# D1.2 Hot spot identification

## Hot spot identification within sparse, imprecise and clustered data

# Document information

| | |
|---|---|
| **Grant Agreement Number** | n° 101112889 |
| **Project Title** | Information-based Strategies for LAND Remediation |
| **Project Acronym** | ISLANDR |
| **Project Coordinator** | GTK/ Juha Kaija |
| **Project Duration** | 1 May 2023 – 30 April 2026 (36 months) |
| **Related Work Package** | WP1 Overview of soil pollution in Europe |
| **Related tasks** | Task 1.3 Interpolation of uncertain data |
| **Lead Organisation** | BRGM |
| **Contributing Partner(s)** | S.Belbeze - Bureau de Recherches Geologiques et minières (BRGM)<br>T.Tarvainen - Geological Survey of Finland (GTK) |
| **Authors** | Stephane Belbeze, Timo Tarvainen |
| **Submission Date** | 29.04.2025 |
| **Dissemination Level** | PU - Public |







Funded by the European Union

# History

| Date | Submitted by | Reviewed by | Version (Notes) |
|---|---|---|---|
| 09.22.2024 | S. Belbeze | T. Tarvainen<br>J. Kaija | Draft |
| 10.10.2024 | S. Belbeze | GTK | Reviewed version |
| 04.01.2025 | S. Belbeze | GTK | Submitted |
| 27.02.2025 | S. Belbeze | MC, T. Tarvainen,<br>J. Kaija | Final draft new template |





Funded by
the European Union

# Summary

The Information-based Strategies for Land Remediation, in short ISLANDR, is a multidisciplinary project, which is foremost aimed at supporting the execution of the EU mission: A Soil Deal for Europe. One of the research objectives in ISLANDR WP1 is to provide methods for the delineation of polluted soils across Europe. The project's input data will be all available geochemical data, including a combination of background measurements, mining surveys, and monitoring of urban and polluted sites. Compared with previous large-scale mapping projects, where sampling focused on a specific soil typology (FOREGS, GEMAS) or a range of soil typologies (LUCAS), ISLANDR's datasets will be unique and more complex, due to the anomaly-background combination of sampled populations. Techniques for anomaly detection in a dataset vary according to the type of anomalous value present: anomalous values on a box plot (range outliers), anomalous values on biplot or multidimensional projection (relationship outlier), or anomalous spatial values of a given extension (spatial outliers). For ISLANDR, which will map the risks associated with diffuse contamination, the spatial approach must be preferred, without neglecting the potential contributions of other methods. To this end, several algorithms have been tested and tuned so anomaly is chosen the same way as experts do. In doing so, the synergy between data interpolation and the detection of spatial outliers was studied, and this research led us to propose an innovative anomaly detection algorithm especially adapted to cases where data is sparse (< 30), clustered and uncertain. Unlike detections made on kriged maps and other deep learner algorithms, our algorithm has a higher detection rate because it has his own interpolator without smoothing behavior. Our algorithm can work with small sets as less than 10 data and is now ready to process as intended any ISLANDR ITA or Large Scale European survey as LUCAS.

## Keywords

European geochemistry; anomaly detection; sparse, clustered; spatial interpolation; uncertainty; soil contamination





Funded by the European Union

# Abbreviations and acronyms

| Acronym | Description |
|---|---|
| A | Anomaly |
| ANN | Artificial neural network interpolation technique |
| ANR | French national research agency |
| AT | Averaging Time is used in health and safety risk assessment |
| BA | Bioaccessibility is used in health and safety risk assessment |
| BASIAS | Potentially polluted site for French authorities |
| BASOL | Known polluted or remediated site for French authorities |
| BRGM | French Geologic Survey |
| BW | Body Weight is used in health and safety risk assessment |
| C-A | Model of concentration variation by area |
| CAX | Learning set for the EEPH algorithm (C are observations, AX is the knowledge base of covariable of arbitrary dimension) |
| CDF | Cumulative distribution function |
| CLORPT | CLimate, Organisms, Relief, Parent material, and Time |
| CODA | COmpositional Data Analysis |
| CoOK | Co ordinari kriging is a multivariate geostatistical technique |
| cov.PI | Coverage of the 90% confidence interval is used in interpolation inter comparison test |
| CS | Concentration in Soil is used in health and safety risk assessment |
| D | A Dose as used in risk assessment |
| DUK | Detrended universal kriging is a geostatistical technique |
| ED | Exposure Duration is used in health and safety risk assessment |
| EDC | Empirical diffusion coefficient |
| EEPH | Enhanced Experimental Probabilistic Hypersurface as created for ISLANDR |
| EF | Exposure Frequency used in health and safety risk assessment |
| ELN | Elastic net regression is a non geostatistical interpolator technique |
| EMEP | Co-operative programme for monitoring and evaluation of the long-range transmission of air pollutants in Europe |
| EML | Ensemble of machine learning models is an hybrid interpolator |
| EPH | Experimental Probabilistic Hypersurface |
| FMCDM | Fuzzy multi-criteria decision-making is a kind of classifier where uncertainty, vagueness, and/or imprecision are present in the decision matrix |





Funded by the European Union

| Acronym | Description |
|---|---|
| FOREGS | The Forum of the European Geological Surveys Directors (FOREGS) was an informal group that provided the directors of geological surveys with a platform for exchanging ideas on the status of each national institute. FOREGS ceased its activities in September 2005, but was taken over by EuroGeoSurveys. FOREGS has carried out a Europe-wide Geochemical Baseline Mapping Programme. |
| FPGN | National French top soil geochemical Background study |
| GEMAS | GEochemical MApping of Agricultural Soil. The GEMAS project collected 2108 Ap horizon soil samples from regularly ploughed fields in 33 European countries, |
| GP | Gaussian processes are a generic non geostatistical interpolator technique |
| GPR | Gaussian process regression is a non geostatistical interpolator technique |
| GRNN | General Regression neural-networks is a non-geostatistical interpolator technique |
| GSLIB | Geostatistical Software Library from Deutsch and Journel (1997) |
| GWR | Geographically weighted regression |
| HBKG | High background of geogenic origin used in fractal study |
| ITAs | ISLANDR Test Areas. |
| K | Simple Kriging is a geostatistical interpolator technique |
| KDE | Kernel density estimation is a non geostatistical interpolator technique |
| KED | Kriging with external drift is a geostatistical interpolator technique |
| LBKG | Low geogenic Background |
| LEPS | Linear error in probability space is a statistical measure |
| LOQ | Limit of Quantification of a measurement or a laboratory analysis |
| LANU | LAnd Use |
| LUCAS | Land use and land cover survey. The data collected by LUCAS provides harmonised and comparable statistics on land use and land cover across the whole of the EU's territory. |
| MAE | Mean absolute error used in interpolation inter comparison test |
| MAF | Maximum autocorrelation function is a multivariate and spatial dimension reduction technique |
| MAXE | Maximum error used in interpolation inter comparison test |
| MBS | Multilevel B-Spline is a non geostatistical technique |
| MBSDE | Multilevel B-splines with external drift is a hybrid interpolator technique |





Funded by the European Union

| | |
|---|---|
| Mcov.PI | Mean absolute deviation of the accuracy plot is used in interpolation inter comparison test |
| MCRPS | Mean continuous ranked probability score is used in interpolation inter comparison test |
| MDS | Multidimentional scaling is a method of dimension reduction |
| ME | Mean error is used in interpolation inter comparison test |
| MIDW | Multifractal Inverse Distance Weighting interpolation is a non geostatistical interpolator technique |
| MISE | Mean integrated squared error (MISE) is used in interpolation inter comparison test |
| MLP | Multi layer preceptron a kind of artificial intelligence and is a non geostatistical interpolator technique |
| MSE | Mean squared error used in interpolation inter comparison test |
| MULTI-MOORA | The MULTIMOORA is a ranking obtained by aggregating the results of the ternary ranking methods Ratio System, Reference Point Approach, and Full Multiplicative Form. It is a technique used in Multicriteria decision-making |
| NN | The Nearest Neighbor algorithm (NN) os |
| NNW | Refers to neural networks in a general sense and are non geostatistical interpolator technique |
| OCS | Organic Carbon Survey in HOUSES Dataset |
| ODC | Optimal diffusion coefficient |
| OK | Ordinary kriging is a geostatistical interpolator technique |
| OLS | Ordinary least squares regression is a non geostatistical interpolator technique |
| PC | Principal component from a Principal Component Analysis |
| PCA | Principal Component Analysis is a statistical methode is a method for Dimensionality Reduction |
| PDS | Exceedance probability is the probability that a certain threshold will be exceeded (PDS for the french "Probabilitée de Dépassement de Deuil") |
| PN | Probability of necessity in Pearl causal inference theory |
| PNS | Probability of necessity and sufficiency in Pearl causal inference theory |
| PS | Probability of sufficiency in Pearl causal inference theory |
| PSO | Particle swarm optimization |
| Q90, Q98 | 90 percentile, 98 percentile |
| QRF | Quantile Regression Forest is a non geostatistical interpolator technique |
| R | R is a scientific and statistical programming language |
| RBF | Radial basis function is a non geostatistical interpolator technique |
| RMSE | Root mean squared error is used in interpolation inter comparison test |





