and hybrid machine learning methods cannot be calibrated with small datasets (on the order of 5–30), non-geostatistical methods are usually used, but the latter do not convey uncertainties associated with, e.g., values below the detection limit (Belbèze et al., 2023; Li and Heap, 2008, 2011; Li, 2012). We therefore adapted an

[*] Corresponding author.
E-mail address: p.negrel@brgm.fr (P. Négrel).

https://doi.org/10.1016/j.gexplo.2025.107868
Received 21 March 2025; Received in revised form 16 July 2025; Accepted 16 July 2025
Available online 21 July 2025

Table 1
Characteristics of key methods used to map European soils.

Encountered approaches	Description	Example
No maps – dot plots only	These authors do not produce interpolated maps because the variability observed in their data excludes this type of treatment or because the density of measuring points does not allow it (Reimann et al., 2008; Rhind et al., 2013).	Reimann et al., 2018; Négrel et al., 2019
Inverse Distance Weighting (IDW)	Inverse Distance Weighting (IDW) is based on the intuitive notion that nearby points have more influence than far-away points. IDW is known to respect the data and any anomalies on the interpolation grid (Grunsky and De Caritat, 2017). It is by far the most widely used interpolation method for geochemical backgrounds.	Négrel et al., 2015
The nearest neighbor algorithm (NN) and Triangulated Irregular Network (TIN)	TIN triangulation is an algorithm that uses Delaunay triangles. It creates triangular surfaces between close neighbors and propagates the contents linearly along the facets of the triangle. These methods are exact interpolators and do not extrapolate. They perform rapidly for densely sampled areas. NN or TIN type of interpolation are useful because they do not smooth out the data and allows for rapid visualization of the studied phenomena trends in an implicitly more precise way if the points are close together and in an imprecise way elsewhere. The choice of a TIN method assumes that the physical phenomenon under consideration consists of a linear trend to which a fluctuating error of small amplitude is added.	Jordan et al., 2018
Simple kriging (K)	Kriging interpolation (Chilès and Delfiner, 2013) is similar in its general form to the IDW, but differs in the way the weights are calculated. While IDW uses an inverse distance determined covariance function, kriging assumes that the data is regionalized (strong hypothesis), uses an expert driven covariance function and tends to eliminate local anomalies. However, the map is still very informative and highlights trends. This process is consistent with the underlying idea of smoothly varying geochemical concentrations.	Tarvainen et al., 2013; Reimann et al., 2014a, 2014b
Kriging with external drift (KED)	Kriging with external drift (KED) has been used by the major European mapping projects FOREGS and LUCAS, as well as for countries which use geostatistical methods of the Paris School of Mines; i.e., France, Australia, Belgium, and Algeria. The methods for establishing drifts before KED are varied and range from simple linear regression to the most advanced partitioning methods (such as multinomial logistic regression, C5 decision tree, and random forest). The book *Digital Soil Mapping* from the Sydney Institute for Agriculture (Malone et al., 2017) is a reference for the implementation of the KDE method, that runs on the R platform.	Tarvainen et al., 2005; Lado et al., 2008; Tóth et al., 2013; Tóth et al., 2016; Heuvelink et al., 2016; Pereira et al., 2012; Maas et al., 2010
Multilevel B-splines with external drift (MBSDE)	In multilevel B-splines with external drift, the MBS performs as well as kriging, but it is computationally faster. The methods for establishing drifts before MBS are the same as for KED. For the LUCAS project, a Cubist model was used as a drift.	Panagos et al., 2014
Quantile Regression Forest (QRF)	Random forest (RF; Breiman, 2001) and its extension quantile regression forest (QRF) (Meinshausen, 2006), are interesting and versatile machine learning algorithms for digital soil mapping. The QRF estimates the probability distribution of the prediction and thus an informative uncertainty is associated with the RF prediction (Khaledian and Miller, 2020). Recently Fendrich et al. (2024) have coupled a semi-parametric GAMLSS model and QRF for a European mapping of arsenic taking into account censored data Some pending questions are the over- and underestimations induced by heterogeneous populations.	Van Eynde et al., 2023; Xiao et al., 2023; Hengl et al., 2021; Wadoux et al., 2020
C-A and S-A fractal methods. Multifractal Inverse Distance Weighting interpolation (MIDW)	In the European projects GEMAS and FOREGS, Italy applied the C-A and S-A fractal methods for establishing background noise. MIDW was used for mapping. In the heterogeneous urban context, such a fractal log-linear relationship can only be established locally. Some pending questions are the calculus self-similarity and its relationship with the method used for the base plan, the geometry of the counting zone and the edges of the calculation domains where the amount of information is decreasing.	Albanese et al., 2007; Civitillo et al., 2016; Petrik et al., 2018
Ensemble of machine learning models	This state-of-the art ensemble approach includes five different models: Cubist	Ballabio et al., 2024

(*continued on next page*)

Table 1 (*continued*)

Encountered approaches	Description	Example
	regression trees (Quinlan, 1993), ordinary least squares (OLS) regression (Andrade et al., 2020), xgbTrees (Friedman, 2001), elastic net regression (Friedman et al., 2010), and Gaussian process regression (GPR). The ensemble's combined output is pooled into a single prediction using a Cubist meta-model. Thus, a concentration can be predicted in various ranges of its value by different models to increase the prediction accuracy.	

information dissemination algorithm that was developed by Zeydina and Beauzamy (2013) to handle SIC data. The resulting algorithm is applied to 4 case studies that each shed light on its capabilities. The first case study is a set of core-sample data from Dahlberg (1975), with only 13 data points. With this type of dataset, it is impossible to construct a variogram and therefore to perform any sort of meaningful kriging or to find hyper parameters for an AI model. The second dataset, from a digital challenge (Dubois and Galmarini, 2005), is a set of approximately two hundred data that are used to detect anomalies. The other two datasets are from real surveys of concentrations of various elements in soil, i.e., the large-scale European soil survey (over 2000 data points), with which some 50 arsenic anomalies were identified (GEMAS Project; Tarvainen et al., 2013) and the soil survey of the city of Toulouse (France) performed for the purpose of defining urban geochemical backgrounds (Belbeze et al., 2019).

This article addresses probabilistic information diffusion mapping and innovative algorithmic developments for the interpolation of SIC data and the production of anomaly maps. The underlying theory and basic equations are described and the algorithm is applied to the above-mentioned datasets. Further development of the proposed Incomplete Imprecise Spatial Data Interpolator Algorithm (IISDIA) and anomaly detector is currently ongoing as part of the Horizon Europe Mission Soil project ISLANDR (https://islandr.eu).

2. Definition of requirements for SIC data interpolation and anomaly detection

2.1. Epistemic choices regarding interpolation of data collected at European scale

Maps showing the concentration of various elements in soil at the continental scale of Europe are regularly proposed. The spatial interpolation techniques used vary depending on authors and project objectives. For example: moving median (MM; Tarvainen et al., 2005), kriging (K; Tarvainen et al., 2013), multilevel B-splines (MBS; Panagos et al., 2014), kriging with external drift (KED; Tóth et al., 2013, 2016; Heuvelink et al., 2016), geographically weighted regression (GWR; Xu and Zhang, 2021; Zhang et al., 2011), quantile regression random forest (QRF; Van Eynde et al., 2023; Xiao et al., 2023) and for one of the most recent, a composite of five different models (Ballabio et al., 2024). Table 1 presents the characteristics of these key methods. This list is by no means exhaustive but highlights a few examples from selected high-visibility projects, to illustrate current practice. For additional information, a more comprehensive review of interpolation techniques used by geochemists to establish urban geochemical backgrounds can be found in Belbèze et al. (2023). The general process used for these different mapping techniques is the following: an expert examines the data, selects a geospatial model and calibrates it. To facilitate this exploration of experimental data, some data (e.g. outliers) are removed and more-or-less complex transformations are typically applied to the data, such as, e.g., logarithm transformations, primarily to ensure Gaussian data characteristics, which is a prerequisite of certain interpolation methods. The chosen model is then calibrated (see below), and soil content values are estimated. In more detail, exploratory spatial data analysis involves removing certain outliers based on expert opinion and assessing the relevance of various methodological choices, such as stationarity, support, additivity, and accumulation. This step may require data transformation techniques like log translation, anamorphosis, or Box-Cox transformations, as well as constructing an external drift using models that may include machine-learning algorithms. Additionally, spatial statistics are calculated, including variogram analysis and tuning of covariance hyperparameters. When calibrating a geostatistical covariance model based on an experimental variogram, only a limited number of derivable functions can be used, such as the exponential, spherical or Mattern model. While the model can be calibrated automatically using methods such as least-squares or maximum likelihood, the investigator still controls and sets the shape beforehand. For machine learning algorithms like QRF or Ensemble Machine Learning models, hyper-parameter calibration is essential and involves multiple runs or anneals, such as determining the number of random trees or resampling parameters. As a co-variable or to construct their external drift, interpolation models use correlations between low-density measured values and a densely sampled variable (such as geology or land use, etc.) in order to increase the output map resolution This is behind the CLORPT concept (CLimate, Organisms, Relief, Parent material, and Time; Jenny, 1994) or SCORPAN concept (SCORPAN = Soil or measured attributes of the soil at a point; Climate; Organisms, including land cover and natural vegetation; Relief, topography including terrain attributes and classes; Parent material, including lithology; Age, the time factor; N, space, spatial or geographic position; McBratney et al., 2003). The objective is to co-model the results from sparse measured soil content data surveys with measurements of covariates (geology, land use, photographs, etc.) that are evenly distributed over the area. However, it should be warned that the approach may suffer from uncertainties related to model calibration and, for the methods mentioned above, to the cascade of covariate scales which may generate a "false reality". This results from epistemic uncertainties (i.e., due to incomplete model knowledge). As illustrated by, e.g., Ferson and Ginzburg (1996) or Loquin and Dubois (2010), such uncertainty should be distinguished from stochastic uncertainty, which stems from the random variability of natural phenomena underlying the measured quantity (e.g., heterogeneity in space and/or time).

To summarize, the various interpolation model approaches often rely on several strong hypotheses and epistemic choices, e.g., that the phenomenon to be interpolated is continuous between two measurements; that data transformation (log, normal score, anamorphosis) does not lead to misinterpretation; that values considered as anomalous can be removed from the calculation, etc. Experts generally select two models, e.g., one for the spatial response of the variable and one for its links to explanatory parameters (covariates), often with the assumption that the greater the number of covariates, the better the result. When these expert models have sufficient data and are well calibrated, they are effective, validated, and published, but if the data are SIC (Sparse, Imprecise, Clustered), the epistemic uncertainty of such mappings increases significantly. This is also the case for advanced geochemical filtering such as the filtering co-kriging developed by Sauvaget et al. (2022) or the MAF/ILR filtering kriging used by Melleton et al. (2021). They are only possible when the quantity of data is large and the spatial structuring is correctly modelled by an expert. Otherwise, a new algorithm tailored for SIC data is necessary, which enables to cover all possible scenarios and, above all, the diversity of measurement sources that other methods cannot always take into account with so many epistemic choices.

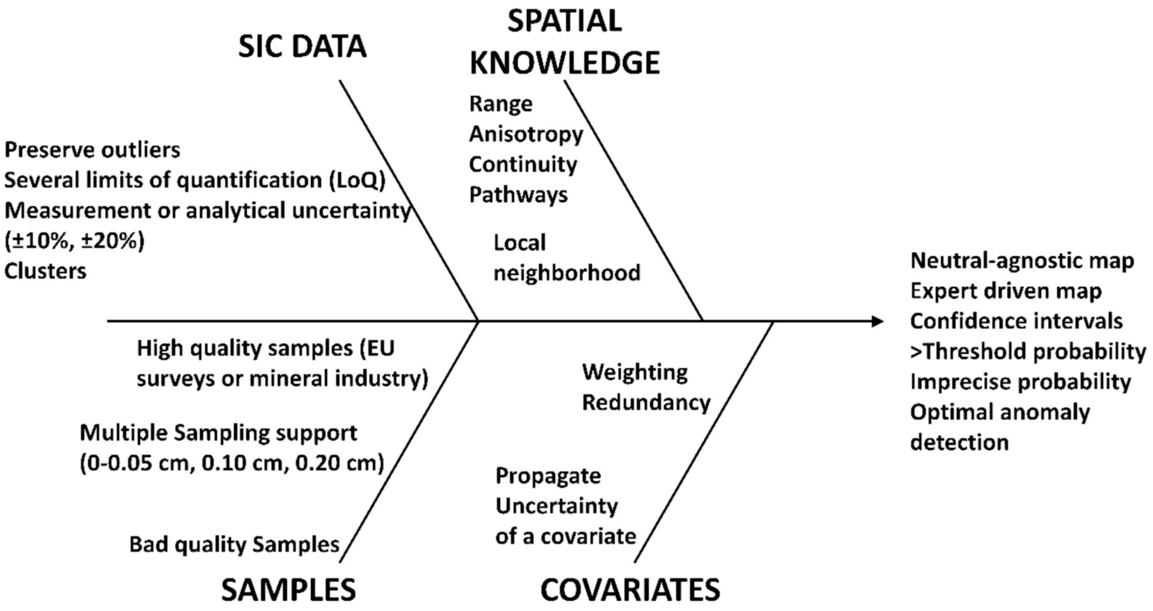
SIC DATA
Preserve outliers
Several limits of quantification (LoQ)
Measurement or analytical uncertainty (±10%, ±20%)
Clusters
SPATIAL KNOWLEDGE
Range
Anisotropy
Continuity
Pathways
Local neighborhood
High quality samples (EU surveys or mineral industry)
Multiple Sampling support (0-0.05 cm, 0.10 cm, 0.20 cm)
Bad quality Samples
SAMPLES
Weighting
Redundancy
Propagate Uncertainty of a covariate
COVARIATES
Neutral-agnostic map
Expert driven map
Confidence intervals
>Threshold probability
Imprecise probability
Optimal anomaly detection

Fig. 1. Fishbone diagram for the IISDIA algorithm. The fishbone diagram is a tool for identifying the main causes of a problem by categorizing ideas, which guides algorithmic development. Here, this diagram indicates the main errors and uncertainties in our data and knowledge that can affect the IISDIA interpolator.

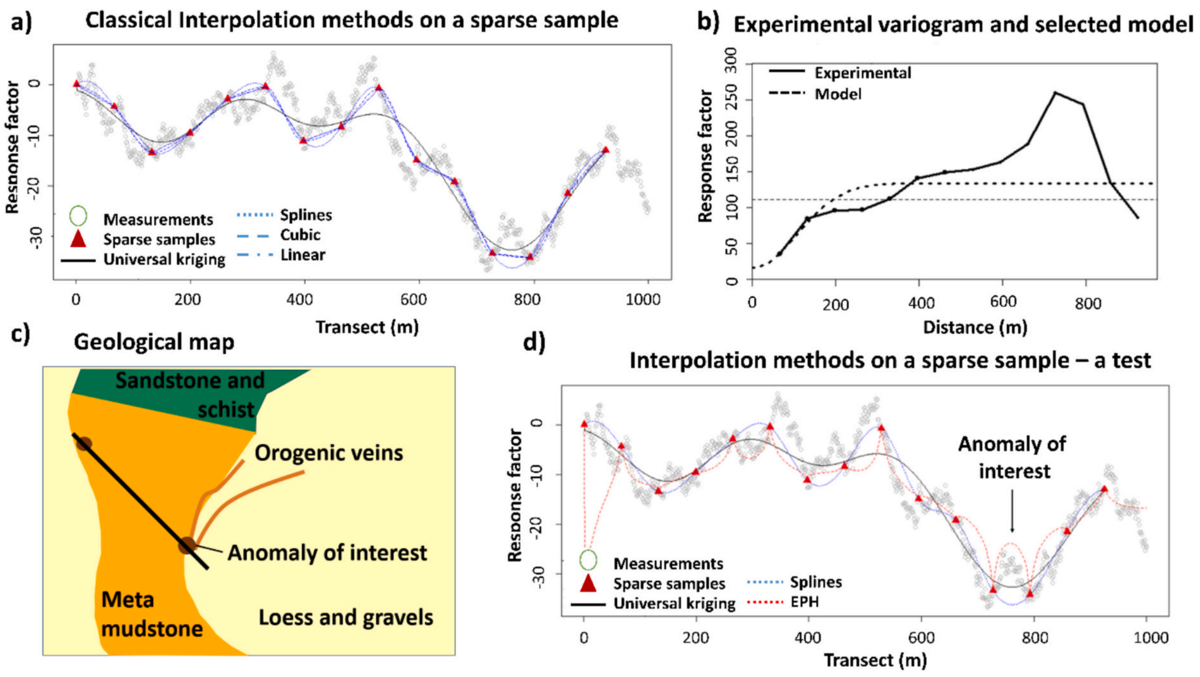

Fig. 2. An interpolation experiment on a sparse set extracted from a geochemical transect intercepting several geological formations and veins that may host a mineral deposit. a) Universal kriging, splines, cubic and linear interpolator applied to 15 data points, b) Experimental and model variogram used for the Universal kriging interpolation, c) Geological map of the area and transect locations, and d) Universal Kriging, splines and Experimental Probabilistic Hypersurface applied to 15 data points. The response factor is the log ratio of the content to its background.

In summary, the desired properties for such a spatial interpolator are i) no transformation of data; ii) no outlier removal; iii) no variogram or hyper parameter expert model (to avoid smoothing anomalies); iv) possible application to moving windows of <10 data in a non-stationarity context. Therefore, the interpolator should be sufficiently robust and applicable for anomaly detection or forensic soil provenance (Aberle et al., 2023). Nevertheless, with the IISDIA interpolator presented herein, the expert still has the possibility to add knowledge regarding physical phenomena and various information such as directions of structures, ranges of variables, sample quality, dissonance, etc. (Fig. 1).