Funded by the European Union

| | |
|---|---|
| RTOP | Interpolation of Data with Variable Spatial Support is a geostatistical interpolator technique |
| SCORPAN | Soil, Climate, Organisms, Relief, Parent material, Age, N is for space, spatial or geographic position |
| SFCM | Spatial fuzzy C-means a clustering algorithm |
| SI | Soil Ingestion (kg/d) used in health and safety risk assessment |
| SIC | Sparse Imprecise Clustered datas |
| SIC2004 | The Spatial InterComparison test for emergency mapping of radiological incidents |
| SSP | Polluted Sites (From the french "Sites et Sols Pollués") |
| SVM | Support vector machine is a non geostatistical interpolator technique |
| TFN | Triangular fuzzy numbers are used in Fuzzy logics mathematics |
| TGK | Trans Gaussian Kriging is a geostatistical interpolator technique |
| TIF | Tukey Inner Fence is statistical term |
| TIN-PPV | Nearest neighbor algorithm and triangulation interpolator is a non geostatistical interpolator |
| TOPSIS | A Technique for Order Preference by Similarity to Ideal Solution used in Multicriteria decision-making |
| TPH | Total Petroleum Hydrocarbon |
| T-SNE | T-distributed stochastic neighbor embedding is a method of dimension reduction |
| UER | Unit Excess Risk are used in risk assessments |
| UMAP | Uniform Manifold Approximation and Projection is a method for Dimensionality Reduction |
| VIKOR | VIKOR is a technique for Multicriteria Optimization and Compromise Solution used in Multicriteria decision-making. The name VIKOR is Serbian “VIseKriterijumska Optimizacija I Kompromisno Resenje” |
| w.PI | The width of the 90% confidence interval is used in interpolation inter comparison test |
| WASS | The WASSerstein distance is a statistical distance measure between observations |
| XGBTREES | A modified random forest algorithm is a non geostatistical interpolator technique |





Funded by the European Union

# Table of Contents







Funded by the European Union

# List of Figures







Funded by the European Union

# List of Tables







Funded by the European Union

# Introduction

The Information-based Strategies for Land Remediation, in short ISLANDR, is a multidisciplinary project, which is foremost aimed at supporting the execution of the EU mission: A Soil Deal for Europe. One of the research objectives in ISLANDR WP1 is to provide methods for the delineation of polluted soils across Europe. The project's input data will be all available geochemical data, including a combination of background measurements, mining surveys, and monitoring of urban and polluted sites. Compared with previous large-scale mapping projects, where sampling focused on a specific soil typology (FOREGS, GEMAS) or a range of soil typologies (LUCAS), ISLANDR's datasets will be unique and more complex, due to the anomaly-background combination of sampled populations. Techniques for anomaly detection in a dataset vary according to the type of anomalous value present: anomalous values on box plot (range outliers), anomalous values on biplot or multidimensional projection (relationship outlier), or anomalous spatial values of a given extension (spatial outliers). An epistemic presupposition of all these methods is the general shape of the population to be cleaned of anomalous values: symmetrical or normal, linear relationships linking elements together, and the spatial extension of local contamination. For ISLANDR, which will map the risks associated with diffuse contamination, the spatial approach must be preferred, without neglecting the potential contributions of other methods. To this end, several algorithms will be tested and tuned so anomaly will be chosen the same way as experts do.

This report presents the ISLANDR anomaly detection algorithm and its associated developments.





Funded by
the European Union

## ISLANDR project in brief

The Information-based Strategies for Land Remediation, in short ISLANDR, is a multidisciplinary project, which is foremost aimed at supporting the execution of the EU mission: A Soil Deal for Europe.

More specifically, the ISLANDR research activities are designed to provide tools and methods so as to support: (1) the delineation of polluted soils across Europe, (2) an evidence-based assessment of the risks posed by polluted soils, (3) the promotion of sustainable and risk-based land management practices, (4) the inclusion of a wider valuation approach in financial and investment cases, and (5) a closer integration of land contamination and spatial planning decision-making. Lessons learnt and experience gained throughout the project duration will be used to (6) deliver key policy-relevant findings related to the Soil Strategy, the proposed Soil Health Law, and other areas of policy where soils are crucial.

In order to road-test the project's findings, seven test areas across Europe have been identified. To begin with, the ISLANDR Test Areas (ITAs) will provide a real-world context for the planned research activities. More concretely, the ITAs have been selected to cover different land use types, such as urban, peri-urban, rural, agro-forestry, mining, wetlands and coastal areas. Furthermore, the ITAs are characterized by both point source and diffuse pollution, as well as by different soil pollution types, such as organic, inorganic, as well as contaminants of emerging concern.

Furthermore, ISLANDR brings a dedicated focus to low input remediation, by including test areas impacted by the consequences of the green transition, such as former mining areas. This will ensure that soil remediation will be facilitated even when the cost of remediation is economically marginal or may even be negative. On the one hand, this necessitates a more thorough understanding of low input remediation approaches from a technological perspective, yet it also requires a wider value proposition for investment cases and financial planning.

Key actors, stakeholders and end-beneficiaries are at the epicentre of ISLANDR. Through roundtables in the respective ITAs, the foremost assignment of local actors will be to provide feedback and offer insights as to the robustness and effectiveness of the strategies, frameworks and decision-support tools, as well as on the wider valuation approaches and financing mechanisms to be developed over the course of the project's lifetime. Thus, the Roundtables are foreseen to bring an iterative feedback loop to the research process, with a view to ensure the wider uptake of the project's outcomes and achievements.

Last but not least, local communities in the respective ITAs will be invited to participate in a survey organized both during the early stages and towards the end of the project, as a means to document soil literacy among society thereby bringing insight as to whether the exposure of society to the project's activities on the ground can bring about a strongly desired 'awareness pull' to the benefits to be reaped from healthy soils, thereby leveraging society at large to subscribe to the projects' motto: ISLANDR for Soil Health!





Funded by the European Union

# 1. Datasets used for testing the algorithms

## 1.1. GEMAS dataset

The GEMAS dataset is a harmonized geochemical data of agricultural soil throughout Europe gathered by the Association of the Geological Surveys of Europe (EuroGeoSurveys) in cooperation with Eurometaux in 2008. The average sampling density was 1 sample from 50 x 50 km2 gird cell and the sampling depths was 0 – 20 cm. Arsenic (As) concentrations in European agricultural topsoils (Ap) have been analysed by using aqua regia extraction of the <2 mm size fraction. For more information, see: *Reimann* et al. (2014).

## 1.2. Toulouse (ISLANDR ITA 3 area)

Toulouse Métropole is one of 20 French metropolises, an intercommunal structure, centred on the city of Toulouse. Located in the Haute-Garonne department, in the Occitanie region, southern France. Toulouse Métropole has been chosen by the French Ministry of the Economy, Finance and Industry as a pilot agglomeration to demonstrate the operational feasibility of the recommendations made by the "Urban Geochemical Background Values" working group, and to provide support for the methodology for excavated soil reuse. Analyses of surface soil samples available for sensitive facilities (40 samples) were supplemented by those obtained during site diagnostics (1442 samples) commissioned by the Metropole, and during two sampling campaigns carried out by BRGM. These campaigns resulted in the collection of 140 high quality surface soil samples and 100 deep soil samples taken every meter in 20 boreholes 5 m deep. 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. For more information, see: *Belbeze* et al. (2019).

## 1.3. Anomaly detection validation dataset

Following the discussions in WP1, algorithms were developed and tested on the European Community's large-scale GEMAS survey conducted in 2005 (Reimann et al., 2014a,b). The GEMAS Ap horizon measurements for arsenic are particularly noteworthy (Figure 1). Tarvainen et al. (2013) identified 52 anomalies (Table 1) that could be considered true ground truth for our algorithms. Most of the anomalies identified came from geogenic sources.





Funded by the European Union

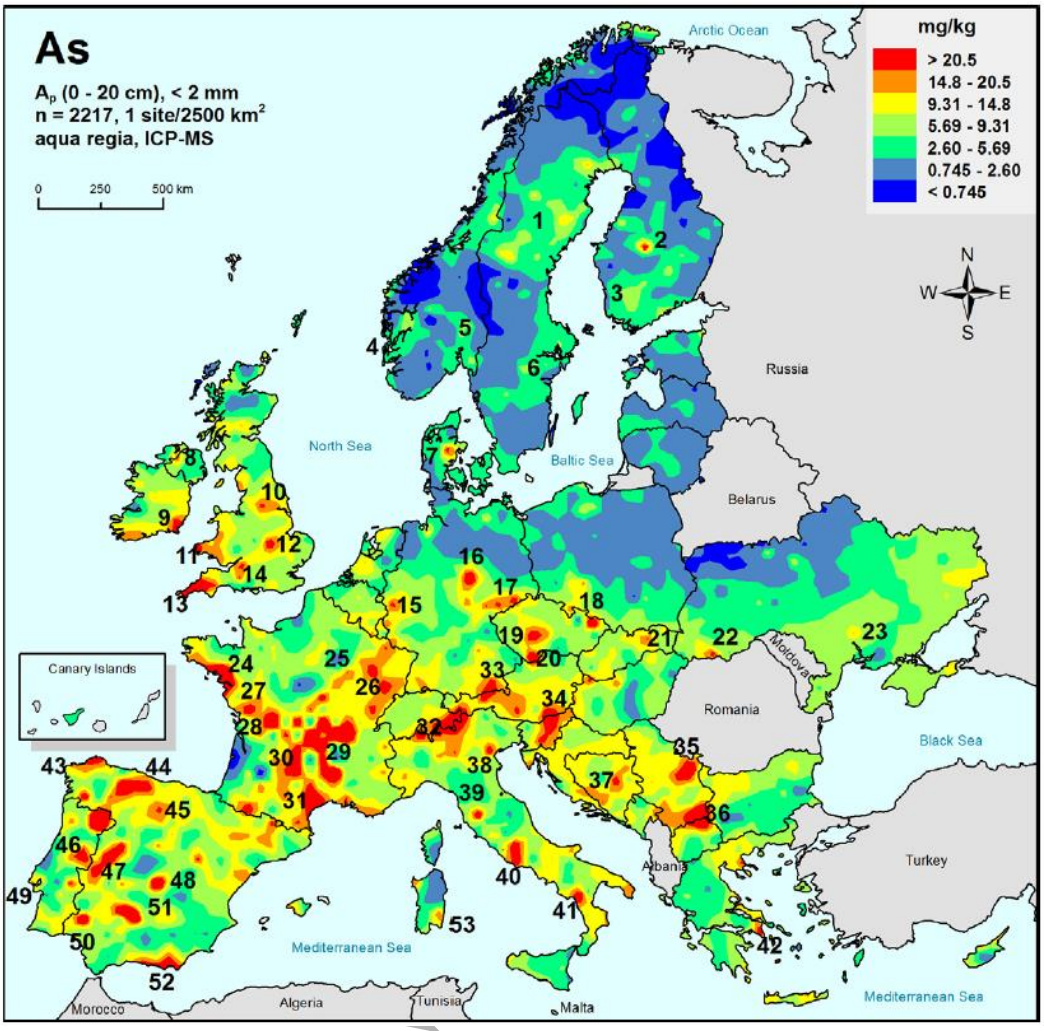


**Figure 1: Distribution of As in European agricultural topsoils. Aqua regia extraction of the <2 mm size fraction from Tarvainen et al. (2013).**