2.2. *A multimethod experiment*

To illustrate the proposed approach, we first consider the example of geochemical data collected on a regular transect intercepting several geological formations and a veins zone that could host a mineral deposit. First a sample of 15 data points is selected to obtain the SIC dataset to be interpolated. Next several interpolation techniques are applied; a geostatistical technique (Chilès and Delfiner, 2013) known as universal

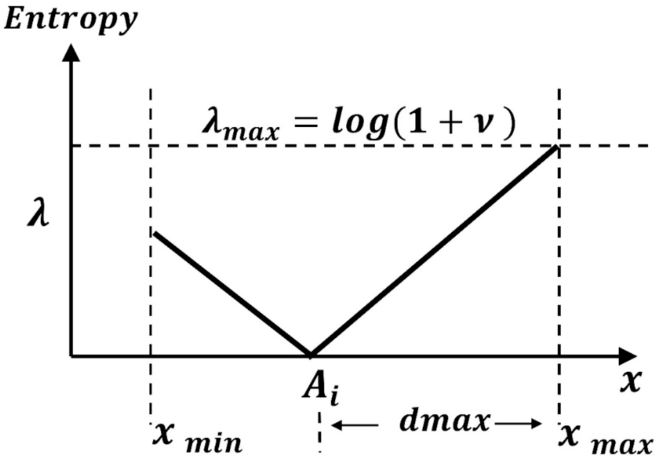

Fig. 3. Entropy variation with the distance to point X to be evaluated. Observation points are denoted by A_i, $i = 1, \ldots, n$ where content C_i has been observed; λ is the entropy of the law of C_i versus distance d. $[x_{min}, x_{max}]$ is the range of coordinates.

kriging (which incorporates an automatic trend), a spline method (Buhmann and Jäger, 2021), a cubic interpolator (Quarteroni et al., 2010) and a linear interpolation method (Fig. 22a). Compared with the reference dataset, a strong smoothing effect is observed due to the scarcity of the data and the algorithms used. The variogram generated for universal kriging shows that the data are not stationary (the mean is not constant since the data are taken from different geological formations). Thus, the calibration of the model is awkward without additional knowledge of the underlying geology (epistemic uncertainty). Kriging and splines seek to establish a spatial structure in the data, in the form of a covariance as a function of distance between observations (Fig. 2b). This limits the applicability of these models when the data are SIC (Belbèze et al., 2023; Malone et al., 2017; Helsel et al., 2012; Reimann et al., 2008). Using now the EPH (Experimental Probabilistic Hypersurface; Zeydina and Beauzamy, 2013) algorithm detailed in the next section, we obtain the Fig. 2d where the anomalies are seen to be sharper and where, unlike with the other methods, an anomaly corresponding to the orogenic veins (Fig. 2c) is detected. The geological interpretation of this transect is that two areas are sampled with two orogenic veins. (Fig. 2c).

3. Interpolator development methodology

3.1. EPH base algorithm

Information diffusion mapping is becoming increasingly popular in situations where there are not enough data points to obtain a statistically relevant variogram for geostatistical methods (Berton, 2018; Huang et al., 2019). Driven by image processing research (Ho et al., 2020), diffusion models produce impressive artificial intelligence-generated image quality (Dhariwal and Nichol, 2021). Two mathematical approaches led to this theoretical development. The first, developed in China, is based on potential and the analogy between information transfer and natural phenomena such as diffusion or vibration and is simply called Information Diffusion. This technique has been applied to various fields, including flood hazard mapping (Huang et al., 1998; Zhou et al., 2000; Yi et al., 2007 etc.), seismicity (Bai et al., 2015), and more recently, precipitation (Huang et al., 2019). The second approach, known as Experimental Probabilistic Hypersurface, was developed by Beauzamy (2004) and relies on the propagation of information entropy (Zeydina and Beauzamy, 2013). This entirely probabilistic method has been utilized for nuclear safety calculations (Godan et al., 2015) and for diverse models such as neutron sensor networks for nuclear reactor operation and territory monitoring for radioactive plumes (Khalipova et al., 2018). It is designed to minimize assumptions, particularly by avoiding a fixed model of data covariance, and in its simplest version, it excludes spatial covariance, showing immunity to outliers, which it magnifies instead. According to Beauzamy (2004), information propagation is based on a general principle of maximum entropy (or minimum information), which is itself an increasing function of distance to the measurement point of the law of C_i versus distance (Fig. 3).

The demonstration is based on entropy propagation of uniform laws and makes use of three proven lemmas: minimal information; maximal entropy; distribution with maximal entropy and fixed variance (Zeydina and Beauzamy, 2013). From a less abstract point of view, the EPH interpolator resembles a spatial kernel method. Several spatializations of the kernel method (Parzen, 1962) have been proposed such that the neighbor effect decreases with distance, as it does in reality (Levine, 2010; Gibin et al., 2007; Xie and Yan, 2008). But EPH is somewhat more complex because the weighting of the kernel and the diffusion coefficients are nonlinear functions of distance to coordinates and parameters. The EPH model produces its estimates in the form of a discrete probability distribution for a given interval and requires two input parameters: min-max limits of each dimension (parameters); min-max

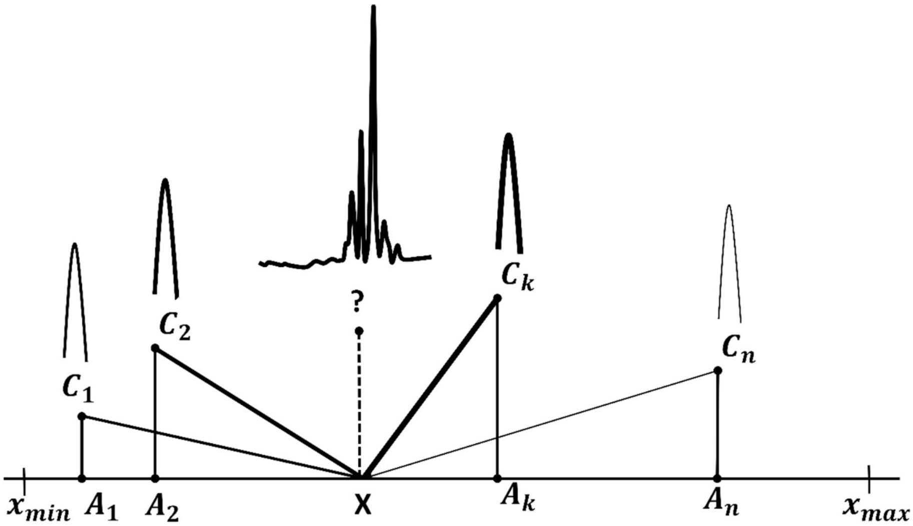

Fig. 4. Information diffusion process attenuated by neighbor distance to X. Point X is to be evaluated. Observation points are denoted by A_i, i = 1, …, n where content $C_i$has been observed.$[x_{min}, x_{max}]$ is the range of coordinates.

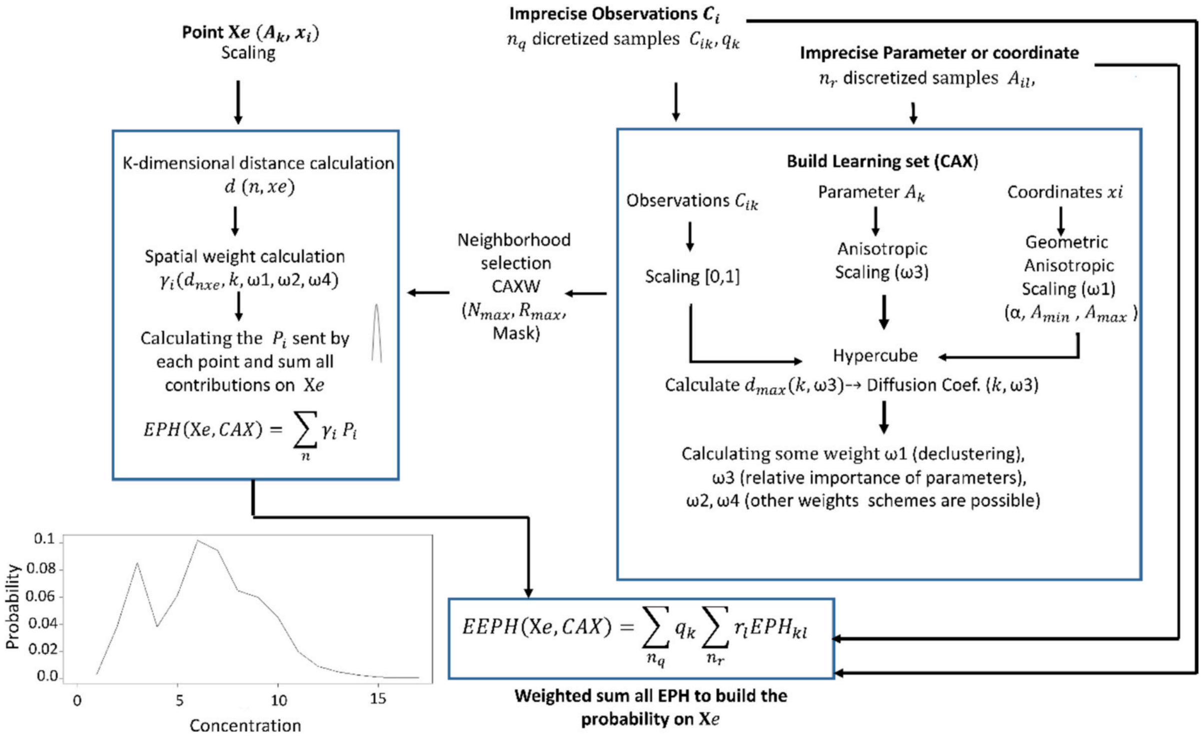
Point Xe (A_k, x_i)
Scaling
Imprecise Observations C_i
n_q dicretized samples C_{ik}, q_k
Imprecise Parameter or coordinate
n_r discretized samples A_{il},
K-dimensional distance calculation
$d\ (n, xe)$
Spatial weight calculation
$\gamma_i(d_{nxe}, k, \omega1, \omega2, \omega4)$
Calculating the P_i sent by each point and sum all contributions on Xe
$EPH(Xe, CAX) = \sum_n \gamma_i P_i$
Neighborhood selection
CAXW
$(N_{max}, R_{max},$
Mask)
Build Learning set (CAX)
Observations C_{ik}
Parameter A_k
Coordinates xi
Scaling [0,1]
Anisotropic Scaling (ω3)
Geometric Anisotropic Scaling (ω1)
$(\alpha, A_{min}, A_{max})$
Hypercube
Calculate $d_{max}(k, \omega3)$→ Diffusion Coef. $(k, \omega3)$
Calculating some weight ω1 (declustering), ω3 (relative importance of parameters), ω2, ω4 (other weights schemes are possible)
$EEPH(Xe, CAX) = \sum_{n_q} q_k \sum_{n_r} r_l EPH_{kl}$
Weighted sum all EPH to build the probability on Xe
Probability
0.1
0.08
0.06
0.04
0.02
0.0
0
5
10
15
Concentration

Fig. 5. Building the Enhanced Experimental Probabilistic Hypersurface (EEPH).

limits of the modelled phenomenon and discretization of this interval (τ steps, μ intervals). The min-max limits may be derived from expert knowledge or physical limits or may be defined by a user after studying the data. In practice, the discretization step τ corresponds to the precision required (ppm, ten ppm, etc.). A small change in the min-max limits has in any case little impact on the final result.

Considering n observation points, noted by A_i, $i = 1, \ldots,$ n where C_i has been observed, X is the point at which a c estimate is to be obtained, and K parameters are available for each measurement point and are available for X. We then have a manifold hyperspace $A_i(a_1, \ldots, a_K)$ and $X(x_1, \ldots, x_K)$. Each n-point A_i contributes to the final result of the density of X according to Eq. 1.

$$P_{A_i,j}(X) = \frac{\tau}{\sigma\sqrt{2\pi}} exp\left[-\frac{(c_j - C_i)^2}{2\sigma^2}\right] \tag{1}$$

Density of this kind takes the form of a Dirac function at the location of a measurement point (the value is known precisely) and becomes increasingly less concentrated with distance (Fig. 4). N-point contributions A_i to point X are recombined to form a single P_{xj} where the various contributions are weighted according to the distance between the target point and each measurement following Eq. 2:

$$P_{xj}(X) = \gamma_1 P_{1,j}(X) + \ldots + \gamma_n P_{n,j}(X) = \sum_{i=1}^{n} \gamma_i P_{A_i,j} \tag{2}$$

where $d_i = d(A_i, X)$ is the distance between the point to be reconstructed and the i-th measurement (Eq. 3 and 4):

$$d_i = d(A_i, X) = \sqrt{\sum_{k=1}^{K} (a_k - x_k)^2} \tag{3}$$

$$\gamma_i = \frac{d_i^{-k}}{\sum_{i=1}^{n} d_i^{-k}} \text{ in dimension k} \tag{4}$$

In fact, each $P_{xj}(X)$ is a conditional probability given the other measurements made according to Eq. 5 and illustrated in Fig. 4:

$$P_{xj}(X) = P_{x|A_1,..A_n,j}(X) = \sum_{i=1}^{n} \gamma_i P_{A_i,j} \tag{5}$$

The diffusion coefficients (σ) can be calibrated by maximum likelihood (Bartkute and Sakalauskas, 2008). The Gaussian form of the information comes from the fact that the distribution with maximum variance for a fixed entropy is Gaussian (Zeydina and Beauzamy, 2013). From this probability distribution in X, we can extract a confidence interval. We can also easily modify the distribution of coefficient calculation γ_i or $P_{A_i,j}(X)$ to include uncertainties. It is also possible to calculate several $P_{xj}(X)$ by m Monte Carlo iterations of a parameter and combine these EPHs to obtain the resultant according to Eq. 6 with q_k being probability of scenario k:

$$P_{xj}(X) = \sum_{k=1}^{m} q_k P_{xj}^{\ k}(X) \tag{6}$$

However, the examination of the algorithm's response to the entire range of parameter variations and their interactions, the latter method rapidly becomes too costly. Therefore it is preferable to use a global Monte-Carlo approach based on a probability of exceeding a threshold with several possible situations.

3.2. Enhanced EPH base algorithm

The proposed enhancement of the EPH method involves the introduction of additional knowledge in the calculation of the original "neutral" EPH (Zeydina and Beauzamy, 2013) such as, e.g., including the scope of phenomena, the weight of explanatory variables, uncertainty on variables, declustering, anisotropy, etc. These modifications required adapting the EPH equations or encapsulating the EPH in a double global Monte Carlo scheme (Fig. 5). The scheme was coded in R (R Core Team, 2022) and designated as EEPH for Enhanced Experimental Probabilistic Hypersurface. All changes made to the EPH that have led to the EEPH are described below since to-date, no probabilistic algorithm of this type has been proposed in the literature.

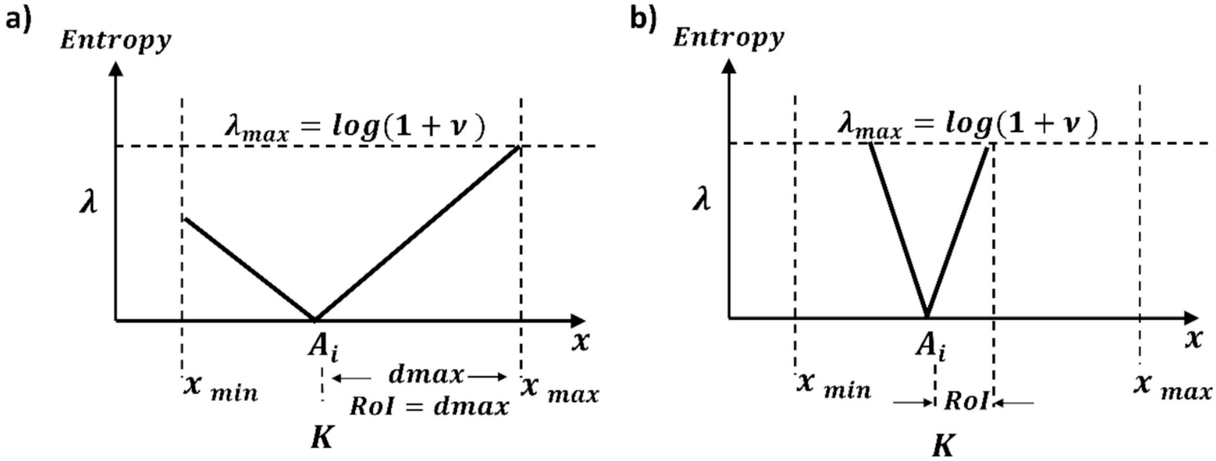

Fig. 6. Modified entropy growth with distance. RoI is the chosen range of influence. When unknown RoI is set to d_{max}.

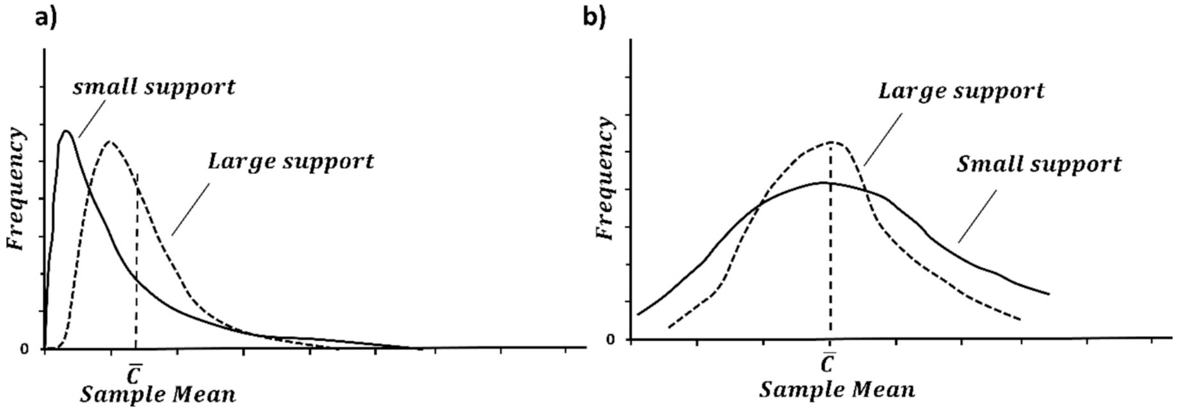

Fig. 7. Effect of support on content histograms with a) based on BRGM ore prospect data and b) based on Sinclair and Blackwell (2002).

As the new interpolator yields a full probability density for each point, even when data are SIC, an entire range of probabilities is explored, while other interpolators rely on assumptions that are often hard to verify in practice such as inverse distance or variogram. Moreover, the new interpolator can account for measurement or parameter uncertainty. If it has been established, for example, by expert opinion or using an experimental directional variogram, that the measurements have a geometric anisotropy, this knowledge can be included in the calculation by setting angles and ratios. There is another type of anisotropy called zonal anisotropy, which is more difficult to integrate into SIC data if it is detected, as it requires knowledge of the populations involved. In that case, we can either distort the data space or partition the populations and handle them separately. To do so, we can make a precise selection of neighbouring data and perform a local EEPH. This is especially true in the case of groundwater, where flowlines change direction at soil permeability interfaces.

While EPH is unaffected by outliers, it is algorithmically very sensitive to data clusters (Berton, 2018), which can bias its spatial probability calculation. As a countermeasure, a cluster bias correction weight W_{clus} has been developed based on the Hclust 3.6.2 version of R Core Team (2022). Different declustering methods exist, and we have adapted a two-point declustering method originally developed by Richmond (2002). An attractive aspect of this technique is that it does not depend on the selected grid configuration, but rather on the position of the detected clusters. The experimental variogram is the tool of choice for studying the range of a physical phenomenon (Chilès and Delfiner, 2013). It is a powerful tool that provides information on the spatial behaviour of soil content. In geostatistics, a variogram is modelled assuming a variety of assumptions which, in the case of SIC data, are not applicable. Nevertheless, the variogram provides information on different zones, their average extension, their anisotropy, and is useful to calculate systematically on all the data or a group of data points thought to have a particular behaviour. If it is established, for example, by expert opinion or by means of an experimental variogram that the neighbourhood measurements no longer have any influence beyond a certain Radius of Influence (RoI), it is possible to modify this dmax quantity in the EEPH, which adapts the slope of the entropy accordingly (Fig. 6).

As with kriging, the map produced then considers a range of phenomena but loses the neutrality on the notion of content continuity and the anomalies detection that are the algorithm's strength. Another option for EEPH would be to introduce various Euclidean distances as parameters, based on the work of Behrens et al. (2018). This involves generating concentration variables as a function of the distance between points. This process is analogous to the construction of an experimental variogram. These concentrations are then injected into the calculation as covariates. The spatial proximity of low, medium, and high values to our probability at a given point then becomes part of the calculation. In the EEPH, distances are already accounted for once via the diffusion coefficient, but considering the differences between concentrations goes even further, revealing a statistical data structure comparable to those seen in variograms. This algorithm has been coded for the EEPH but, like the variogram, it is highly dependent on the number of points (a minimum of 50–100 pts. is required). Therefore, when used on a dataset with a significant number of measurement points, this algorithm produces results comparable to kriging for a quantile random forest algorithm

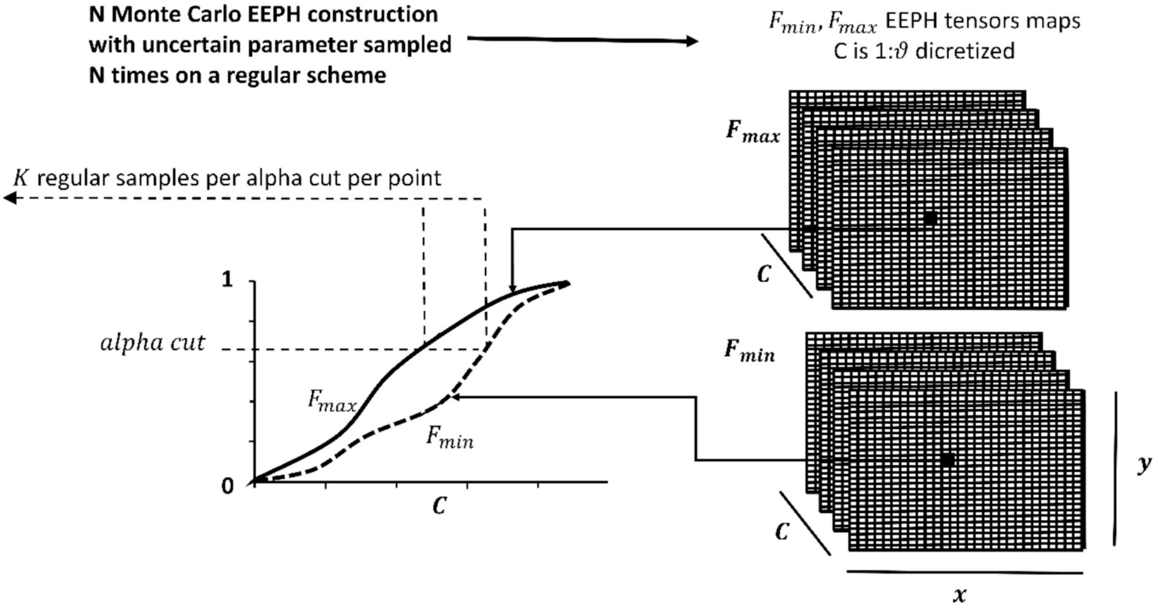
N Monte Carlo EEPH construction with uncertain parameter sampled N times on a regular scheme
F_{min}, F_{max} EEPH tensors maps C is 1:ϑ dicretized
K regular samples per alpha cut per point
F_{max}
C
1
alpha cut
F_{max}
F_{min}
0
C
F_{min}
C
y
x

Fig. 8. Keeping only the min and max for each point on each N Monte Carlo run to feed the risk calculus. F is a possible spatial cumulative distribution function of soil content C. F_{min} (plausibility) and F_{max} (credibility) are probability bounds of Fs (Shafer, 1976).

(QRF; Hengl et al., 2018).

While SIC data does not typically carry continuity or even trend information; a strong epistemic hypothesis is made when an expert-selected continuous spatial covariance is imposed on the data. To establish such “continuity” in EPH, we need to introduce at least one parameter to convey it. If the main causal factor of soil content is the geological nature of the subsoil, using this information as a parameter will restore a continuity of phenomena that cannot be deduced from the data alone. Taking the example of the GEMAS project (Tarvainen et al., 2013), a surface soil sample is a 2–2.5 kg sample collected from a 10 m grid by adding and mixing (compositing) 5 samples (subsites) on the grid. This protocol assigns the analysis result to the mesh size, which means that the average of the samples is attributed to it. It should be noted that the notion of scale applies here, as it does to all mapping (Lindeberg, 1997). This is because soil content is subject to physical laws, among which spatial additivity. The average over an area (A) or volume (V) must be equal to that of its sub-sections v_i. This volume V or v_i is referred to as the “support” for the content information. Apart from the mean, statistics such as variance will vary depending on the support. Changing supports (large samples/small samples, boreholes/grids, etc.) in geostatistics is of the utmost importance, and is directly related to the physical laws of sampling described above. Difficulties arise particularly when merging two differently-sampling campaigns. Fig. 7 shows the effect of sampling the content support.

We observe in Fig. 7 that although the mean remains fairly stable between supports, the variance decreases sharply. In geostatistics there are methods to account for changes of support, particularly when soil measurements are performed at a precise location (such as soil cores or trenches), and the contents must be estimated on panels or supports that are much larger than those observed. The geostatistical approach involves modifying the spatial function models (covariance, variogram) that describe the spatial dependence between observations to account for the change in support (Chilès and Delfiner, 2013). These techniques require excellent variogram modelling and therefore a significant amount of data. Once the models have been calibrated, a theoretical point location model is calculated; the calculations are performed before being re-transformed into output support (Kasmaeeyazdi et al., 2020). For the EEPH, three main cases of support change have been considered, depending on the type of sampling encountered.

Firstly, if the sampling is of good quality, in line with mining standards between sampling campaigns, supports are known and documented. The supports are based on strict equi-probabilistic sampling rules that guarantee additivity. In this case, content is corrected for its support to give accumulations (Chilès and Delfiner, 2013), and maps are made for these before being switched back to map support. In addition, this type of mining-inspired sampling always has extensive error management, with duplicates, etc. This makes it possible to efficiently fill in the error function as a % of the mean introduced in the EEPH equations.

Secondly, the sampling may be of average quality. For example, there is no correspondence between sample lengths and the lithology measured, and it is not possible to work in accumulation. Nevertheless, sampling has been carried out according to a fixed protocol, as is customary for studies of polluted sites and soils by environmental consultants. The proposed approach is as follows: we place ourselves in a mesh comprising samples from two different supports (small support S1, large support S2) supposed to represent the same content (in the case of sampling, we consider them to represent the same profiles or collocated boreholes). If they had been identical, their average would have given the grade of the mesh. As this is rarely the case, we assign to the smallest support a virtual value that it would have in the large support S2 associated with significant local error (Lajaunie, 1996). For our EEPH, the virtual content of the small support becoming S2 would be between the value obtained for the large support and that of the small support. This interval will be taken uniformly and produce the sample corrected map.

Thirdly, and last considered case, the sampling is of poor quality. Either the mass sampled is insufficient according to the Visman equation (Visman, 1969) or the sampler did not sample the soil correctly (termed “Grab sampling” in Minkkinen and Esbensen, 2009). A composite of samples is necessary to restore correct averaging, which means degrading the resolution of the datasets to make them compatible. An example of such a drastic countermeasure can be taken from Ingamells and Pitard (1986). In this case, a dozen boreholes were drilled in a cobalt mine in the 1980s. The results of the sample appear to indicate that there is little cobalt in the panel boreholes, with the exception of two samples. The mine might have been abandoned, but the excavation was carried out based on the geological expert's opinion. It yielded minable ore with 0.2 % cobalt. The cobalt was concentrated in pockets and grains that were only randomly intersected by drilling. The pattern of the histogram can help predict such a phenomenon: it's not a log-normal but an

Table 2
Triangular fuzzification (TFN) of anomaly detection techniques index on the GEMAS dataset.

5-item Likert scale	Reimann statistics	Zero probability Bands Index	Index C-A fractal	Singularity index	Moran index	Anomaly cluster Index	Nemerow index	TFN
Virtually certain	TIF	3	4	0.5	10	3	6	(81.4, 100, 100)
Very likely	Q98	2	3	1	3	2	3	(61.6, 81.4, 100)
Likely	Q95	1	2	1.5	2	1	1	(22.6, 41.5, 61.6)
About as likely as not	–	0	1	2	1	0	0	(0, 22.6, 41.5)
Unlikely	–	–	0	2.5	0	–	–	(0, 0, 22.6)

Table 3
Anomaly detection TFN meta-ranking weights selected by author.

Anomaly index	Type of outlier	Expert opinion	Expert 5-item Likert scale	TFN w_i
Reimann statistics	Range outlier	Very efficient if data is not multimodal.	Very good	(0.3, 0.38, 0.45)
Zero probability bands index	Range outlier Relationship outlier	Detects outliers on small sample sets like ITA3 but less effective for big surveys like the GEMAS	Poor	(0.005,0.01,0.03)
C-A fractal index	Range outlier Spatial outlier	A statistical fractal method. Choosing Limits on a C-A curve is subjective.	Average	(0.05, 0.04, 0.1)
Singularity index	Spatial outlier	Window-based fractal statistics. Efficient on GEMAS dataset	Good	(0.17, 0.24, 0.3)
Moran index	Spatial outlier	Window-based variogram based statistics. Efficient on big anomalies but didn't detect light-signal anomalies	Average	(0.02, 0.03, 0.1)
Anomaly cluster Index	Spatial outlier Relationship outlier	Window-based cluster statistics.	Good	(0.17, 0.3, 0.4)
Nemerow index	Spatial outlier Range outlier	Not possible on our GEMAS dataset but very efficient on ITA3	Very good	(0.3, 0.38, 0.45)

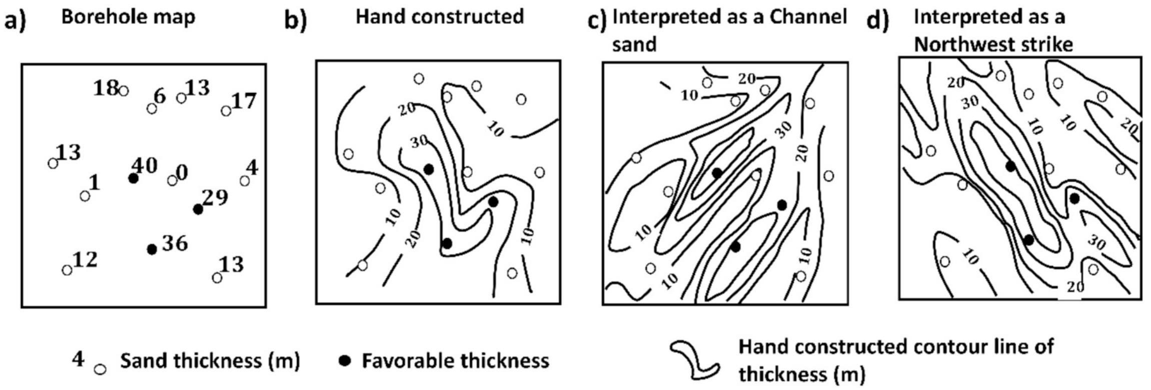
a) Borehole map
b) Hand constructed
c) Interpreted as a Channel sand
d) Interpreted as a Northwest strike
18
6
13
17
13
40
0
4
1
29
36
12
13
10
20
30
4 ○ Sand thickness (m)
● Favorable thickness
Hand constructed contour line of thickness (m)

Fig. 9. Four different SIC dataset interpretations adapted from Dahlberg (1975). a) 13 sand thickness data points in m, b) manual geologist triangulation, c) geologist interpreting profiles as Channel sand deposits and d) geologist map interpreting profiles as regional northwest strike and southwest paleoslope fluvial deposits.

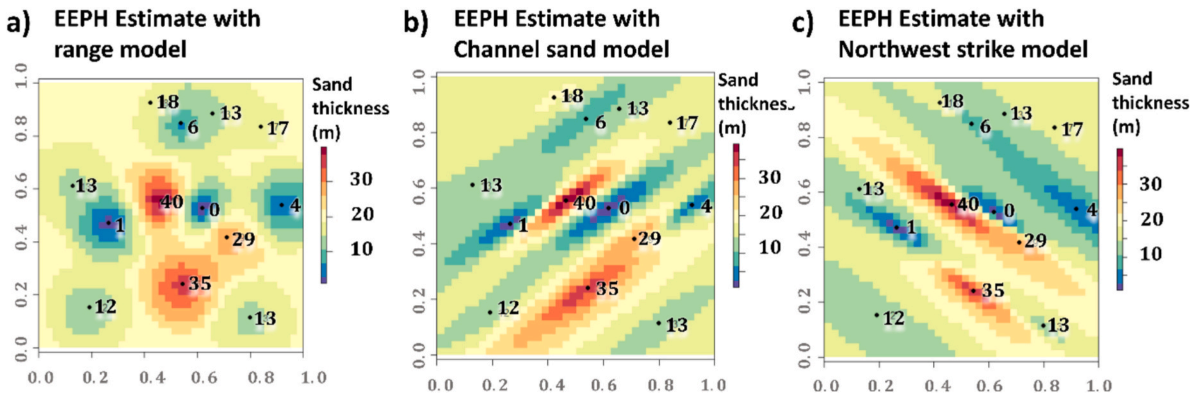
a) EEPH Estimate with range model
b) EEPH Estimate with Channel sand model
c) EEPH Estimate with Northwest strike model
Sand thickness (m)
30
20
10
18
6
13
17
13
40
0
4
1
29
35
12
13
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 10. EEPH-Expectation generated maps with 13 data points from Dahlberg (1975). a)EEPH Expected value with altered range to 0.1 distance units b) EEPH expected value map with 2:1 N60E anisotropy assuming channel sand model c) EEPH expected value map with 2:1 S60W anisotropy assuming a regional northwest strike with a southwest paleoslope.