Funded by the European Union

**Table 1: Explanation of arsenic anomalies as in Tarvainen et al. (2013). Numbers refer to the map in Fig. 1.**

| No. in Ap map | Country | Name | Remarks |
|---|---|---|---|
| 1 | Sweden | Central Jämtland, Skellefteå-Västerbotten, Boden | Mineralization; may be partially enhanced by mining |
| 2 | Finland | Saarijärvi | Unknown source; clay-rich soil |
| 3 | Finland | Tampere region | Geology; major geogenic arsenic province in metavolcanic/metsedimentary area |
| 4 | Norway | | Geology; arsenopyrite in glacial sediments |
| 5 | Norway | Oslo Rift, Kongsberg | Geology/mineralization; black shale as additional As source |
| 6 | Sweden | Zinkgruvan in Southern Bergslagen | Mineralization/geology; some of the high values may also be due to marine clays |
| 7 | Denmark | | Unknown source |
| 8 | UK | N Ireland | As-rich strata in West Tyrone, but geological source still not fully established |
| 9 | Ireland | SE Ireland | Geology/mineralization, elevated As levels in Ordovician metavolcanic formations, which host VMS (volcanogenic massive sulfide) and gold deposits locally |
| 10 | UK | | Mineralization/geological control |
| 11 | UK | | Mineralization/geological control and processing of As-rich metal ores in Swansea in southern Wales |
| 12 | UK | | Geology; As-rich Jurassic ironstones |
| 13 | UK | Cornwall | Mineralization/mining; related to granitic intrusions and the Sn deposits in Cornwall |
| 14 | UK | | Processing of As-rich metal ores are evident near the Bristol Channel (Avonmouth smelter) |
| 15 | Germany | Aachen–Stollberg | Mineralization/mining; related to Pb–Zn vein-type deposits of the Variscan Rhenish Massif (Maubach, for example) |
| 16 | Germany | Harz Mountains | Mineralization/mining; related to Pb–Zn–Cu deposits in the Variscan Harz Mountains (Rammelsberg, Bad Grund) |
| 17 | Germany + Czech Rep. | Erzgebirge (ore mountains) | Mineralization/mining; mineralized area (Pb–Zn–Ag–U) of Joachimsthal-Annaberg-Freiberg etc. in the Variscan Erzgebirge region, anthropogenic overprint documented; brown coal combustion, vein mineralization in the Czech Republic |





Funded by the European Union

| | | | |
|---|---|---|---|
| 18 | Poland/Czech Republic | Ostrava | Predominantly anthropogenic (?); mining and combustion of coal, metallurgical industry |
| 19 | Czech Republic | Kutna Hora | Mineralization/mining; Ag–Pb–Zn mining district Kutna Hora |
| 20 | Czech Republic | Bohemian massif | Geology; occurrence of high-potassic plutonites = DURBACHITES |
| 21 | Slovakia | | Mineralization/mining |
| 22 | Ukraine | Transcarpathia | Mineralization/mining |
| 23 | Ukraine | Near Mykolajiv | Unknown source |
| 24 | France | | Mineralization/geology; Armorican shear zone with As, Sb, Au mineralizations |
| 25 | France | NE axis SE Paris basin | 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 |
| 26 | France | Southern Vosges | 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 |
| 27 | France | | Mineralization/geology; As, Co, U vein type mineralization and main shear zone (SW Armorican) |
| 28 | France | | Geology; Hercynian granite in Jurassic black marls |
| 29 | France | | Geology/mineralization; Argentat deep fault, perigranitic thermal aureoles and epithermal mineralization in the Auvergne quaternary volcanics |
| 30 | France | | Mineralization/mining; As anomalies related to the La Baume (Pb–Zn) and Carmaux (Coal) abandoned mines |
| 31 | France | | 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 |
| 32 | Switzerland/Italy/Austria | Austria: Silvretta and Oetztal Alps | Mineralization; in Austria, stratiform mineralizations within the crystalline units of the Campo–Sesvenna–Silvretta nappes and the Oetztal nappes; in Switzerland, natural mineralizations along complex tectonic features |
| 33 | Germany/Austria | Austria: Western Greywacke Zone (Tyrol) | Mineralization/mining; fahlore deposits in Schwaz–Brixlegg |
| 34 | Austria/Slovenia | Austria: Styrian basin | Geology; there is also a higher As-level in the soil survey of Styria in that area |





Funded by the European Union

| | | | |
|---|---|---|---|
| 35 | Serbia | Bor-Majdanpek area | Mineralization/mining; mining district, pollution related to tailings |
| 36 | Serbia/FYROM/Bulgaria | Transboundary zone | Mineralization/mining; transboundary metallogenic province, including, for example, Lece–Halkidiki zone and Besna Kobila–Osogovo zone |
| 37 | Bosnia and Herzegovina | | |
| 38 | Italy | Po river plain | Enrichment due to adsorption onto Fe-hydroxides in fine-grain-sized fraction of soils in the alluvial plain; mixed source (agriculture in the plain and arsenopyrite from metamorphic rocks) |
| 39 | Italy | Piana di Scarlino (Tuscany) | Geology; Cu–Pb–Zn and Fe mineralization and thermal water |
| 40 | Italy | Roman–Neapolitan Comagmatic Province (RNCP) | Geology; volcanic soils |
| 41 | Italy | Southern Campania (Cilento) | Geology; volcanic soils generated from wind transported pyroclastics (from the RNCP) mantling carbonatic rocks |
| 42 | Greece | Vavrona (Attiki Prefecture) | Mineralization/mining; related to polymetallic sulphide mineralizations; minor anthropogenic contribution by insecticides or CCA |
| 43 | Spain | Galicia | Mineralization; two anomalies that are related to As disseminations close to mineralization zones in a shear zone |
| 44 | Spain | Asturias Arc Domain | Mineralizations |
| 45 | Spain | Castilla–León Tertiary Basin | Geology; most probably related to elevated As content in some marls and limestones |
| 46 | Portugal | | Mineralization/mining |
| 47 | Spain | | Geology; several small anomalies related to leucocratic granitic facies, late magmatic alternation, W–Sn–As mineralizations |
| 48 | Spain | Toledo Mountains | Mineralization |
| 49 | Portugal | Lisbon | Anthropogenic |
| 50 | Spain | Iberian Pyrite Belt | Geology/mineralization; mineralizations in the Iberian Pyrite Belt or Pb–Zn–Cu district of Sierra de Aracena (limestones, metavulcanites) |
| 51 | Spain | Pedroches batholit | Mineralization/mining |
| 52 | Spain | Sierra de Gador | Mineralization/mining |
| 53 | Italy/Sardinia | Southern Sardinia | Mineralization/mining; arsenopyrite veins |





Funded by the European Union

This anomaly detection work was subsequently taken up by Reiman (2018), who tested various thresholds for northern and southern Europe (range outliers). Table 2 shows the Reimann et al., (2018) statistics and Figure 2 provides their map.

**Table 2: As statistics on GEMAS Ap sample using threshold by Reiman et al., (2018)**

| | Q95<br>95th percentiles | Q98<br>98th percentiles | TIF<br>Upper whisker of the box plot |
|---|---|---|---|
| North | 7.31 | 10.3 | 16.6 |
| South | 26.8 | 45.7 | 38.2 |

*Note: TIF is the highest value that inner whisker of the box plot can reach, also called Tukey Inner Fence* (Tukey, 1977)

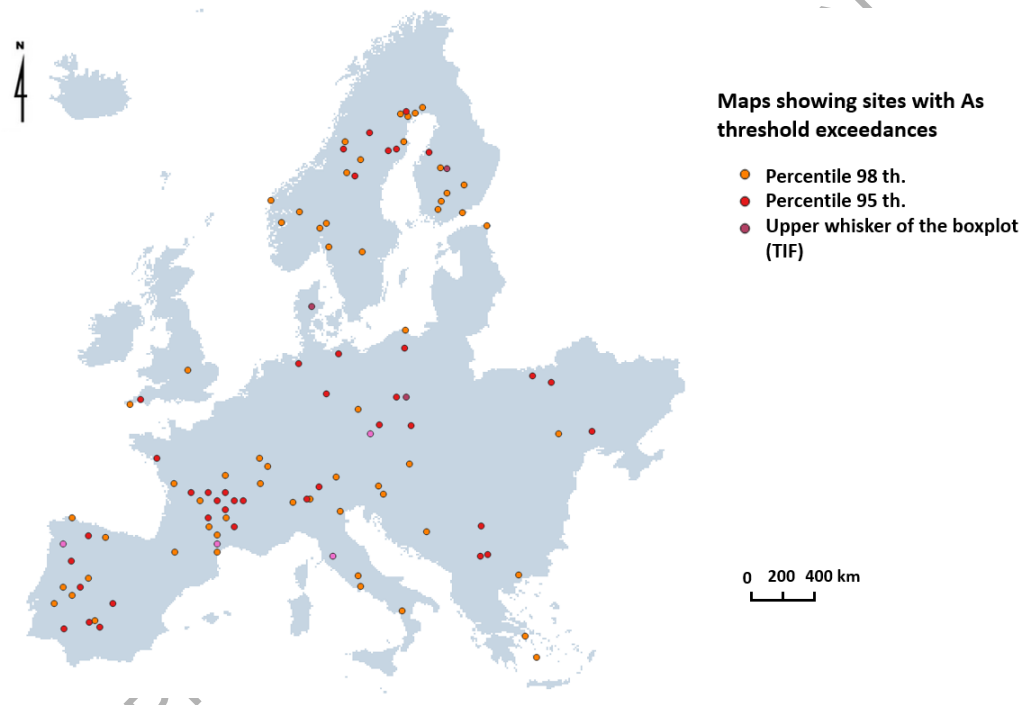


**Figure 2: Anomalies in GEMAS As Ap samples as pinpointed by Reimann et al., (2018)**

## 1.4. Arsenic Baseline with GEMAS data

To test other detection systems, which are based on interpolated surfaces rather than measurement points, two As base maps have been created to better appreciate the efficiency of our interpolation anomaly detection algorithms (Figure 3 and Figure 4).

To do this, we used an advanced geostatistical technique called Trans-gaussian Kriging (Chiles and Delfiner, 2012) and developed our own interpolator especially suited for anomaly detection and called EEPH (Appendix 1: Interpolation algorithm). Among the many interpolation models available, we have opted for an information dissemination algorithm that can handle sparse, imprecise, and clustered (SIC) data that was developed



Funded by the European Union

by Beauzamy in the 2000s (Extended Probabilistic Hypersurface). This probabilistic model has been enriched (anisotropy, declustering, auto-variography, multi-support, treatment of covariates, treatment of possibilities, uncertainties, and censored data) in a way that fully meets our needs and can be used in conjunction with the risk calculations of WP2, whatever their mathematical form. In addition, this intrinsically diffusive model presents analogies with the processes of forming background concentrations in soils and their observation through changes in scale.

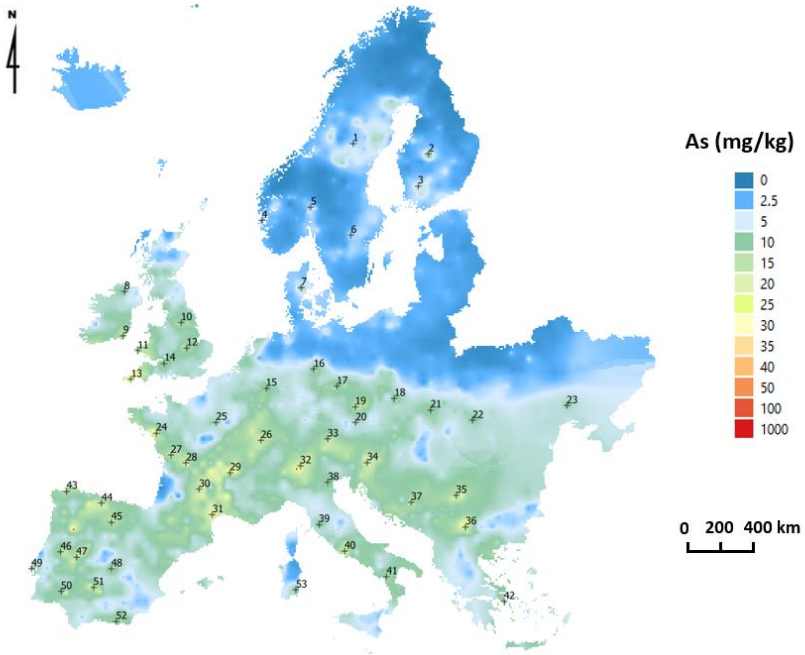


**Figure 3: Transgaussian kriging GEMAS Survey, Ap (0–20 cm), < 2mm, n=2217, 1 site/2500 km2, aqua regia, ICP-MS**

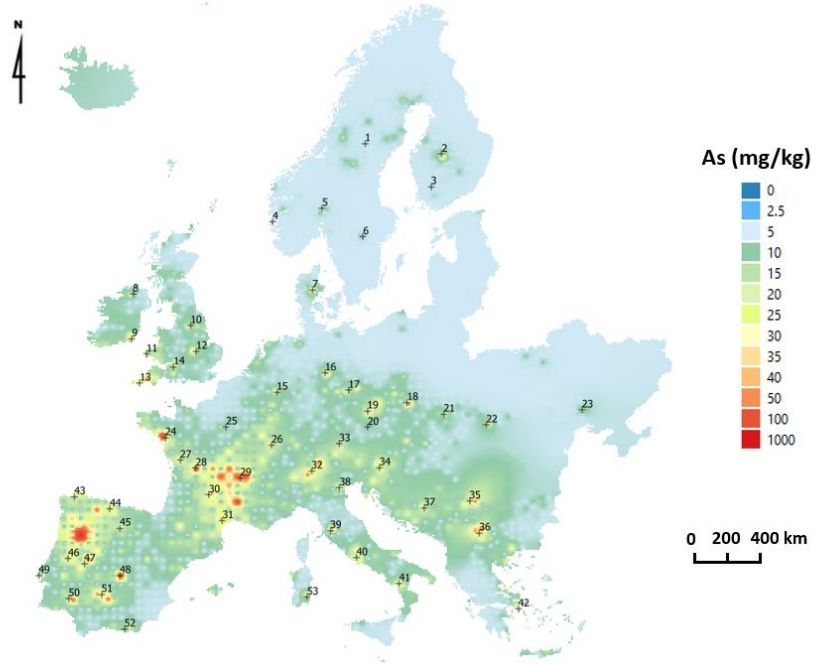


**Figure 4: Expected value EEPH with neutral diffusion coefficient for enhanced anomaly detection, GEMAS Survey, Ap (0–20 cm), < 2mm, n=2217, 1 site/2500 km2, aqua regia, ICP-MS**





Funded by the European Union

## 1.5. Proxy interpretation maps

To help interpret the anomalies identified, and especially to reflect the anthropogenic or diffuse nature of our anomalies, a variety of interpretation proxies are needed, such as land use, urban perimeter, and geology. The data used as interpretation proxies can vary in precision and should therefore be used with caution and as an indicator only. We have selected the data below (Table 3).

**Table 3: Interpretation proxy collected for ISLANDR and used as geogenic/anthropic clue**

| Proxy name | Description |
|---|---|
| Lithology | Dürr et al., 2005 publication and appendices.<br>https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2005GB002515<br>A digital map of the lithology of the continental surfaces is proposed in vector mode (n ≈ 8300, reaggregated at 0.5° × 0.5° resolution) for 15 rock types (plus water and ice) targeted to surficial Earth system analysis (chemical weathering, land erosion, carbon cycling, sediment formation, riverine fluxes, aquifer typology, coastal erosion). |
| Land use | CLC 2018 raster<br>Generated using information from the European Union's Copernicus Land Monitoring Service<br>https://doi.org/10.2909/960998c1-1870-4e82-8051-6485205ebbac<br>Provides pan-European CORINE Land Cover inventory for 44 thematic classes for the 2018 reference year. The dataset has a minimum mapping unit (MMU) of 25 hectares (ha) for areal phenomena and a minimum mapping width (MMW) of 100 m for linear phenomena and is available as vector and as 100 m raster data. |
| Urban area | "Europe." Downloaded from http://tapiquen-sig.jimdo.com. Carlos Efraín Porto Tapiquén. Orogénesis Soluciones Geográficas. Porlamar, Venezuela 2015.<br>Based on shapes from Environmental Systems Research Institute (ESRI). Free distribution.<br>Within Europe, countries, major cities |
| Moss survey | Harmens et al. (2010). Heavy metals in moss survey results, 2010 on EMEP grid.<br>https://icpvegetation.ceh.ac.uk/sites/default/files/Heavy%20metals%20and%20nitrogen%20in%20mosses%20-%20spatial%20patterns%20in%202010-2011.pdf<br>Heavy metal 2010 sampling sites, nitrogen 2010 sampling sites, aluminum, arsenic, cadmium, chromium, copper, iron, mercury, nitrogen, nickel, lead, antimony, vanadium, and zinc. Concentration of heavy metals in moss samples can be used to estimate the atmospheric deposition of metals. |
| Dust survey | Targa et al. (2023). ETC HE Report 2023/1: European air quality status report for 2021, using validated data. https://www.eea.europa.eu/publications/europes-air-quality-status-2023<br>Online PM10, PM2.5, ozone, NO2, BaP results, one µg/m$^3$. |
| Mineral knowledge | Simplified, user-friendly, and efficient access to all available and new data related to mineral resources through the Minerals4EU Knowledge Data Platform<br>http://minerals4eu.brgm-rec.fr/ |





Funded by the European Union

One avenue that ISLANDR has successfully explored has been to construct metal(oid) content proxies for mosses and dust (PM2.5 emitted mainly from the combustion of solid fuels for domestic heating, industrial activities and road transport) to better identify a potential diffuse component, drawing on the work of Harmens et al. (2010) and Frontasyeva et al. (2020).

Therefore, an elevated trace element content in mosses and GEMAS soil data is likely to be of geogenic or mining origin. An anomaly in mosses alone is indicative of anthropogenic activity. Figure 5 shows an overlay of arsenic anomalies identified in GEMAS samples with 2010 moss data from EMEP monitoring (Harmens et al., 2010). When raw data was unavailable or not publicly available, published maps had to be digitized. Existing digitalization methods under R such as HistMapR by Auffret et al. (2017) proved ineffective on this data. We then sampled the data on fishnet points and an interpolation using our EEPH. These mappings have a significant error (for example, estimated at 66% for arsenic) due to the fact that we only know the levels via a legend, but our interpolation algorithm handles this well.