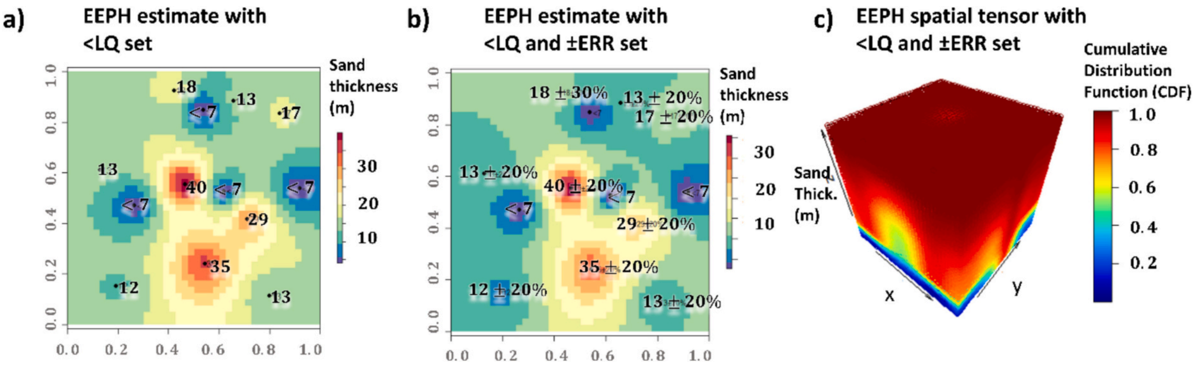

Fig. 11. 13 data points from Dahlberg (1975) and corresponding EEPH-generated map. a) EEPH expected value with 4/13 points values censored to be below 7 m (<LoQ) b) EEPH expected value with 4/13 < LoQ and 8/13 point values set with analytical uncertainty (ERR) of 20 % and 1/13 of 30 %. c) full spatial cdf tensor generated by EEPH with 4/13 < LoQ and 8/13 with analytical uncertainty of 20 % and 1/13 of 30 %.

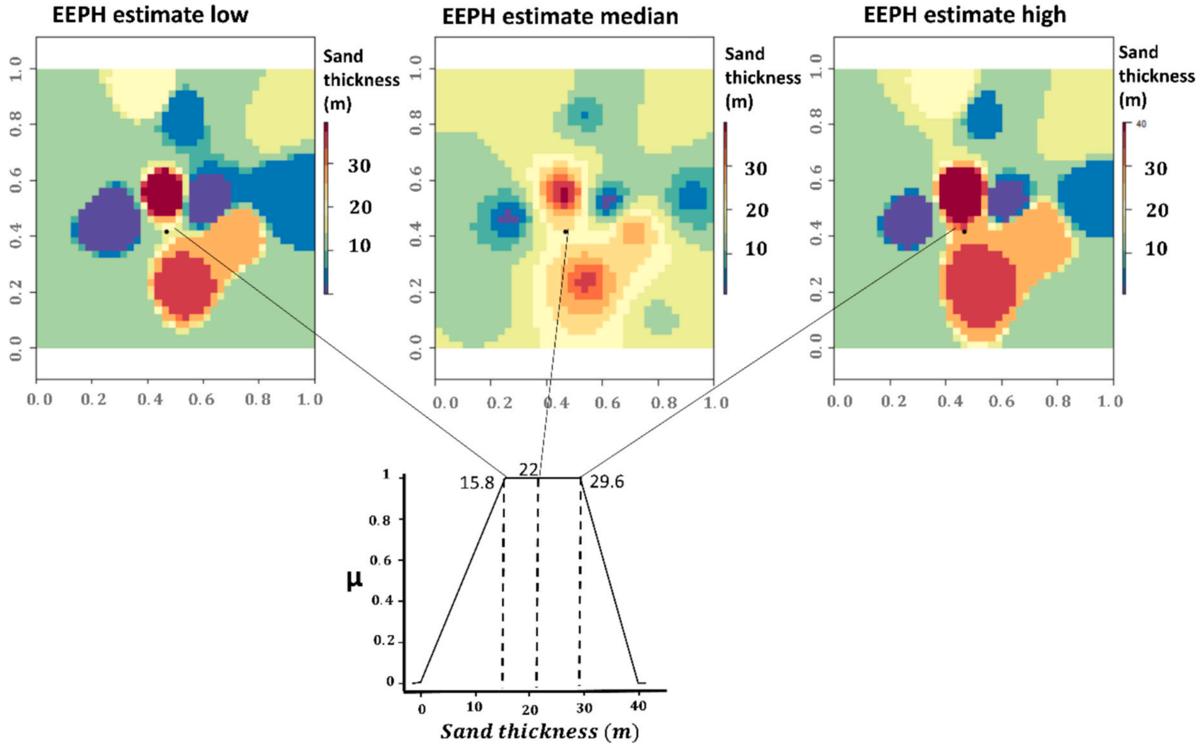

Fig. 12. 13 data points from Dahlberg (1975) three slices of the full spatial cdf tensor of Fig. 11 and the corresponding Trapezoidal fuzzy number generated on an uncertain sample point.

approximate double Poisson distribution that shows some cyclicity.

3.3. Addressing uncertainties in the interpolation

Uncertainties appear at all stages of a geochemical survey, i.e., sampling stage (e.g., core loss, imprecision regarding geolocalization of samples, etc.), analysis stage (analytical uncertainty), interpretation stage (e.g., uncertainty regarding covariates), etc. To account for such uncertainties in the spatial EEPH scheme, the approach developed in HYRISK (Baudrit et al., 2006; https://github.com/BRGM/HYRISKdev) was used. HYRISK is an R package for addressing uncertainty in risk assessments that accommodates uncertainties of both epistemic and stochastic origin (Dubois and Guyonnet, 2011), using, e.g., probability distributions, fuzzy numbers, simple intervals or probability distributions with imprecise moments, that the modeler may select depending on the nature of available information. The uncertainty propagation procedure combines Monte Carlo random sampling with fuzzy interval analysis (see also Baudrit et al., 2007). Sensitivity analysis is included based on the pinching method of Ferson and Troy Tucker (2006). As illustrated in the Dahlberg (1975) application case shown below, with this version of the EEPH algorithm, for each measurement point or parameter, the user can specify a type of uncertainty (possibility, probability, imprecise probability, …). The distribution is then sampled and the q_k weight is calculated. The EEPH calculation is then run in full tensor mode of probability, which will generate a possibilistic content input at each point.

When the uncertainty of a parameter is known in discrete, interval, fuzzy number, or distribution interval form, we can consider various values for it and, as we have seen, perform as many EEPHs as necessary, which will be recombined to yield the final result (Fig. 8). This system is particularly suitable for uncertainties that cannot be incorporated directly into the EEPH calculation. This would be the case for X and Y positioning errors or other parameters. The presence of what are known as censored values corresponds, e.g., to values below the limit of

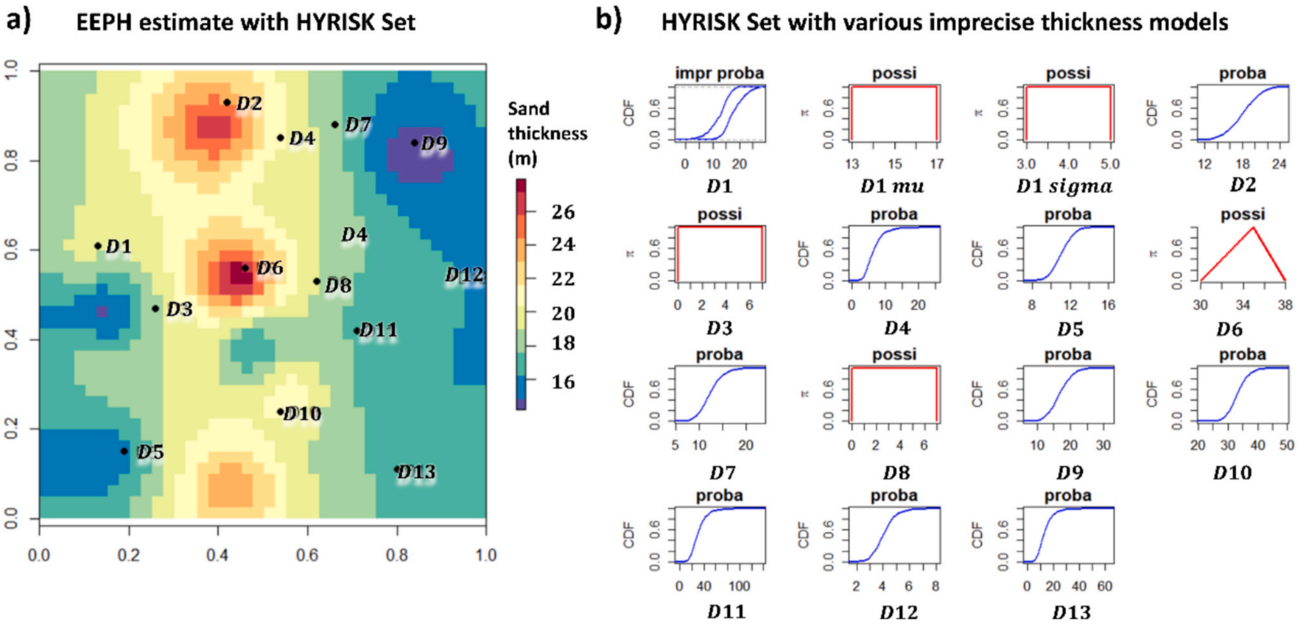

Fig. 13. 13 data points from Dahlberg (1975) with uncertainty prior (content corrected by core loss measurement and sample representativeness) in each point set in Hyrisk formalism and the corresponding EEPH map generated. a) expected value map of the SIC set and b) SIC set in Hyrisk format. "proba" is a probability distribution prior; "impr proba" is an imprecise probability distribution prior; "possi" is a possibility distribution prior.

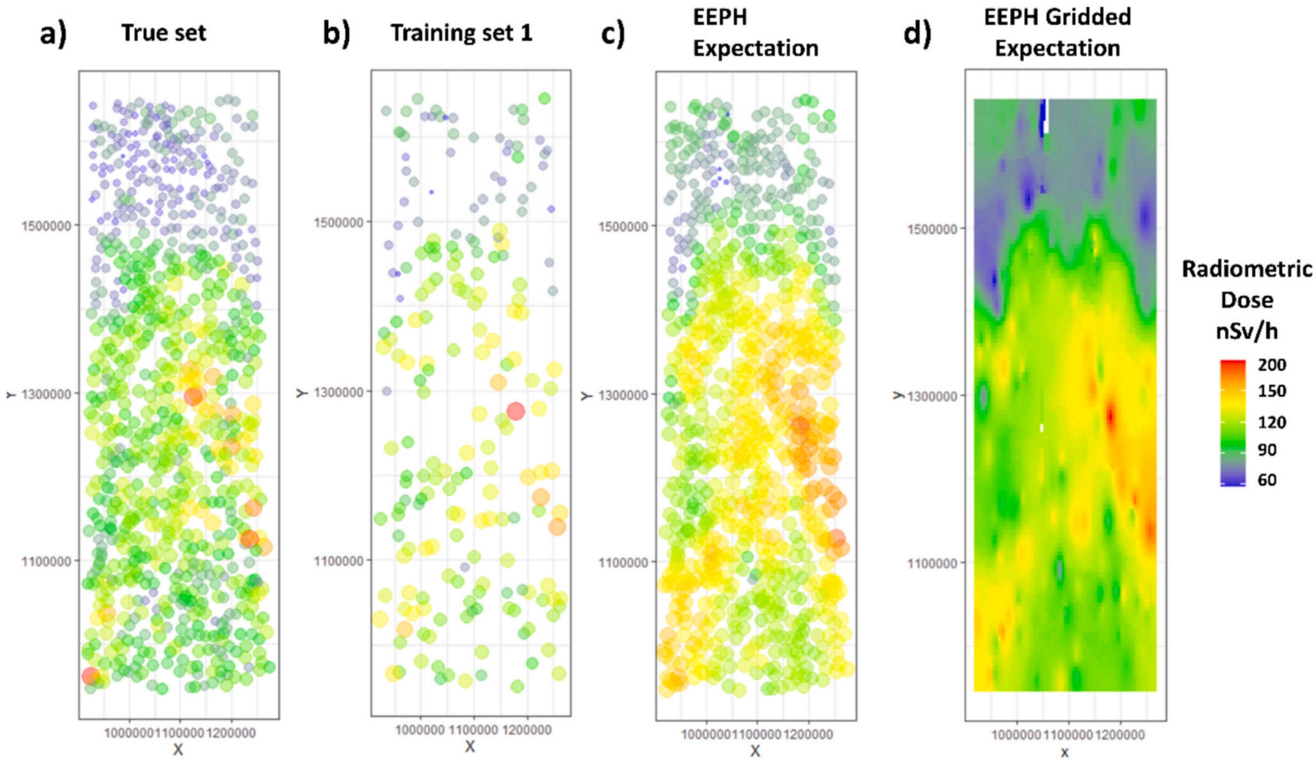

Fig. 14. EEPH tests on Dataset 1 from the SIC2004 exercises obtained in 0.15 s (dataset from Dubois and Galmarini, 2005). a) true set 1, b) training set 1, c) EEPH expectation on validation set and d) EEPH expectation on gridded domain.

quantification (<LoQ). Their presence changes the shape of content distribution functions and can bias our estimates of probability density by EPH. To remedy this while adding a minimum of assumptions to the calculation, we propose a discretization approach in the EEPH. We replace our unquantified values with discretized (and un-simulated) m values of index k between 0 and LoQ, calculate the EPH of each, then agglomerate these EPHs weighted by the probability of occurrence q_k of the discretized content. By default, this is a draw on a uniform distribution (interval), to avoid giving any particular shape to our uncertainty for the value below LoQ. However, the algorithm remains compatible with a possibilistic approach (Baudrit et al., 2006, 2007), and the LoQ distribution could just as well be a possibility instead of a uniform probability. With SIC data, there is no reason to prefer one distribution over another.

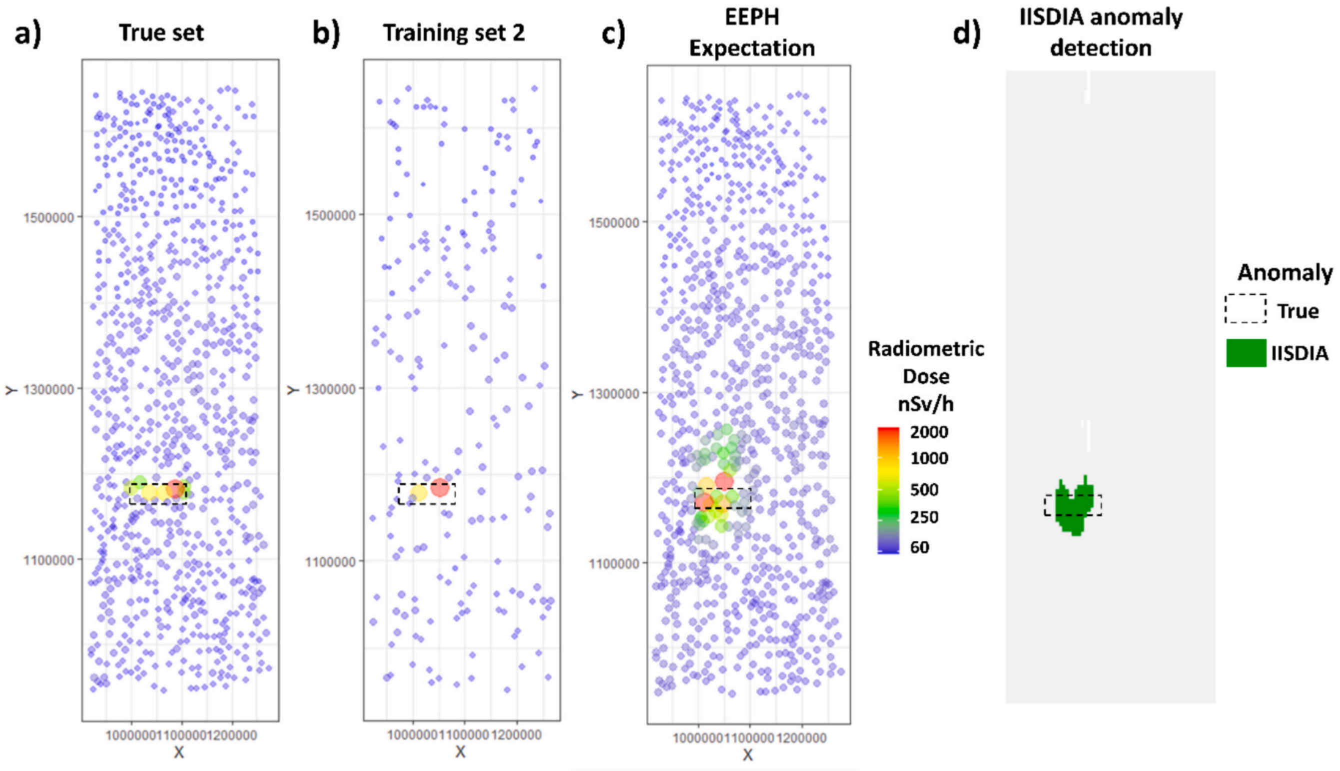

Fig. 15. EEPH tests on Dataset 2 "joker set" from the SIC2004 exercises obtained in 1 s (dataset from Dubois and Galmarini, 2005). a) true set 1, b) training set 1, c) EEPH expectation on validation set, and d) IISDIA spatial anomaly detection.

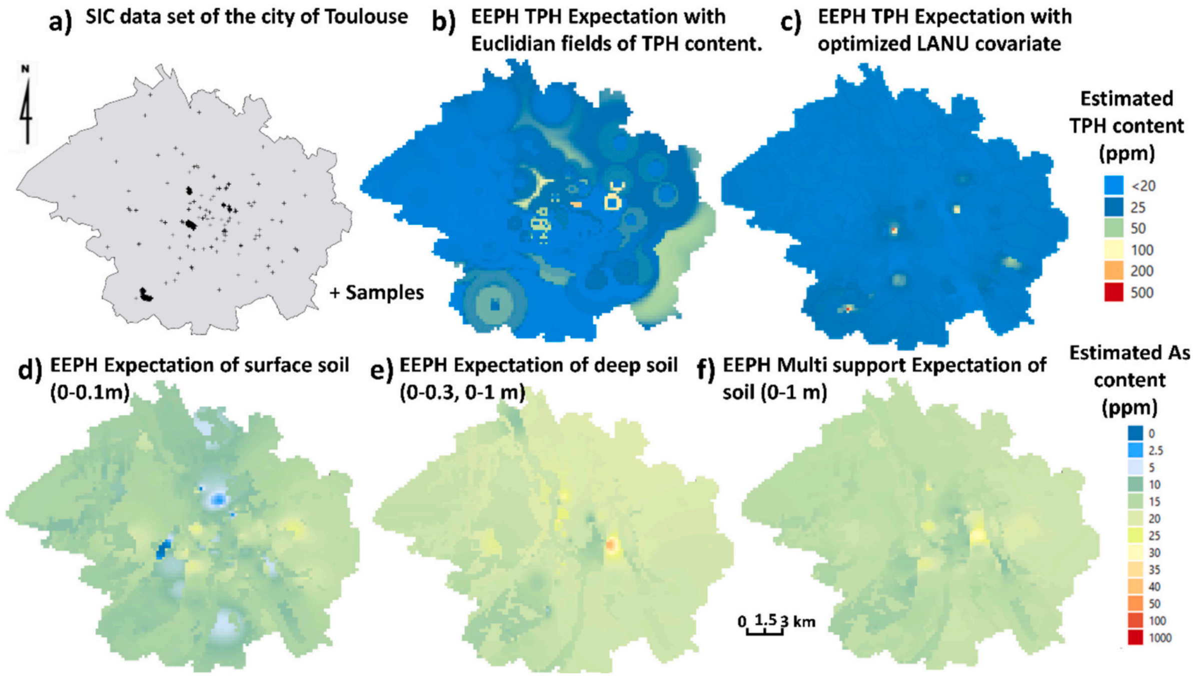
a) SIC data set of the city of Toulouse
N
+ Samples
b) EEPH TPH Expectation with Euclidian fields of TPH content.
c) EEPH TPH Expectation with optimized LANU covariate
Estimated TPH content (ppm)
<20
25
50
100
200
500
d) EEPH Expectation of surface soil (0-0.1m)
e) EEPH Expectation of deep soil (0-0.3, 0-1 m)
f) EEPH Multi support Expectation of soil (0-1 m)
Estimated As content (ppm)
0
2.5
5
10
15
20
25
30
35
40
50
100
1000
0 1.5 3 km

Fig. 16. Two EEPH interpolation methods for mapping total petroleum hydrocarbon (TPH) and EEPH Multi-support soil arsenic concentration in the city of Toulouse. a) SIC dataset b) EEPH interpolation with Euclidean fields from classes of TPH concentrations, c) EEPH expectation of TPH with optimized LANU covariate, d) EEPH expectation of As with surface soil sample only (0–0.1 m), EEPH expectation of As with deeper samples only (0–0.3, 0-1 m), f) EEPH expectation of As multi-support in soil (all samples between 0 and 1 m) with geologic covariable. Topsoil samples from Belbeze et al. (2019), $n = 139$, 0-10 cm, TPH analysis by multiple laboratories, LOQ: 10 mg/kg (8 samples), 20 mg/kg (57 samples). Samples from Belbeze et al. (2019), $n = 822$, 0–0.10 cm, 0–0.30 cm, 0–1 m, As analysis by multiple laboratories, LOQ 1 mg/kg (33 samples), LOQ 10 mg/kg (789 samples).

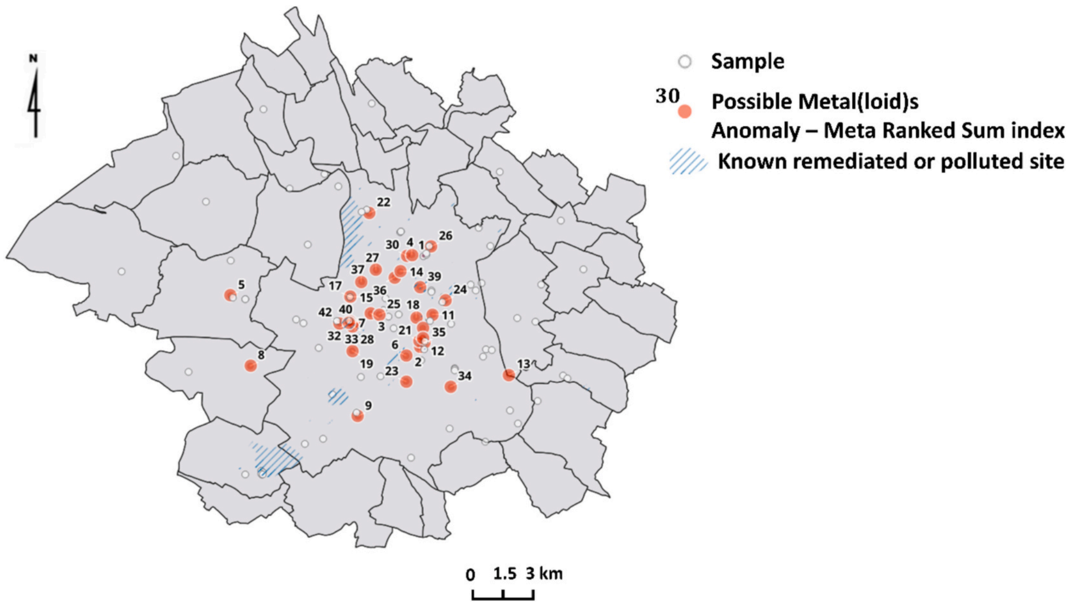

Fig. 17. Metal(oid) anomalies (43) in ITA3, 138 surface samples, as pinpointed by the IISDIA detection algorithm. This map highlights the contamination of highly urbanized areas. This observation is consistent with the presence of historic urban anthropogenic deposits.

Table 4
Metal(oid) anomalies in ITA3 as located on 1950s aerial surveys.

Anomaly n°	1950s Aerial Survey
10,43,38,40,7,32,31,33,28	Large ammunition factory
42	Waste disposal and/or fire test of ammunition Factory
27	Waste disposal proximity
36,29,15,16,17	Backfilled areas
39	Old railway facilities
8,9,4,30,1,26,13	Agricultural
34	Agricultural with traces of spraying
5,25,14,23	Old buildings
3,6	Gardens near main river
24,18,12,41,20,21,35,2	City Parcs and green zones

3.4. *Anomaly detection*

Anomaly detection is one of the key steps in environmental studies and geochemical exploration (Carranza, 2009). An anomaly is an observation that falls outside of the range of expected values and therefore can be indicative of, e.g., a mining prospect, a contamination hotspot or a measurement error. In the latter case, such values are dubbed “outliers”. Because there is a large variety of anomaly detection methods, some authors suggest it is beneficial to combine some of them (e.g., Ballabio et al., 2024). Three fundamental techniques are used to detect anomalous values of element concentrations (Lalor and Zhang, 2001; Zhang et al., 2009): (i) detection on boxplots (‘range outliers’); (ii) detection on biplots or multidimensional projections (‘relationship outliers’), and (iii) anomalous patches of values (‘spatial outliers’). A range outlier is an elevated value that is higher than an anomaly threshold. A relationship outlier is a value that does not follow the multivariate relation found with the remaining data. A spatial outlier is a high value patch surrounded by low values.

In order to build a high-performance detection system, we have selected seven methods to identify anomalies of range and spatial outliers. One method involves applying an anomaly threshold as described by Reimann et al. (2018), utilizing quantile 95 (Q95), quantile 98 (Q98), and the Tukey theoretical percentile (Tukey Inner Fence, or TIF) as effective tools for detecting anomalous values. Another technique incorporates Zero Probability Bands as employed in Belbeze et al. (2019), where the multidimensional space of results, referred to as a mathematical manifold, is projected into two-dimensional views. In this representation, populations appear as balls or clusters, with outliers gravitating around them and being manually selected. What distinguishes outliers from populations in this approach is the presence of an empty space in the projected space. This inspired the idea of identifying spatial populations within a band of zero probability or, alternatively, locating a plane that separates the populations into two distinct regions. To achieve this, a support vector machine (SVM) algorithm, as introduced by Vapnik (1995), is employed.

On a neutral EEPH interpolation map that magnifies anomalies before applying algorithms, the C-A fractal, as suggested by Carranza (2009), is used in the context of fractal theory as proposed by Cheng (1999a, 1999b). This model aligns with the theory of concentration variations by area as a function of concentration, acting as a refined technique to separate geochemical background from anomalies based on the anomaly surface, with the background assumed to dominate the collected data. Using EEPH as the baseline, it becomes feasible to plot multiple C-A curves corresponding to calculated quantiles, thereby validating the technique's conclusions. Additionally, the singularity index, as described by Xiao et al. (2016), is derived within the framework of fractal theory and is computed using a sliding window approach. Furthermore, the local Moran index, based on Anselin (1995), shares similarities with the variogram of geostatistics and is notably sensitive to deviations from data normality. It is typically calculated on values transformed by methods such as Box-Cox or normal score, using a predefined neighbourhood distance, as noted by Zhang et al. (2009). All clustering methods exhibit sensitivity to anomalous values, which predominantly occupy most of the identified clusters. As a result, clustering can be conducted in a manner where the initial outputs highlight anomalies. For the ISLANDR project, an effective spatial clustering method commonly employed in brain imaging, known as spatial fuzzy C-means (SFCM), was adopted based on the work of Cai et al. (2007) and Zhao et al. (2013). This method was implemented using *geocmeans*, an R-based application developed by Gelb and Apparicio (2021), which demonstrates strong performance on large PC configurations. Additionally, the Nemerow index, also referred to as ratios, enrichment factors, or response factors, is a prevalent tool in geochemistry due to its simple design and minimal underlying assumptions. This index typically involves identifying a geochemical background in deep soil or nearby controls and measuring its enrichment using a ratio, with interpretation

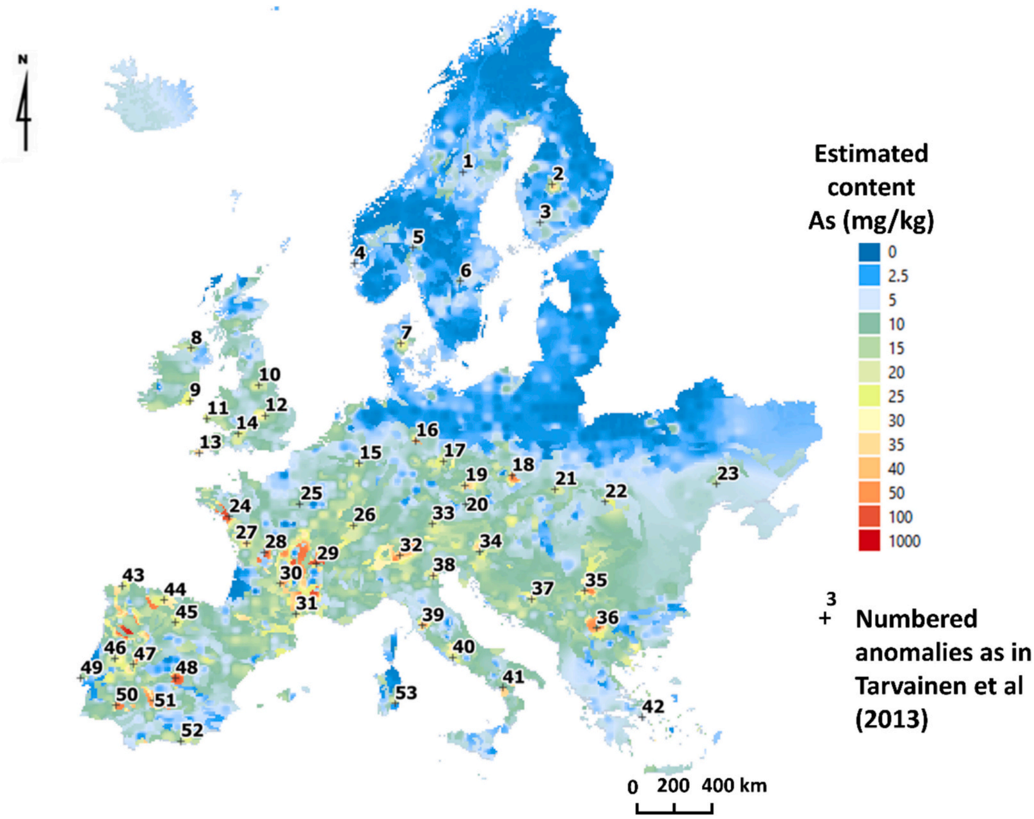

Fig. 18. EEPH Colour surface Continuous map of As with geological covariable. Numbered anomalies as in Tarvainen et al. (2013). GEMAS Survey, Ap (0–20 cm), < 2 mm, $n = 2217$, 1 site/2500 km2, aqua regia, ICP-MS. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

criteria varying based on the index's construction methodology.

By maintaining and applying these techniques to datasets, the detection algorithms will behave like a panel of experts whose consensus can be sought. Anomaly detection tests were conducted on the GEMAS dataset (Tarvainen et al., 2013) with seven techniques, which all produced informative maps (see Supplement 1). Each point is associated with seven anomaly descriptors, which are as many criteria, denoted C_1, C_2, …, C_n. The criteria are partitioned using a 5-item Likert scale (Likert, 1932) fuzzification (Table 2).

Tested on real datasets, each technique has its own advantages and disadvantages that must be considered. This uncertain information can be modelled using, e.g., triangular fuzzy numbers (Bouchon-Meunier and Marsala, 2003). Since one criterion may be more relevant than others, an overall weight of w_1, w_2, …, w_n is assigned to each criterion (Table 3). It is a subjective choice made by the authors based on results observed on test data sets (Supplement 1). These weights may be subject to adjustment according to the degree of sensitivity to strong and weak anomalies.

Fuzzy multi-criteria decision-making methods (FuzzyMCDM – FMCDM) are then used for classification problems where uncertainty, vagueness, and/or imprecision are present in the decision matrix (Ceballos et al., 2016, 2018). There are several ways to apply these multi-criteria sorting equations to fuzzy numbers, including the VIKOR method (Opricovic, 2011), the TOPSIS method (Wang and Chang, 2007) and the multi-MOORA method (Baležentis and Baležentis, 2014). By construction, the sorting results produced by TOPSIS and multi-MOORA are generally similar, whereas those produced by the various VIKOR variants show variability. In a way, this variability reproduces the natural variability of expert responses to criteria questionnaires. In this way, the consensus can be reproduced by a majority vote pass, known as meta-ranking.

4. Experiments and results

In this section, we describe the computational experiments performed to assess our EEPH and anomaly-detection methods on two synthetic dataset and three real datasets.

4.1. Core sample from Dahlberg (1975)

To refine these algorithms and illustrate their effects, a dataset of core samples from Dahlberg (1975), with only 13 data points, was adapted as a basis for working with small numbers of imprecise data points (Fig. 9a). With this type of dataset, it is impossible to construct a variogram and therefore to perform any sort of kriging. The EEPH was tested on this dataset to visually evaluate the EEPH against manual interpretations (Fig. 10b). Specifying an anisotropy in the EEPH interpolation of the 13 data points allows us to recover the two possible interpretations made by geologists on this dataset (Dahlberg, 1975; compare Fig. 9c and d to Fig. *1*10b and 10c). It should also be noted that the map produced with altered range (Fig. 10a) is equivalent Dahlberg's manual smoothing (Fig. 9b). The calculation on the 13 datasets (Dahlberg, 1975), four of which have been censored at 7 m thickness (LoQ; limit of quantification), is shown in Fig. 11. This figure shows the strong effect of <LoQ values on mapping. The expected value calculated by EPH decreases, reflecting the wide spread of the probability density. Similarly, our content may be affected by uncertainty, usually expressed as a relative percentage ± E, for example a Sand (S) content of 5 m ± 10 %. The values therefore fall within a high-low range, calculated using the relative error, which can vary from sample to sample. To remedy this while adding a minimum of assumptions to the calculation, as for the LoQ problem we propose a discretization approach. The concentration range is replaced by m discretized values (regularly sampled along the interval, 0 - LoQ) between high and low range, each yielding an EEPH