European PM25 dust contents (Targa et al., 2023) were treated the same way, resulting in a set of proxies for interpreting our anomalies covering 10 metal(oid)s: As, Cd, Cr, Cu, Fe, Hg, Ni, Pb, Sb, and Zn. These are possible upper and lower values due to the imprecise legend used for the EMEP European maps (error is 66% for the mean).





Funded by
the European Union

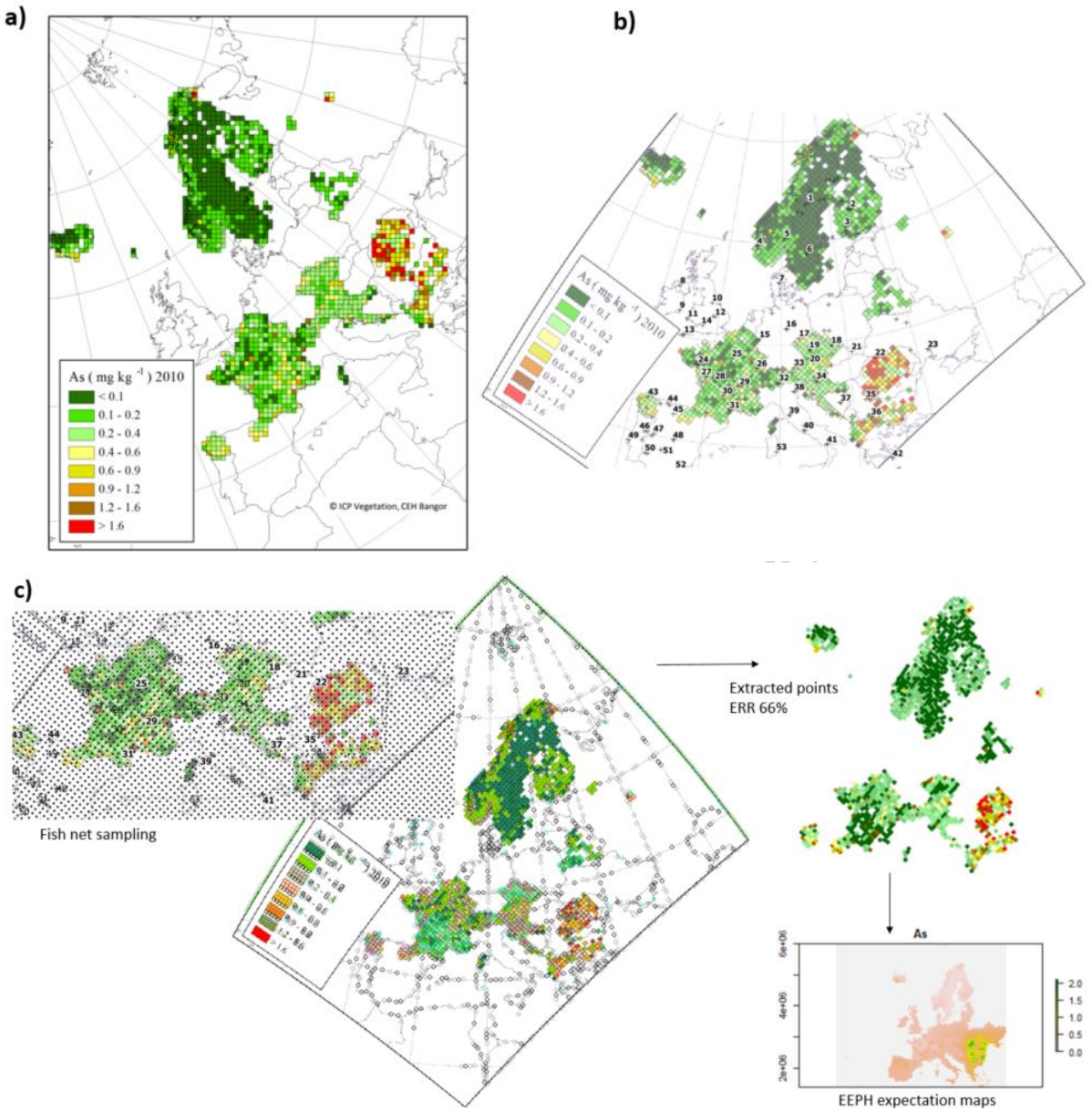


**Figure 5: Digitization of published moss maps to imprecise maps used as interpretation proxy for ISLANDR** ***With a) published EMEP 2010 As in moss (Harmens et al (2010), b) re-projected EMEP map with GEMAS numbered anomalies, and c) EEPH moss map by fishnet point capture and interpolation.***





Funded by the European Union

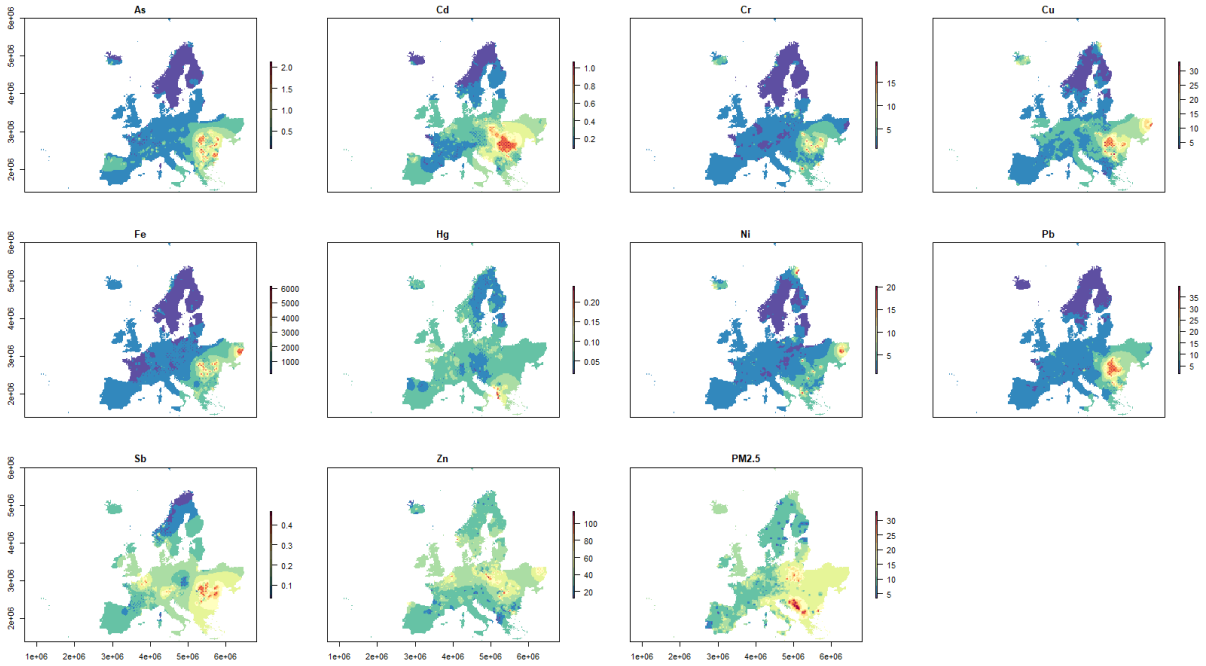


**Figure 6: Metalloids in mosses in ppm and PM2.5 dust in µg/m3 Interpretation proxies generated for ISLANDR**

Finally, the Minerals4EU database has been queried to extract all mining and mining waste sites (Figure 7 and Figure 8).

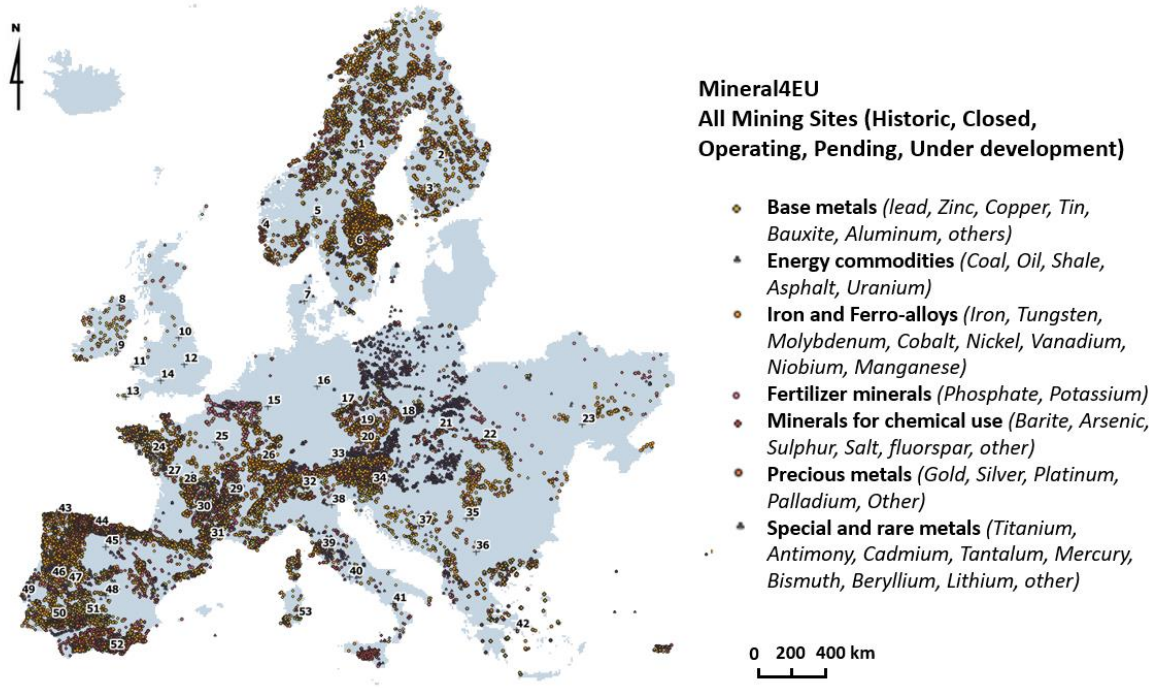


**Figure 7: Mining sites in Europe, Minerals4EU database, all continental mining sites (historic, closed, operating, pending, under development)**