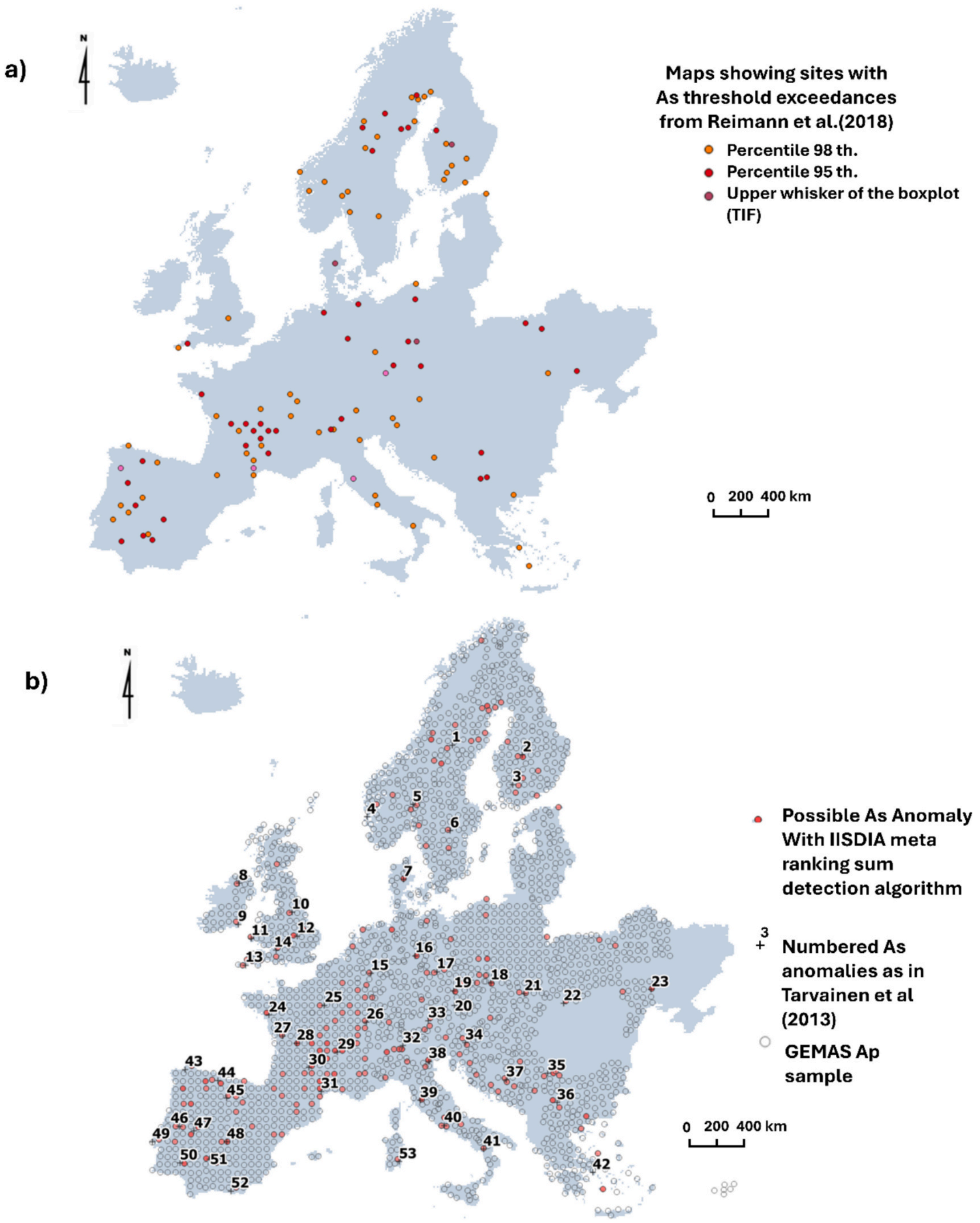

Fig. 19. a) Anomalies in GEMAS As Ap samples as pinpointed by Reimann et al. (2018) and b) IISDIA detection algorithm. GEMAS Survey, Ap (0–20 cm), < 2 mm, n = 2217, 1 site/2500 km^2, aqua regia, ICP-MS. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

calculation, then an agglomeration with a weight q_k that is equal to its probability of occurrence. In this case, the draw is based on a uniform distribution to avoid giving the uncertainty any particular shape. Nevertheless, the algorithm is still compatible with a possibilistic approach, and the error distribution could take many shapes. As uncertainty increases, the expected value calculated by the EEPH decreases, reflecting the spread of the probability density under the effect of uncertainty (Fig. 11b). While the maps presented here are expectations, at any time quantiles can be extracted from the spatial probability tensor or the spatial cdf tensor (Fig. 11c). In fact, it is possible to construct fuzzy numbers or possibilities for each pixel (Fig. 12). The Fig. 13b shows a dataset of HYRISK inputs for our dataset of thirteen core samples. It is purely hypothetical and serves only to demonstrate that, through expertise, core loss measurement, and representativeness, the measurement distributions can be specified as possibilistic priors. Once the data has been imported, the HYRISK version of the EEPH calculates the expectation or probability distribution, based on the probability-possibility distribution of the content at each point. When compared with previous figures for the same site, this Fig. 13a shows the strong influence of the hypothesized uncertainty on the sand deposit's overall structure.

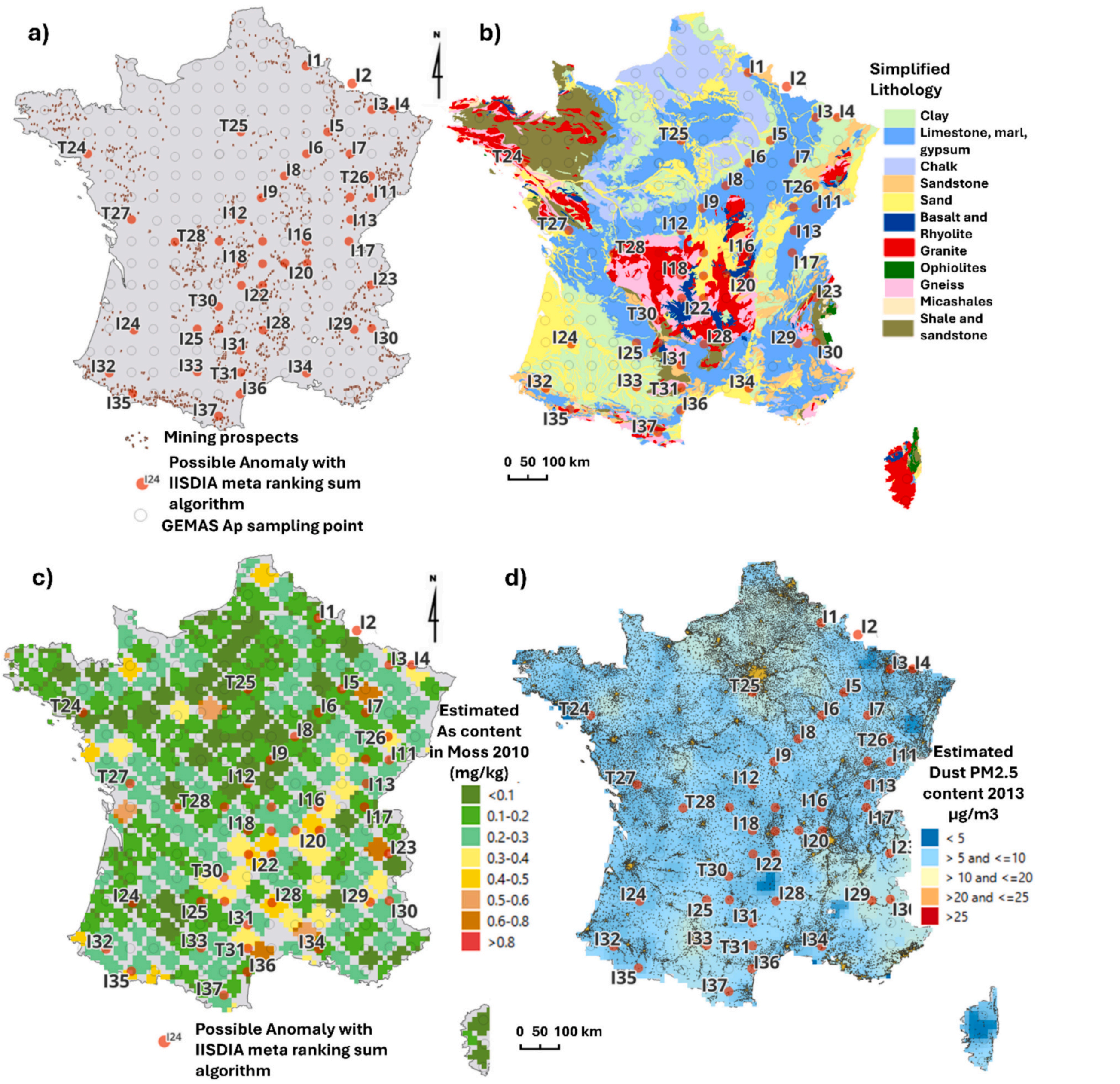

Fig. 20. a) Anomalies in GEMAS FRANCE As Ap samples as pinpointed by IISDIA algorithm and background maps used in interpretation based on mining prospect anomaly databases (Billa et al., 2016), b) a simplified lithology digital geological map (https://infoterre.brgm.fr/), c) Heavy metals in moss survey result (Harmens et al., 2010), and d) EEPH of dust measurement from Targa et al. (2023). Selected France GEMAS Survey, Ap (0–20 cm), < 2 mm, n = 2217, 1 site/2500 km^{2x}, aqua regia, ICP-MS. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2. The spatial interpolation intercomparison test (SIC2004)

It is possible to situate EEPH in its neutral version in relation to other techniques without taking uncertainty into account, applying it to complex intercomparison datasets such as SIC2004 (Dubois and Galmarini, 2005). This test, designed for emergency mapping of radiological incidents, has a dataset of 1008 data points from which 200 training values were extracted (Dataset 1), and another dataset of 1008 data points where a radiological incident "anomaly" was added before extracting the training (Dataset 2, designated as the "joker dataset"). Various more-or-less automatic techniques were then tested by the participants to estimate the 808 remaining from the 200 and their respective deviations from the mean. The trials were statistically analyzed in relation to this objective, and an overall indicator was created by adding up the scaled deviation statistics. This test is useful for comparison with the EEPH methodology because it requires fast, neutral interpolation without covariates, Also, the joker dataset has a radiological background, is a natural dataset, and is not derived from a model. The joker dataset presents an anomaly within the background, that must be identified (an objective of the ISLANDR project). The test as carried out at the time focused only on the restitution of the mean in the presence of the outlier.

Among the participants in the SIC2004 intercomparison test, two apply methods that are related to the proposed EEPH: one uses deviations between measurements instead of EEPH distances (Fang, 2005) and another uses a neural network equipped with a Nadaraya-Watson Kernel (Timonin and Savelieva, 2005), which also takes the prize with its overall score. Nevertheless, with 200 structured data points, these datasets are neither sparse nor clustered, which does not favour EEPH (it only delivers its full power if $n < 50$), but they do highlight its role as an

Table 5
Arsenic (As) anomalies in GEMAS France data.

Map number	Anomaly type
T24	Mineralization/geology; Armorican shear zone with As, Sb, Au mineralizations
T25, i5, i6, i8, i9, i3, i7	Mineralization/geology; from SW to NE: Permo-triassic sandstone enriched in As (unconformity), black marl of Middle Jurassic enriched in As (disseminated sulphides), albo-cenomanian contact: glauconitic sandstones and black marls and chalk enriched in As
T26	Geology/mineralization; partly inherited from the Hercynian per-granitic mineralization (W, Cu, etc.) and late tectonic sulphide veins but also As in Jurassic black marls
T27	Mineralization/geology; As, Co, U vein type mineralization and main shear zone (SW Armorican)
T28	Geology; Hercynian granite in Jurassic black marls
T29	Geology/mineralization; Argentat deep fault, perigranitic thermal aureoles and epithermal mineralization in the Auvergne quaternary volcanics
T30	Mineralization/mining; As anomalies related to the La Baume (Pb—Zn) and Carmaux (Coal) abandoned mines
T31	Mineralization/geology/anthropogenic; the NW part is clearly related to the major gold–arsenopyrite deposit of Salsigne (mesothermal gold) and the SE part is related to pesticides used in orchards and vineyards
i2	Mineralization prospect; ranked A1; Fe-TI; Virton
i28	Mineralization prospect; ranked A1; W-(Ba); Fustugères
i20	Mineralization prospect; ranked A1; Cu-Co-Ni-Pb-Zn; Beaujolais
i18	Mineralization prospect; ranked A2; W-(As—Pb)
i35	Mineralization prospect; ranked A2; Co-Zn-Ni
i7	Mineralization prospect; ranked A3; Ni-Cr-(Zn)
i14	Mineralization prospect; ranked A3; As
i16	Mineralization prospect; ranked A3; As-Ba
i21	Mineralization prospect; ranked A3; Pb
i22	Mineralization prospect; ranked A3; As—Pb
i23	Mineralization prospect; ranked A3; Cr-(K—Ba)
i26	Mineralization prospect; ranked A3; Ba
i31	Close proximity with mineral prospect; ranked A3; As-Ag
i1	Anthropogenic possible in a forest which was an agricultural field in 1950's
i3, i36	Anthropogenic near big city
i4	Anthropogenic possible steelmaking and metallurgical processing
i10, i11, i12, i13, i15, i17,i19, i24,i25, i27, i29,i30, i32, i33, i34	Possible Anthropogenic with diffuse proximity to cities, land spreading, pesticides or other unknow causes
I27, i37	Possible anthropogenic related to pesticide in vineyards or orchards as seen in the 1950's

Notes: Priority 1 (noted A1) anomalies are already known and have been the subject of additional investigation during the inventory. However, metals currently being investigated may warrant a reassessment of their potential. Priority 2 (A2) anomalies are often less well-known and would warrant further field checks and re-sampling or re-analysis for those that have been the subject of past geochemical analyses. Priority 3 (A3) anomalies are often associated with non-strategic metals (such as Pb—Zn) with a geological environment that gives little chance of finding large accumulations at depth.

anomaly enhancer and its extremely fast and reliable mapping capabilities.

The tests were conducted with EEPH in a moving window of 10 neighbors to anticipate any non-stationarity, and with sub-second computation time (0.15 s for 808 interpolations for Dataset 1 (Fig. 14) and 1 s for Dataset 2 with anomaly detection (Fig. 15) on a portable PC i5. The formal results for Dataset 1 yield a score of 26, very close to the 24–25 obtained by experts in geostatistical modelling (note that, for the experts, the outliers are removed and the covariance or structure is determined by a human). On the dataset with the joker outlier, the score is not as good (120), but this is the desired effect for the neutral EEPH: to move away from the endless calculation of the mean and amplify the data just enough to find the anomalies we are looking for (Fig. 14).

4.3. Topsoil samples from the TOULOUSE metropolis (Belbeze et al., 2019)

Toulouse metropolis is located in the Haute-Garonne department, in the Occitanie region, southern France. It is one of 20 largest French metropolises, with an intercommunal structure and is focused on the city of Toulouse. Toulouse Metropolis was chosen as a pilot agglomeration to demonstrate the operational feasibility and provide support for the methodology for excavated soil reuse. Analyses of surface soils sampled in schools (40 samples) were supplemented by samples obtained during potentially contaminated site diagnostic studies (1442 samples) commissioned by the Metropolis, and during two additional sampling campaigns carried out by BRGM. The latter campaigns resulted in the collection of 138 high-quality surface soil samples and 100 deep soil samples taken every meter in 20 m-deep boreholes. Sampling density was approx. One sample per km^2 (Fig. 16a). Analyses covered 24 parameters: metallic trace elements, PAHs, total cyanides, phenol index, PCB, BTEX, Sum of light hydrocarbons C5 —C10 hydrocarbons, sum of C10-C40 hydrocarbons and dioxins; see Belbeze et al. (2019) for more information. We consider the results for Total Petroleum Hydrocarbon (TPH), which are sparse, imprecise and clustered. This real-world SIC dataset is therefore useful for assessing the algorithm's self-learning potential in terms of spatial range. The main contributors to surface soil concentrations in the city are road traffic and polluted sites.

A first test was carried out using EEPH in autovariography mode. For SIC data such as this, results appear less convincing, displaying characteristic circular bullseye profiles (Fig. 16b), when compared with an EEPH with optimized covariates (Fig. 16c). This phenomenon is linked to the small amount of data (sparsity). It is interesting to note that experts tend to impose a continuous covariance function on their data, not because the natural covariance is continuous, but because it is necessary for the calculation. Nevertheless, this type of calculation makes it possible to visually observe the sampling gaps between the various circles of influence. Furthermore, we will see in EEPH continuity processing (Fig. 16c), when the concentration variation is genetically linked to a known covariate such as land use, that EEPH captures the data structure on this covariate. There is no need for autovariography. If we want to estimate the soil content of the first metre of the city, we need to mix various samples taken from the surface (Fig. 16d) and at depth (Fig. 16e), enabling us to test the EEPH in multi-support mode (Fig. 16f). This mapping is consistent with the geochemical background of the area as calculated with conventional methods by Belbeze et al., 2019.

Finally, we tested the anomaly ensemble prediction on all metals and metalloid from surface analyses and obtained the Fig. 17.

Examining this figure shows that when a sample is taken from a known remediated or polluted zone, it detects it as an anomaly. If the polluted zone is not covered by a sample, it is not detected, as the notion of scale survey does not allow it to be seen. Nevertheless, new zones have been identified compared to Belbeze et al. (2019) and merit further verification using historical databases. As shown in Table 4, the city deconstructs and rebuilds on itself, preserving in some of its soils trace pollutants from the past.