Funded by the European Union

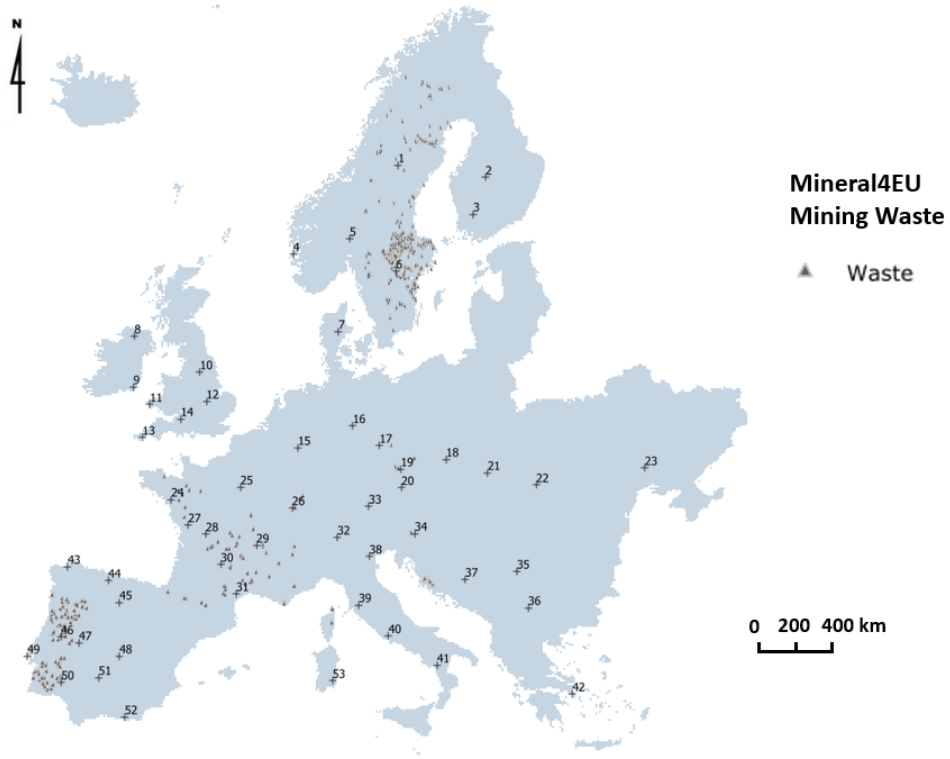


**Figure 8: Mining sites in Europe Minerals4EU continental mining wastes**

*Note: Numbered points are anomalies in As Ap sample from Tarvainen et al. (2013).*





Funded by
the European Union

# 2. Hot spot identification algorithms

## 2.1. Anomaly threshold analysis values

For GEMAS, Reimann et al. (2018) tested various values for the total dataset, then Northern and Southern Europe (range outliers). Although dependent on the degree of symmetry of the element distributions studied, Reimann deemed quantile 95 (Q95), quantile 98 (Q98), and the Tukey theoretical percentile (Tukey Inner Fence, or TIF) to be powerful tools for detecting anomalous values.

Just a reminder:

$$TIF = Q75 + 1.5\,(Q75 - Q25)$$

The 1.5 factor comes from the normal distribution. Reimann et al. (2018) specifies that this TIF should be calculated on variables whose symmetry has been restored by Log or another transformation. These quantiles, whose TIF will be systematically calculated for all ITAs, are known as range outliers.

## 2.2. Local Moran index

The local Moran index (Anselin, 1995) for determining anomalous values is calculated as follows.

$$I_i = \frac{z_i - \overline{z}}{\sigma^2} \sum_{j=1, j \neq i}^{n} \left[ w_{ij} (z_j - \overline{z}) \right]$$

With $I$ the Moran index, $z_i$ the value of the variable, $\overline{z}$ the mean of the n samples, $\sigma$ its variance, and $w_{ij}$ is a spatial weight that can associate a given neighborhood on a distance band or simply the inverse of the distance (IDW weight) between points.

Like the geostatisticians' variogram, the Moran index is highly sensitive to deviations from data normality and is usually calculated by these on values transformed by Box-Cox or normal score and a given neighborhood distance (Zang et al., 2008).

For ISLANDR, the spatial Moran index was tested on raw data and the normal score of kriging data (Figure 9 and 10) and on raw values for the EEPH (Figure 11 and Figure 12).

The Moran index can be interpreted as follows:

- A high positive index indicates spatial clusters HH (strong values in a strong neighborhood) and LL (weak value in a weak neighborhood).
- A high negative index indicates an outlier that differs markedly from its neighbors; HL is a strong value within weak values, and LH a weak value within strong values.





Funded by the European Union

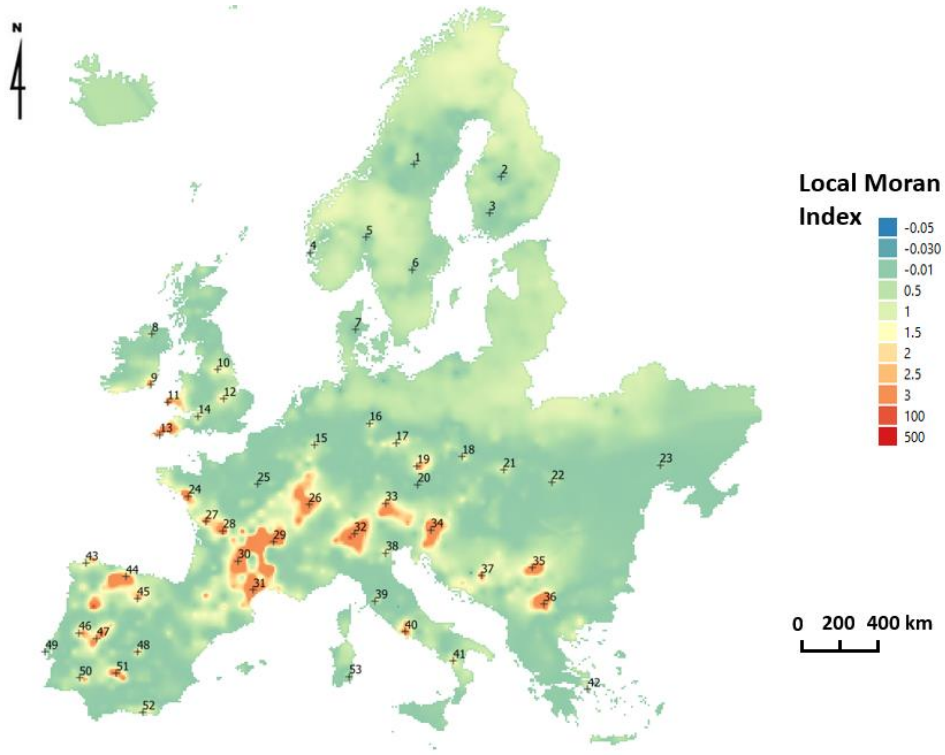


**Figure 9: Local Moran index on transguaussian kriged map, GEMAS Survey, Ap (0–20 cm), < 2mm, n=2217, 1 site/2500 km2, aqua regia, ICP-MS**

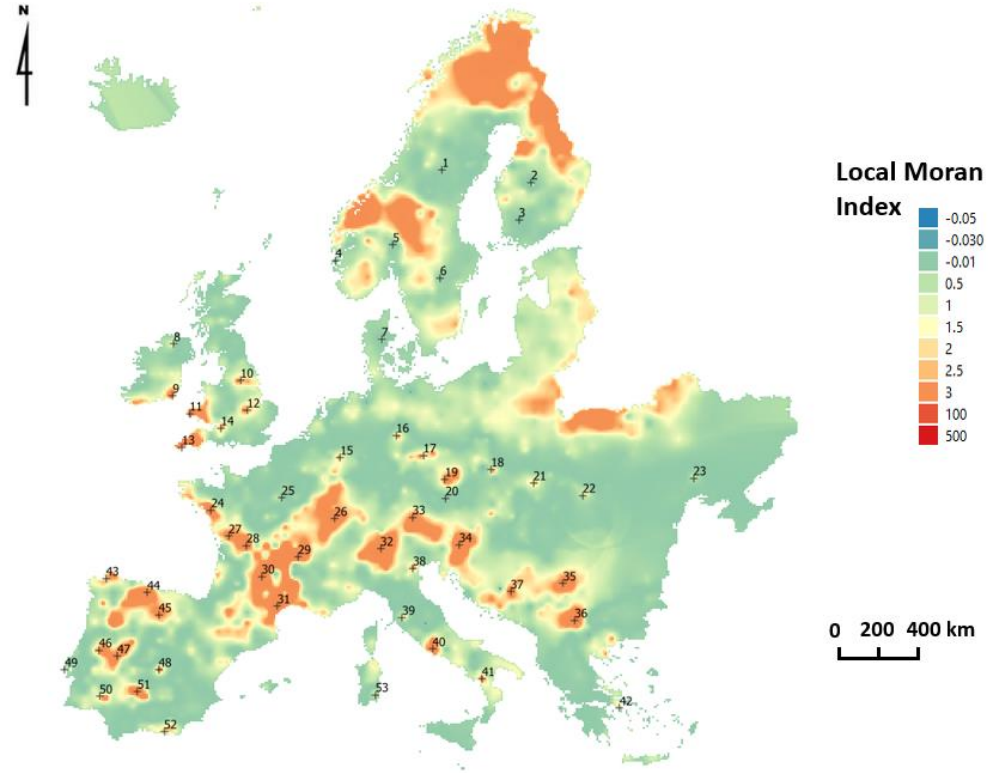


**Figure 10: Local Moran index on normal score trans-gaussian kriged map, GEMAS Survey, Ap (0–20 cm), < 2mm, n=2217, 1 site/2500 km2, aqua regia, ICP-MS**



Funded by the European Union

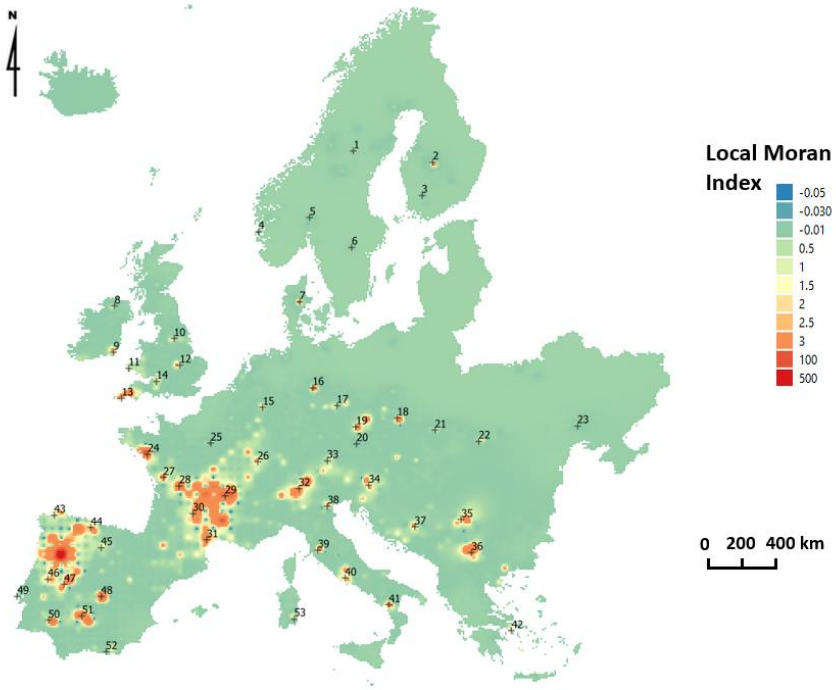


**Figure 11: Local Moran index on expected value EPH map, GEMAS Survey, Ap (0–20 cm), < 2mm, n=2217, 1 site/2500 km², aqua regia, ICP-MS**

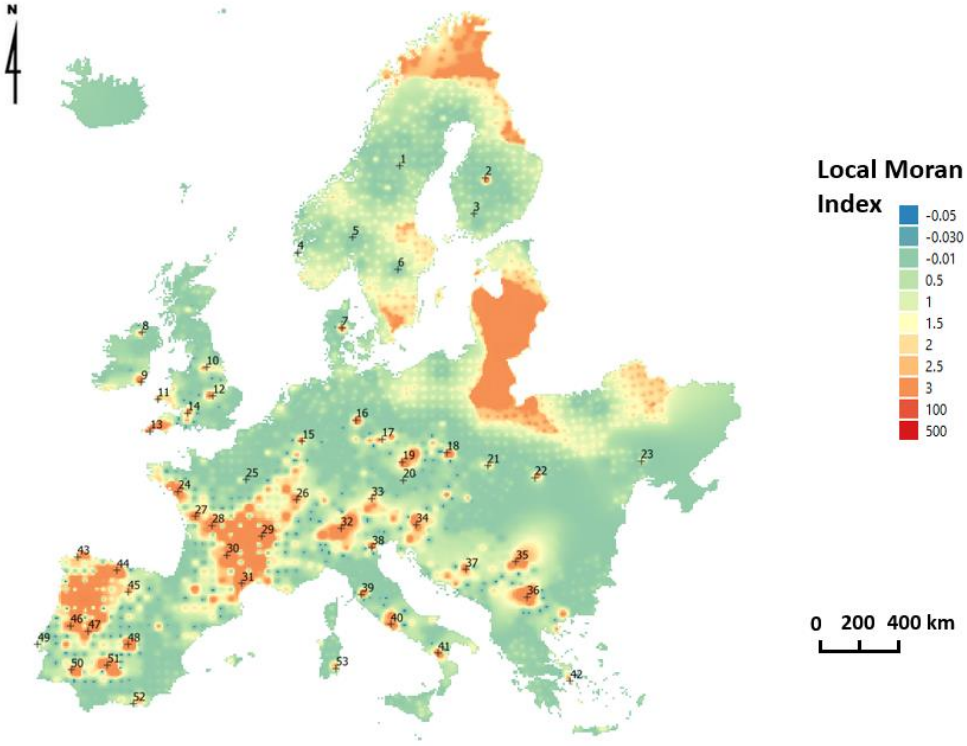


**Figure 12: Local Moran index on normal score of expected value EEPH map and using a neighborhood of 10000m, GEMAS Survey, Ap (0–20 cm), < 2mm, n=2217, 1 site/2500 km2, aqua regia, ICP-MS**



Funded by the European Union

The Moran index requires an elaborate transformation of the maps to deliver its full power of detection. Working as a variogram on the differences between neighbours, it may miss certain anomalies on the complete dataset that are "too flat" due to its variographic approach (centered on the mean). On the other hand, if we separate Northern and Southern Europe like Reimann et al. (2018) did, its detection power can be improved.

## 2.3. The fractal singularity index

Much like the discussion on urban fill thickness (Ducommun et al., 2022), microcontaminant levels observed in soil are produced by cascades of anthropogenic (aerial deposition, industrial discharge) and geochemical (dilution, diffusion, etc.) processes. Cheng (2012) considers that an area A of chronically polluted soil has a concentration of element C resulting from a dynamic process involving area A. As an initial approach, if this process is linear with a rate of change of λ, we have:

$$\frac{\partial C}{\partial A} = -\lambda \, \mathrm{C}$$

The solution to this equation is: $C = a\, e^{-\lambda \, \mathrm{A}}$ with a constant. In this case, the coefficient λ is proportional to the logarithm of the concentration and inversely proportional to the area. This type of relationship makes it possible to explain deposits linked to river flooding or decay by biological degradation following a past pollution episode.
In a second approach, the variation in concentration per area may depend on the concentration with a variable rate of change:

$$\frac{\partial C}{\partial A} = -\frac{\lambda}{A} \, \mathrm{C}$$

The solution to this equation is: $C = a\, \sqrt{A}^{\Delta\alpha}$ (a log-linear relationship) with $\Delta\alpha = 2 - \lambda$ and a coefficient that may or may not be constant. $\alpha$ is referred to as a singularity in the theory of the same name (Cheng 2007 a,b and Cheng 2012 a,b).

These equations are used in mining research and have, for example, been calibrated for the Pulacayo zinc deposit (Agteberg, 2012) with complex formation geochemistry. For this type of model:

- An area of given geogenic concentration where there is no element accumulation or dilution ($\lambda = 0$)) has a constant concentration regardless of its area A), which corresponds closely to the physical meaning of the natural geochemical background noise in a given area.
- In another area, where an anthropogenic contribution of an element is added to the geogenic concentration, the concentration will become more or less proportional to the measurement area: this is an anomaly in the natural geochemical background, meaning point-source pollution or urban geochemical background. Both can be distinguished by their $\Delta\alpha$ (singularity indices).





Funded by the European Union

For ISLANDR, the local fractal singularity index was calculated for kriging Figure 13) and EEPH (Figure 14) based on a sliding window (Xiao et al., 2016), and is interpreted as follows:

$$\alpha \begin{cases} > 2 \text{ dilution} \\ < 2 \text{ accumulation} \end{cases}$$

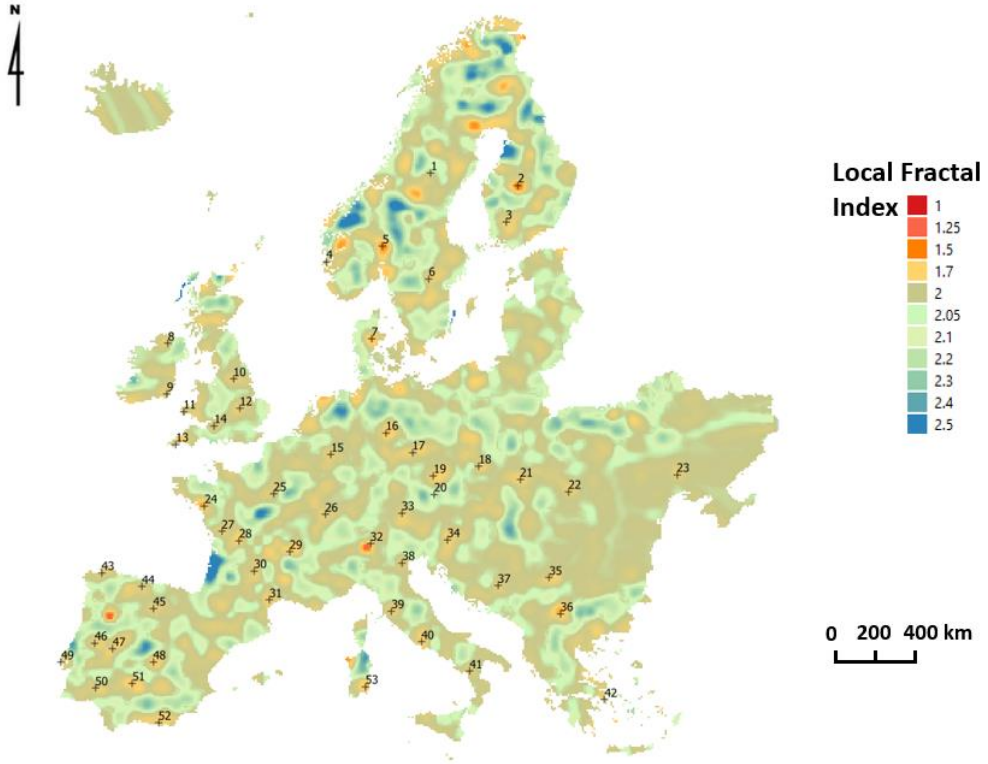


**Figure 13: Local fractal index on transgaussian kriging and using a neighborhood of 10000 m adapted from Xiao et al. (2016).**