4.4. GEMAS Project (Tarvainen et al., 2013)

The GEMAS dataset is a harmonized geochemical dataset of agricultural soil throughout Europe gathered by the Association of Geological Surveys of Europe (EuroGeoSurveys) in cooperation with Eurometaux in 2008. The average sampling density was 1 sample from a

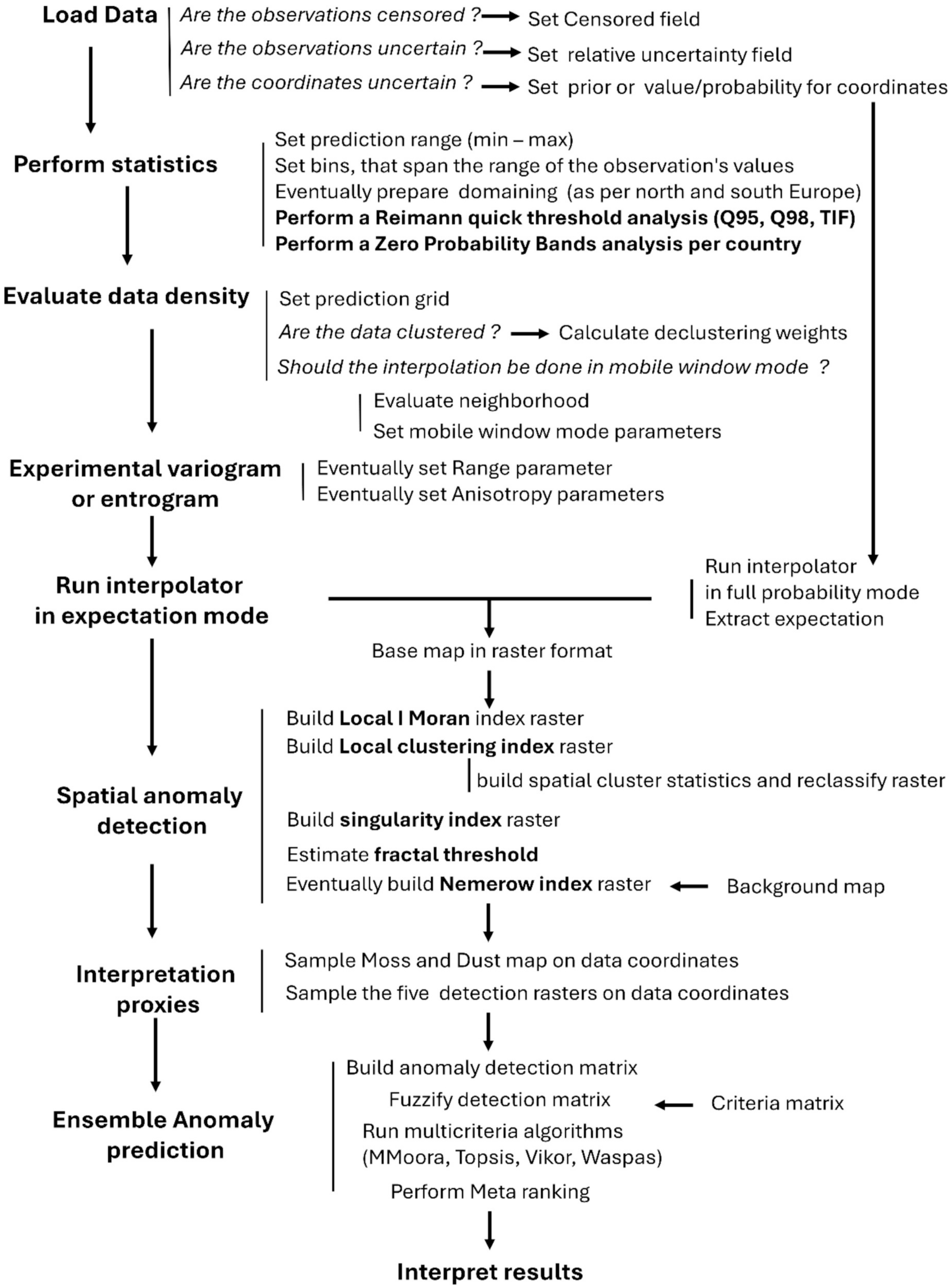
Load Data
Are the observations censored ?
Set Censored field
Are the observations uncertain ?
Set relative uncertainty field
Are the coordinates uncertain ?
Set prior or value/probability for coordinates
Perform statistics
Set prediction range (min – max)
Set bins, that span the range of the observation's values
Eventually prepare domaining (as per north and south Europe)
Perform a Reimann quick threshold analysis (Q95, Q98, TIF)
Perform a Zero Probability Bands analysis per country
Evaluate data density
Set prediction grid
Are the data clustered ?
Calculate declustering weights
Should the interpolation be done in mobile window mode ?
Evaluate neighborhood
Set mobile window mode parameters
Experimental variogram or entrogram
Eventually set Range parameter
Eventually set Anisotropy parameters
Run interpolator in expectation mode
Run interpolator in full probability mode
Extract expectation
Base map in raster format
Build Local I Moran index raster
Build Local clustering index raster
build spatial cluster statistics and reclassify raster
Spatial anomaly detection
Build singularity index raster
Estimate fractal threshold
Eventually build Nemerow index raster
Background map
Interpretation proxies
Sample Moss and Dust map on data coordinates
Sample the five detection rasters on data coordinates
Build anomaly detection matrix
Fuzzify detection matrix
Criteria matrix
Ensemble Anomaly prediction
Run multicriteria algorithms (MMoora, Topsis, Vikor, Waspas)
Perform Meta ranking
Interpret results

Fig. 21. IISDIA detection algorithm as applied to GEMAS Survey, Ap (0–20 cm), < 2 mm, n = 2217, 1 site/2500 km2, aqua regia, ICP-MS. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

50 × 50 km grid cell and sampling depth was 0–20 cm. Arsenic (As) concentrations in European agricultural topsoils (Ap – ploughing soil layer) have been analyzed using aqua regia extraction of the <2 mm size fraction. For more information, see Reimann et al. (2014a, 2014b). Algorithms were developed and tested on the European Community's large-scale GEMAS survey using a moving window so that a small data set (SIC) was processed. The GEMAS Ap measurements for arsenic are particularly noteworthy. Tarvainen et al. (2013) identified 52 anomalies for EU including 7 for France that can be tested against the IISDIA algorithm. Most of the identified anomalies originated from geogenic

sources. Thus, if the EEPH of arsenic levels is calculated with geological information reduced to a 10 km grid as parameter, we obtain the map presented in Fig. 18, which presents an improved continuity that is remarkably congruent with the anomaly marking carried out by Tarvainen et al. (2013).

This map is not as smooth as the usual maps, such as KED and GWE, but it is a map showing all the anomalies (no data has been discarded) and with minimized underlying assumptions. Geologists will note that the shape of the anomalies matches the orientation of the layers and fractures. It is also worth noting that the anomalies identified by Tarvainen et al. (2013) and Reimann et al. (2018) are included (Fig. 19), as well as new low-signal anomalies, the detection of which is one of the main objectives of the ISLANDR project.

For France (Fig. 20), data interpretation is facilitated by several sources of information: a comprehensive mining prospect anomaly database (Billa et al., 2016); a polluted and remediated sites database (Darmendrail, 2003); a simplified lithology digital geological map of France (https://infoterre.brgm.fr/), heavy metals in a moss survey (Harmens et al., 2011), and dust measurement from Targa et al. (2023). It is then possible to propose a first interpretation of the 17 new low-level anomalies detected in addition to the already 20 anomalies from Reimann et al. (2018).

The IISDIA algorithm allowed us to detect twelve (12) major arsenic mineralizations, three (3) prospect mineralization ranked A1 (already known and needing re-assessment), two (2) ranked A2 (less well-known and needing field-check) and eight (8) A3 (non-strategic). The remaining 20 anomalies cannot be explained by the extensive geological knowledge of the French territory and seem to be associated with anthropic contamination linked to dispersion from agriculture, backfills and pesticides, especially in vineyards and fruit tree orchards. Compared with previous studies (Tarvainen et al., 2013; Reimann et al., 2018), the proposed algorithm enables the detection of new arsenic anomalies in the GEMAS France data, including 20 new ones that could come from the diffuse anthropogenic background (Table 5). This example illustrates the performance of the proposed IISDIA algorithms on a large-scale survey of agricultural arsenic levels such as GEMAS.

5. Conclusions and perspectives

We propose an innovative interpolation and anomaly detection algorithm especially adapted to cases where data is sparse (⟨30), clustered and/or uncertain (Fig. 21). Because one of the main objectives of the proposed algorithm is to detect anomalies in spatial data, it avoids the «smoothing» effect of more classical methods such as kriging with variograms and unlike kriging and other deep-learning algorithms, the proposed IISDIA algorithm can function with small datasets (on the order of, e.g., 10 data points). The proposed algorithm is currently under development to allow additional capabilities, e.g., detect stream sediment geochemical anomalies, carry uncertain information for measurements in three dimensions, trace complex pollutant patterns and help build geological models. This research also focuses on the issue of epistemic uncertainties in practical situations faced by geologists and geochemists.

With SIC data, there is little chance of knowing the true joint distributions. As for any model, we must therefore be cautious regarding the parameters we introduce. Compared to kriging (Berton, 2018), EEPH shows superior performance in the case of sparsely-sampled phenomena (typically $n < 30$). Kriging has been shown to work best when there is a significant amount of data available (generally speaking, any weighted sum of neighbouring content tends on average towards the true mean value when n is large—the so-called law of large numbers, and the dependency between data depends only on the distance between points). With the proposed IISDIA algorithm, which is focused largely on anomaly detection, we avoid (i) attributing false continuity to available soil pollution measurements and (ii) smoothing the data. A recommended approach for mapping sites is therefore to first examine the data and the associated variogram or other spatial measures, to see if kriging or machine learning algorithms are applicable. If not, EEPH is a viable alternative. It is reminded that EEPH makes no assumptions regarding data variability, particularly in terms of continuity and is sensitive to outliers, which it magnifies. The possibility of inserting uncertainty treatment and optimization into EEPH are promising and will be pursued. Also, it may be possible to improve covariate and Dirac function propagation, using proper entrogram (directional entropy propagation estimator) management (Bianchi and Pedretti, 2018).

CRediT authorship contribution statement

Stéphane Belbèze: Supervision, Methodology, Formal analysis, Conceptualization. **Jérémy Rohmer:** Validation, Investigation, Conceptualization. **Dominique Guyonnet:** Validation, Methodology. **Philippe Négrel:** Validation, Writing – review & editing. **Timo Tarvainen:** Methodology. **Timo TARVAINEN:** Writing – review & editing. **Stéphane BELBÈZE:** Writing – original draft. **Jérémy ROHMER:** Writing – review & editing. **Dominique GUYONNET:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the ISLANDR project (Information-based Strategies for LAND Remediation). ISLANDR has received funding from the European Union's Horizon Europe Research and Innovation Program under Grant Agreement 101112889. The interpolation benchmarking presented herein was supported in part by the French National Research Agency within the HOUSES project (ANR-22-CE56-0006). We would like to thank our ISLANDR reviewers for the timely and very helpful reviews of the initial manuscript. The authors would like to thank the two anonymous reviewers and the Editorial board for helping to improve the quality of this paper with their comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gexplo.2025.107868.

Data availability

Data will be made available on request.

References

Aberle, M.G., Robertson, J., Hoogewerff, J.A., 2023. Voronoi Natural Neighbours Tessellation: an interpolation and grid agnostic approach to forensic soil provenancing. For. Chem. 35, 100522. https://doi.org/10.1016/j.forc.2023.100522.

Albanese, S., De Vivo, B., Lima, A., Cicchella, D., 2007. Geochemical background and baseline values of toxic elements in stream sediments of Campania region (Italy). J. Geochem. Explor. 93, 21–34. https://doi.org/10.1016/j.gexplo.2006.07.006.

Andrade, R., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Faria, W.M., Mesquita, L.F., Guilherme, L.R.G., Curi, N., 2020. Assessing models for prediction of some soil chemical properties from portable X-ray fluorescence (pXRF) spectrometry data in Brazilian Coastal Plains. Geoderma 357, 113957. https://doi.org/10.1016/j.geoderma.2019.113957.

Anselin, L., 1995. Local Indicators of Spatial Association—LISA. Geogr. Anal. 27, 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Bai, C.-Z., Zhang, R., Hong, M., Qian, L., Wang, Z., 2015. A new information diffusion modelling technique based on vibrating string equation and its application in natural disaster risk assessment. Int. J. Gen. Syst. 44, 601–614. https://doi.org/10.1080/03081079.2014.980242.

Baležentis, T., Baležentis, A., 2014. A survey on Development and applications of the Multi-criteria Decision making Method MULTIMOORA: a SURVEY ON

DEVELOPMENT AND APPLICATIONS OF MULTIMOORA. J. Multicrit. Decis. Anal. 21, 209–222. https://doi.org/10.1002/mcda.1501.

Ballabio, C., Jones, A., Panagos, P., 2024. Cadmium in topsoils of the European Union – an analysis based on LUCAS topsoil database. Sci. Total Environ. 912, 168710. https://doi.org/10.1016/j.scitotenv.2023.168710.

Bartkute, V., Sakalauskas, L., 2008. Experimental probabilistic hypersurface construction by Gaussian fields. In: Institute of Mathematics and Informatics. Vilnius 08663, Lithuania.

Baudrit, C., Dubois, D., Guyonnet, D., 2006. Joint Propagation and Exploitation of Probabilistic and Possibilistic Information in Risk Assessment. IEEE Trans. Fuzzy Syst. 14, 593–608. https://doi.org/10.1109/TFUZZ.2006.876720.

Baudrit, C., Guyonnet, D., Dubois, D., 2007. Joint propagation of variability and imprecision in assessing the risk of groundwater contamination. J. Contam. Hydrol. 93, 72–84. https://doi.org/10.1016/j.jconhyd.2007.01.015.

Beauzamy, B., 2004. Méthodes probabilistes pour l'étude des phénomènes réels. Les mathématiques du réel. Société de calcul mathématique, Paris.

Behrens, T., Schmidt, K., Viscarra Rossel, R.A., Gries, P., Scholten, T., MacMillan, R.A., 2018. Spatial modelling with Euclidean distance fields and machine learning. Eur. J. Soil Sci. 69, 757–770. https://doi.org/10.1111/ejss.12687.

Belbeze, S., Djemil, M., Béranger, S., Stochetti, A., 2019. Détermination de FPGA - Fonds Pédo-Géochimiques Anthropisés urbains. Agglomération pilote : Toulouse Métropole (public No. RP-69502-FR). BRGM.

Belbèze, S., Rohmer, J., Négrel, P., Guyonnet, D., 2023. Defining urban soil geochemical backgrounds: a review for application to the French context. J. Geochem. Explor. 254, 107298. https://doi.org/10.1016/j.gexplo.2023.107298.

Berton, G., 2018. Comparison between two interpolation methods: kriging and EPH. J. Phys. Conf. Ser. 1141, 012130. https://doi.org/10.1088/1742-6596/1141/1/012130.

Bianchi, M., Pedretti, D., 2018. An Entrogram-based Approach to Describe Spatial Heterogeneity with applications to Solute Transport in Porous Media. Water Resour. Res. 54, 4432–4448. https://doi.org/10.1029/2018WR022827.

Billa, M., Gloaguen, E., Melleton, J., 2016. Consolidation des anomalies géochimiques et géophysiques du territoire métropolitain. (public No. RP-66416-FR). BRGM.

Bouchon-Meunier, B., Marsala, C., 2003. Logique floue, principes, aide à la décision. Hermès Science publications, Paris.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Buhmann, M.D., Jäger, J. 2021. Quasi-interpolation, Cambridge monographs on applied and computational mathematics. Cambridge University Press, Cambridge, United Kingdom; New York, NY.

Cai, W., Chen, S., Zhang, D., 2007. Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. Pattern Recogn. 40, 825–838. https://doi.org/10.1016/j.patcog.2006.07.011.

Carranza, E.J.M., 2009. Geochemical Anomaly and Mineral Prospectivity Mapping in GIS, 1st, ed. ed. Handbook of exploration and environmental geochemistry, Elsevier, Amsterdam Boston.

Ceballos, B., Lamata, M.T., Pelta, D.A., 2016. A comparative analysis of multi-criteria decision-making methods. Prog. Artif. Intell. 5, 315–322. https://doi.org/10.1007/s13748-016-0093-1.

Ceballos, B., Pelta, D.A., Lamata, M.T., 2018. Rank Reversal and the VIKOR Method: an Empirical Evaluation. Int. J. Info. Tech. Dec. Mak. 17, 513–525. https://doi.org/10.1142/S0219622017500237.

Cheng, Q., 1999a. Markov Processes and Discrete Multifractals. Math. Geol. 31, 455–469. https://doi.org/10.1023/A:1007594709250.

Cheng, Q., 1999b. Multifractality and spatial statistics. Comput. Geosci. 25, 949–961. https://doi.org/10.1016/S0098-3004(99)00060-6.

Chilès, J.P., Delfiner, O., 2013. Geostatistics: Modeling Spatial Uncertainty, 2nd. Wiley.

Civitillo, D., Ayuso, R.A., Lima, A., Albanese, S., Esposito, R., Cannatelli, C., De Vivo, B., 2016. Potentially harmful elements and lead isotopes distribution in a heavily anthropized suburban area: the Casoria case study (Italy). Environ. Earth Sci. 75, 1325. https://doi.org/10.1007/s12665-016-6093-4.

Dahlberg, E.C., 1975. Relative effectiveness of geologists and computers in mapping potential hydrocarbon exploration targets. Math. Geol. 7, 373–394. https://doi.org/10.1007/BF02080496.

Darmendrail, D., 2003. The French Approach to Contaminated-Land Management (public No. RP-52276-FR). BRGM.

Dhariwal, P., Nichol, A., 2021. Diffusion models beat GANs on image synthesis. https://doi.org/10.48550/ARXIV.2105.05233.

Dubois, D., Guyonnet, D., 2011. Risk-informed decision-making in the presence of epistemic uncertainty. Int. J. Gen. Syst. 40, 145–167. https://doi.org/10.1080/03081079.2010.506179.

Dubois, G., Galmarini, S., 2005. Introduction to the Spatial Interpolation Comparison (SIC) 2004 Exercise and Presentation of the Datasets. Applied GIS 1. https://doi.org/10.2104/ag050009.

Fang, K.K.B., 2005. Multi-Dimension and Real-Time Interpolation (Dirac-Monte Carlo Method) (un-published). FANG, INC., Stanton, CA 90680 U.S..

Fendrich, A.N., Van Eynde, E., Stasinopoulos, D.M., Rigby, R.A., Mezquita, F.Y., Panagos, P., 2024. Modeling arsenic in European topsoils with a coupled semiparametric (GAMLSS-RF) model for censored data. Environ. Int. 185, 108544. https://doi.org/10.1016/j.envint.2024.108544.

Ferson, S., Ginzburg, L.R., 1996. Different methods are needed to propagate ignorance and variability. Reliab. Eng. Syst. Saf. 54, 133–144. https://doi.org/10.1016/S0951-8320(96)00071-3.

Ferson, S., Troy Tucker, W., 2006. Sensitivity analysis using probability bounding. Reliab. Eng. Syst. Saf. 91, 1435–1442. https://doi.org/10.1016/j.ress.2005.11.052.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33. https://doi.org/10.18637/jss.v033.i01.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29. https://doi.org/10.1214/aos/1013203451.

Gelb, J., Apparicio, P., 2021. Apport de la classification floue c-means spatiale en géographie : essai de taxinomie socio-résidentielle et environnementale à Lyon. Cybergeo. https://doi.org/10.4000/cybergeo.36414.

Gibin, M., Longley, P., Atkinson, P., 2007. Kernel Density Estimation and Percent Volume Contours in General Practice Catchment Area Analysis in Urban Areas. Presented at the GIScience Research UK Conference (GISRUK), Maynooth - Ireland.

Godan, F., Zeydina, O., Richet, Y., Beauzamy, B., 2015. Reactor Safety and Incomplete Information: Comparison of Extrapolation Methods for the Extension of Computational Codes, in: Paper 15377. Presented at the ICAPP 2015, Nice, France, p. 5.

Grunsky, E.C., de Caritat, P., 2017. Advances in the use of geochemical data for mineral exploration, in: Proceedings of Exploration 17. Presented at the Sixth Decennial International Conference on Mineral Exploration, Toronto, pp. 451–456..

Harmens, H., Foan, L., Simon, V., Mills, G., 2011. Mosses as biomonitors of atmospheric POPs pollution: A review. Report for Defra contract No. AQ08610, Ecology and Hydrology. Environment Centre Wales, Bangor, UK.

Helsel, D.R., Helsel, D.R., Helsel, D.R., 2012. Statistics for Censored Environmental Data Using Minitab and R, 2nd ed. Wiley, Hoboken, N.J. https://doi.org/10.1002/9781118162729.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518. https://doi.org/10.7717/peerj.5518.

Hengl, T., Miller, M.A.E., Križan, J., Shepherd, K.D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S.M., McGrath, S.P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F.B.T., Yemefack, M., Wendt, J., MacMillan, R.A., Wheeler, I., Crouch, J., 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. Sci. Rep. 11, 6130. https://doi.org/10.1038/s41598-021-85639-y.

Heuvelink, G.B.M., Kros, J., Reinds, G.J., De Vries, W., 2016. Geostatistical prediction and simulation of European soil property maps. Geoderma Reg. 7, 201–215. https://doi.org/10.1016/j.geodrs.2016.04.002.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. https://doi.org/10.48550/arXiv.2006.11239.

Huang, H., Liang, Z., Li, B., Wang, D., 2019. A new spatial precipitation interpolation method based on the information diffusion principle. Stoch. Env. Res. Risk A. 33, 765–777. https://doi.org/10.1007/s00477-019-01658-2.

Huang, S., Zhou, C., Wan, Q., 1998. Primary analysis on flood disaster risk evaluation. Geogr. Res. 17, 71–77.

Ingamells, C.O., Pitard, F.F., 1986. Applied geochemical analysis, Chemical analysis. J. Wiley and sons, (New York Chichester Brisbane [etc.).

Jenny, H., 1994. Factors of Soil Formation: A System of Quantitative Pedology. Dover, New York.

Jordan, G., Petrik, A., De Vivo, B., Albanese, S., Demetriades, A., Sadeghi, M., 2018. GEMAS: Spatial analysis of the Ni distribution on a continental-scale using digital image processing techniques on European agricultural soil data. J. Geochem. Explor. 186, 143–157. https://doi.org/10.1016/j.gexplo.2017.11.011.

Kasmaeeyazdi, S., Raspa, G., De Fouquet, C., Tinti, F., Bonduà, S., Bruno, R., 2020. How different data supports affect geostatistical modelling: the new aggregation method and comparison with the classical regularisation and the theoretical punctual model. Int. J. Min. Reclam. Environ. 34, 34–54. https://doi.org/10.1080/17480930.2018.1507609.

Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. Appl. Math. Model. 81, 401–418. https://doi.org/10.1016/j.apm.2019.12.016.

Khalipova, V., Damart, G., Beauzamy, B., Bruna, G., 2018. Malfunctions in radioactivity sensors' networks. EPJ Web Conf. 170, 08002. https://doi.org/10.1051/epjconf/201817008002.

Lado, L.R., Hengl, T., Reuter, H.I., 2008. Heavy metals in European soils: a geostatistical analysis of the FOREGS Geochemical database. Geoderma 148, 189–199. https://doi.org/10.1016/j.geoderma.2008.09.020.

Lajaunie, C., 1996. In: Ecole nationale des mines de Paris (Ed.), Documentation of the Mixed Support Kriging Programs.

Lalor, G.C., Zhang, C., 2001. Multivariate outlier detection and remediation in geochemical databases. Sci. Total Environ. 281, 99–109. https://doi.org/10.1016/S0048-9697(01)00839-7.

Levine, N., 2010. Crimestat Iii: A Spatial Statistics Program for the Analysis of Crime Incident Locations (Version 3.3). Ned Levine & Associates/Washington. National Institute of Justice, Houston.

Li, J., 2012. Predicting Seabed Sand Content across the Australian Margin Using Machine Learning and Geostatistical Methods. Geoscience Australia, Canberra.

Li, J., Heap, A.D., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists. Geoscience Australia, Canberra.

Li, J., Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. Eco. Inform. 6, 228–241. https://doi.org/10.1016/j.ecoinf.2010.12.003.

Likert, R., 1932. A technique for the measurement of attitudes. Arch. Psychol. 22, 5–55.

Lindeberg, T. 1997. Scale-Space Theory in Computer Vision, 3. printing. ed, The Kluwer international series in engineering and computer science. Kluwer Acad. Publ, Boston.

Loquin, K., Dubois, D. 2010. Kriging and Epistemic uncertainty: A critical Discussion, in: Jeansoulin, R., Papini, O., Prade, H., Schockaert, S. (Eds.), Methods for Handling

Imperfect Spatial Information, Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 269–305. https://doi.org/10.1007/978-3-642-14755-5_11.

Maas, S., Scheifler, R., Benslama, M., Crini, N., Lucot, E., Brahmia, Z., Benyacoub, S., Giraudoux, P., 2010. Spatial distribution of heavy metal concentrations in urban, suburban and agricultural soils in a Mediterranean city of Algeria. Environ. Pollut. 158, 2294–2301. https://doi.org/10.1016/j.envpol.2010.02.001.

Malone, B.P., Minasny, B., McBratney, A.B. 2017. Using R for Digital Soil Mapping, Progress in Soil Science. Springer, Cham. https://doi.org/10.1007/978-3-319-44327-0.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Melleton, J., Belbeze, S., Vic, G., Auger, P., Chevillard, M., 2021. Établissement du fond pédogéochimique dans la région de l'ancien seur minier de Salsigne (Aude) (public No. RP-7067-FR). BRGM.

Minkkinen, P.O., Esbensen, K.H., 2009. Grab vs. composite sampling of particulate materials with significant spatial heterogeneity—A simulation study of "correct sampling errors.". Anal. Chim. Acta 653, 59–70. https://doi.org/10.1016/j.aca.2009.08.039.

Négrel, Ph., Sadeghi, M., Ladenberger, A., Reimann, C., Birke, M., 2015. Geochemical fingerprinting and source discrimination of agricultural soils at continental scale. Chem. Geol. 396, 1–15. https://doi.org/10.1016/j.chemgeo.2014.12.004.

Négrel, Ph., Ladenberger, A., Reimann, C., Birke, M., Demetriades, A., Sadeghi, M., 2019. GEMAS: Geochemical background and mineral potential of emerging tech-critical elements in Europe revealed from low-sampling density geochemical mapping. Appl. Geochem. 111, 104425. https://doi.org/10.1016/j.apgeochem.2019.104425.

Opricovic, S., 2011. Fuzzy VIKOR with an application to water resources planning. Expert Syst. Appl. 38, 12983–12990. https://doi.org/10.1016/j.eswa.2011.04.097.

Panagos, P., Meusburger, K., Ballabio, C., Borrelli, P., Alewell, C., 2014. Soil erodibility in Europe: a high-resolution dataset based on LUCAS. Sci. Total Environ. 479, 189–200. https://doi.org/10.1016/j.scitotenv.2014.02.010.

Parzen, E., 1962. On Estimation of a Probability Density Function and Mode. Ann. Math. Stat. 33, 1065–1076. https://doi.org/10.1214/aoms/1177704472.

Pereira, B., Titeux, H., Schneider, A., Sonnet, P., 2012. Rapport Final du Projet Pollusol 2 Partie « sols », SPAQuE. UCL-ELI.

Petrik, A., Thiombane, M., Albanese, S., Lima, A., De Vivo, B., 2018. Source patterns of Zn, Pb, Cr and Ni potentially toxic elements (PTEs) through a compositional discrimination analysis: a case study on the Campanian topsoil data. Geoderma 331, 87–99. https://doi.org/10.1016/j.geoderma.2018.06.019.

Quarteroni, A., Gervasio, P., Saleri, F. 2010. Scientific Computing with MATLAB and Octave, Texts in Computational Science and Engineering. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-12430-3.

Quinlan, J.R., 1993. Combining instance-based and model-based learning, in: Proceedings. Presented at the Tenth International Conference on Machine Learning, Amherst, pp. 236–243.

R Core Team, 2022. R: A Language and Environment for Statistical Computing.

Reimann, C., Filzmoser, P., Garrett, R., Dutter, R. (Eds.). 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R. John Wiley & Sons, Chichester, England; Hoboken, NJ.

Chemistry of Europe's agricultural soils – Part A: Methodology and interpretation of the GEMAS data set. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P. (Eds.), 2014a. Geologisches Jahrbuch (Reihe B 102). Schweizerbarth, Hannover, p. 528.

Chemistry of Europe's agricultural soils – Part B: General background information and further analysis of the GEMAS data set. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P. (Eds.), 2014b. Geologisches Jahrbuch (Reihe B 103). Schweizerbarth, Hannover, p. 352.

Reimann, C., Fabian, K., Birke, M., Filzmoser, P., Demetriades, A., Négrel, P., Oorts, K., Matschullat, J., De Caritat, P., Albanese, S., Anderson, M., Baritz, R., Batista, M.J., Bel-Ian, A., Cicchella, D., De Vivo, B., De Vos, W., Dinelli, E., Ďuriš, M., Dusza-Dobek, A., Eggen, O.A., Eklund, M., Ernsten, V., Flight, D.M.A., Forrester, S., Fügedi, U., Gilucis, A., Gosar, M., Gregorauskiene, V., De Groot, W., Gulan, A., Halamić, J., Haslinger, E., Hayoz, P., Hoogewerff, J., Hrvatovic, H., Husnjak, S., Jähne-Klingberg, F., Janik, L., Jordan, G., Kaminari, M., Kirby, J., Klos, V., Kwećko, P., Kuti, L., Ladenberger, A., Lima, A., Locutura, J., Lucivjansky, P., Mann, A., Mackovych, D., McLaughlin, M., Malyuk, B.I., Maquil, R., Meuli, R.G., Mol, G., O'Connor, P., Ottesen, R.T., Pasnieczna, A., Petersell, V., Pfleiderer, S., Poňavič, M., Prazeres, C., Radusinović, S., Rauch, U., Salpeteur, I., Scanlon, R., Schedl, A., Scheib, A., Schoeters, I., Šefčik, P., Sellersjö, E., Slaninka, I., Soriano-Disla, J.M., Šorša, A., Svrkota, R., Stafilov, T., Tarvainen, T., Tendavilov, V., Valera, P., Verougstraete, V., Vidojević, D., Zissimos, A., Zomeni, Z., Sadeghi, M., 2018. GEMAS: establishing geochemical background and threshold for 53 chemical elements in European agricultural soil. Appl. Geochem. 88, 302–318. https://doi.org/10.1016/j.apgeochem.2017.01.021.

Rhind, S.M., Kyle, C.E., Kerr, C., Osprey, M., Zhang, Z.L., Duff, E.I., Lilly, A., Nolan, A., Hudson, G., Towers, W., Bell, J., Coull, M., McKenzie, C., 2013. Concentrations and geographic distribution of selected organic pollutants in Scottish surface soils. Environ. Pollut. 182, 15–27. https://doi.org/10.1016/j.envpol.2013.06.041.

Richmond, A., 2002. Two-point declustering for weighting data pairs in experimental variogram calculations. Comput. Geosci. 28, 231–241. https://doi.org/10.1016/S0098-3004(01)00070-X.

Sauvaget, B., De Fouquet, C., Le Guern, C., Renard, D., Roussel, H., 2022. Geostatistical filtering to map a 3D anthropogenic pedo-geochemical background for excavated soil reuse. J. Geochem. Explor. 240, 107031. https://doi.org/10.1016/j.gexplo.2022.107031.

Shafer, G., 1976. A Mathematical Theory of Evidence. Princeton University Press.

Sinclair, A.J., Blackwell, G.H. 2002. Applied Mineral Inventory Estimation, 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511545993.

Targa, J., Ripoll, A., Banyuls, L., González Ortiz, A., Soares, J., 2023. Status report of air quality in Europe for year 2021, using validated data (no. ETC-HE Report 2023/1) European Topic Centre on Human health and the environment (ETC-HE).

Tarvainen, T., Salminen, R., Vos, W.D. 2005. Geochemical Atlas of Europe. Part 1: Background Information, Methodology and Maps. Geological Survey of Finland, Espoo.

Tarvainen, T., Albanese, S., Birke, M., Poňavič, M., Reimann, C., 2013. Arsenic in agricultural and grazing land soils of Europe. Appl. Geochem. 28, 2–10. https://doi.org/10.1016/j.apgeochem.2012.10.005.

Timonin, V., Savelieva, E., 2005. Spatial Prediction of Radioactivity using General Regression Neural Network. Applied GIS 1. https://doi.org/10.2104/ag050019.

Tóth, G., Jones, A., Montanarella, L., 2013. LUCAS Topsoil Survey: Methodology, Data and Results., European Commission. Joint Research Centre. Institute for Environment and Sustainability. Publications Office, LU.

Tóth, G., Hermann, T., Szatmári, G., Pásztor, L., 2016. Maps of heavy metals in the soils of the European Union and proposed priority areas for detailed assessment. Sci. Total Environ. 565, 1054–1062. https://doi.org/10.1016/j.scitotenv.2016.05.115.

Van Eynde, E., Fendrich, A.N., Ballabio, C., Panagos, P., 2023. Spatial assessment of topsoil zinc concentrations in Europe. Sci. Total Environ. 892, 164512. https://doi.org/10.1016/j.scitotenv.2023.164512.

Vapnik, V.N. 1995. The Nature of Statistical Learning Theory, 1st ed. Springer New York, New York.

Visman, J., 1969. A general sampling theory. American Society for Testing and Materials (ASTM), Materials Research and Standards 9, 8–13.

Wadoux, A.M.J.-C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359. https://doi.org/10.1016/j.earscirev.2020.103359.

Wang, T.-C., Chang, T.-H., 2007. Application of TOPSIS in evaluating initial training aircraft under a fuzzy environment. Expert Syst. Appl. 33, 870–880. https://doi.org/10.1016/j.eswa.2006.07.003.

Xiao, F., Chen, Z., Chen, J., Zhou, Y., 2016. A batch sliding window method for local singularity mapping and its application for geochemical anomaly identification. Comput. Geosci. 90, 189–201. https://doi.org/10.1016/j.cageo.2015.11.001.

Xiao, S., Ou, M., Geng, Y., Zhou, T., 2023. Mapping soil pH levels across Europe: an analysis of LUCAS topsoil data using random forest kriging (RFK). Soil Use Manag. 39, 900–916. https://doi.org/10.1111/sum.12874.

Xie, Z., Yan, J., 2008. Kernel Density Estimation of traffic accidents in a network space. Comput. Environ. Urban. Syst. 32, 396–406. https://doi.org/10.1016/j.compenvurbsys.2008.05.001.

Xu, H., Zhang, C., 2021. Investigating spatially varying relationships between total organic carbon contents and pH values in European agricultural soil using geographically weighted regression. Sci. Total Environ. 752, 141977. https://doi.org/10.1016/j.scitotenv.2020.141977.

Yi, C., Huang, C., Pan, Y., 2007. Flood Disaster Risk Analysis for Songhua River Basin Based on Theory of Information Diffusion, in: Shi, Y., Van Albada, G.D., Dongarra, J., Sloot, P.M.A. (Eds.), Computational Science – ICCS 2007, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1069–1076. https://doi.org/10.1007/978-3-540-72588-6_171.

Zeydina, O., Beauzamy, B., 2013. Probabilistic information transfer, Les mathématiques du réel. Société de calcul mathématique, Paris.

Zhang, C., Tang, Y., Luo, L., Xu, W., 2009. Outlier identification and visualization for Pb concentrations in urban soils and its implications for identification of potential contaminated land. Environ. Pollut. 157, 3083–3090. https://doi.org/10.1016/j.envpol.2009.05.044.

Zhang, C., Tang, Y., Xu, X., Kiely, G., 2011. Towards spatial geochemical modelling: use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Appl. Geochem. 26, 1239–1248. https://doi.org/10.1016/j.apgeochem.2011.04.014.

Zhao, F., Jiao, L., Liu, H., 2013. Kernel generalized fuzzy c-means clustering with spatial information for image segmentation. Digit. Signal Process. 23, 184–199. https://doi.org/10.1016/j.dsp.2012.09.016.

Zhou, C., Wan, Q., Huang, S., Chen, D., 2000. A GIS-based approach to flood risk zonation. Acta Geograph. Sin. 55, 15.