Funded by the European Union

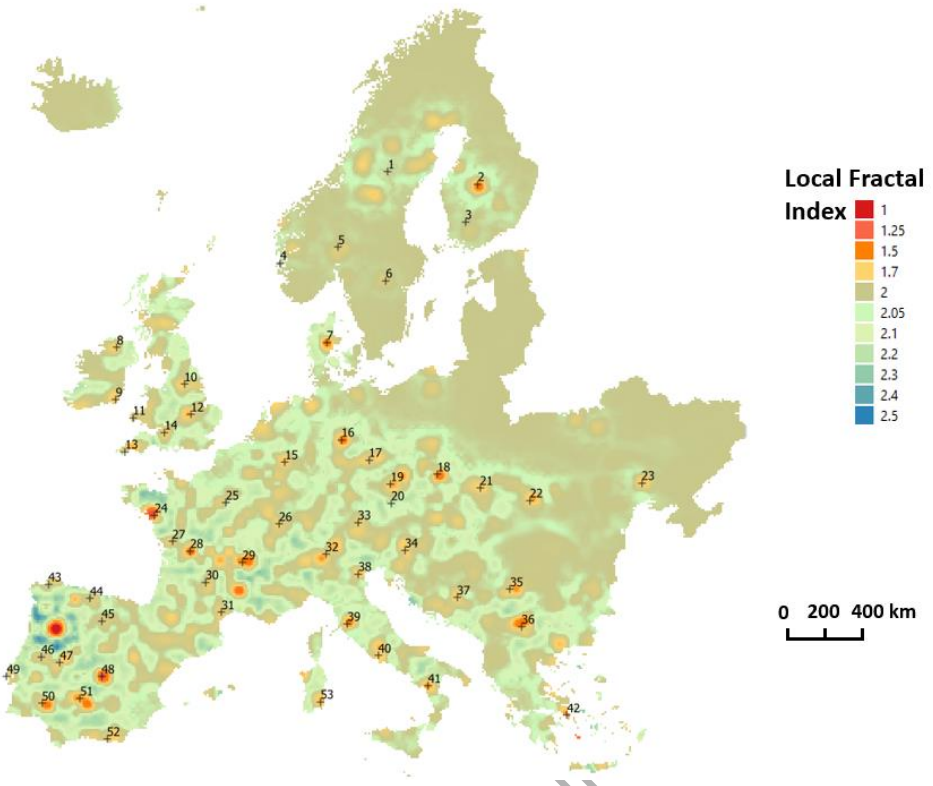


**Figure 14: Local fractal index on expected value EEPH with 10,000 m sliding window.**

Thus, the fractal singularity method provides excellent results for anomaly detection, whether using kriging or, even better, EEPH, and should therefore be used for ISLANDR, especially for EEPH.

## 2.4. C-A fractal analysis

In the context of fractal theory (Cheng, 1999 a, b) used by Carranza (2008), the C-A model is suggested, in line with the model of concentration variation by area as a function of concentration described above:

$$A[>C] \propto C^{-\beta}$$

This is a model of isotropic scale invariance. This model has been modified to work in frequency on the fourier transforms of geophysical surveys to give the S-A model (Cheng 1999 a, b),

$$A[>S] \propto S^{-2d/\beta}$$

where S is the spectral energy as a function of the fourier transform wavelength. In both cases, the exponent can be mapped to reflect the degree of process self-similarity. These maps can be used to define geographically coherent land units (LU) and the threshold values that separate them. In the European GEMAS and FOREGS projects, Italy has adopted C-A and S-A fractal methods to establish background noise: Albanese et al. (2007), Cicchella et al. (2012), Civitillo et al. (2016), and Petrik et al. (2018).





Funded by the European Union

This is a well-established technique (Caranza, 2009), which elegantly separates a geochemical background from anomalies based on the anomaly surface. The background is assumed to be in the majority in our collected data. Using EEPH as a baseline, it is possible to plot as many C-A curves as calculated quantiles, which supports the technique's conclusions (Figure 15).

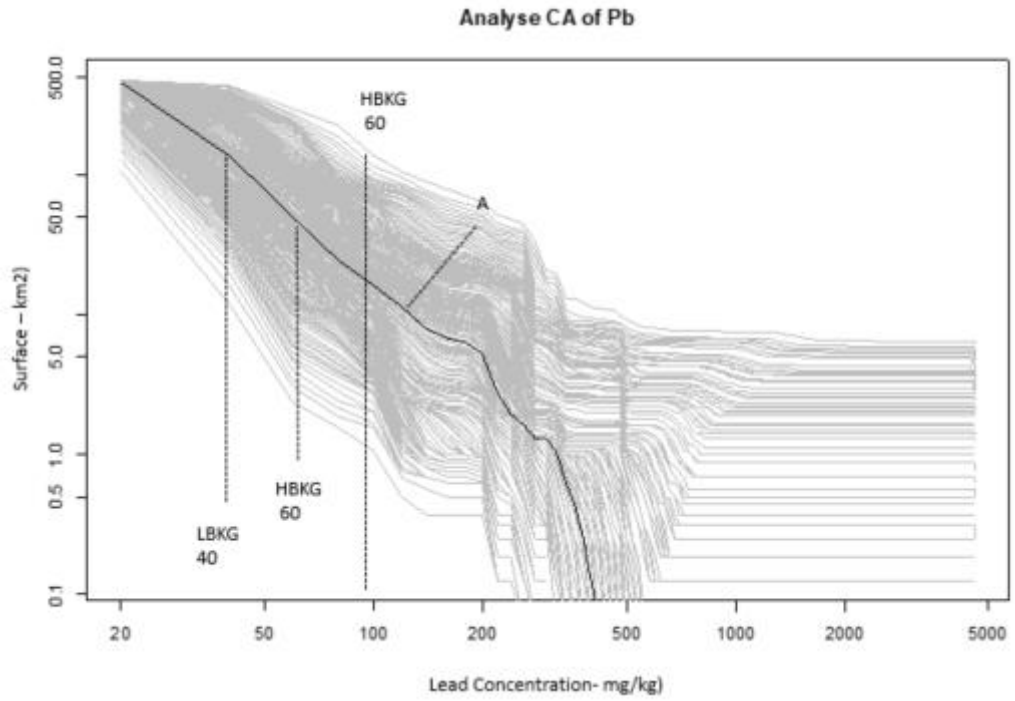


**Figure 15: Example of a C-A analysis of a low background value (LBKG), high background value (HBKG), urban background value proposition of anomaly threshold (A) of multiple quantile calculus. City of Toulouse ITA, 138 surface samples, (Belbeze et al., 2019).**

## 2.5. Zero probability bands

Belbeze et al. (2019) proposed a method for detecting anomalous values in which the multidimensional space of results (known as a mathematical manifold) is projected in 2-dimensional views. With such a display, populations form balls or clusters around which outliers gravitate, and which are also selected interactively (Figure 16). Although this method is effective, it suffers from an intractable exponential complexity as the number of data points increases. This is called the "curse of dimensionality." Consequently, it was not possible to test it in its current form on ISLANDR's thousands of datasets. However, what separated outliers from populations in this method was an empty space in the projected space. This gave rise to the idea of searching the spatial populations identified around a band with zero probability (Figure 17) or, from another angle, for a plane separating the populations in two. A support vector machine (SVM) algorithm (Vapnick, 1993) can be used for this purpose.





Funded by the European Union

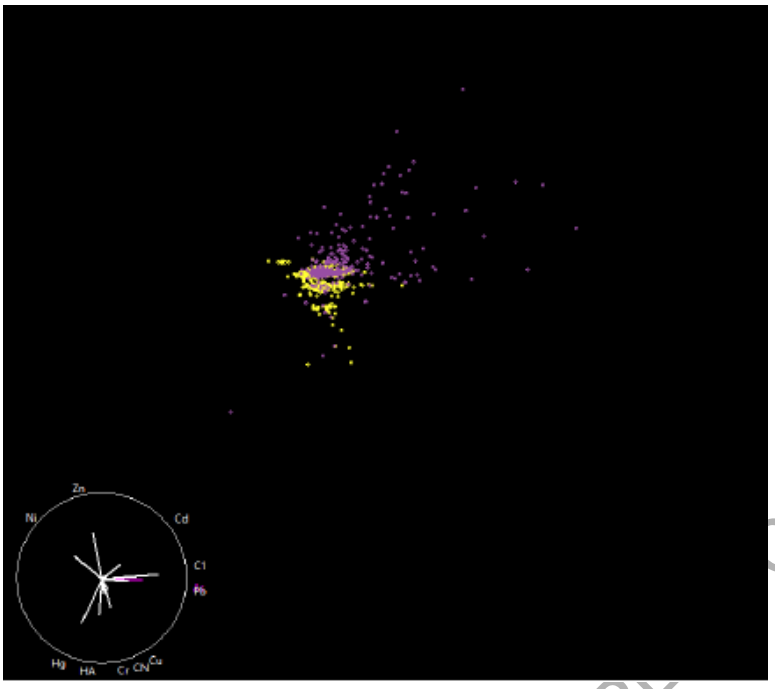


**Figure 16: Identification of multi-element anomalies (purple dots) from Belbeze et al. (2019)**

In concrete terms, for a given country or geology, we construct a histogram of the data.

Let A = $\{x_1, x_2, \dots, x_N\}$ Let A = ZZ be a set of N data in the interval [min(A), max(A)] usually denoted [a, b].

Let $\tau$ be the chosen step size, usually the data quantification limit or min(A). We are going to generate $\upsilon$ intervals of length $\tau$ or boxes in which to place our data:

We choose K the number of boxes such that integer K and Kt ≥ b
$B_1 = [0, \mathrm{t}[\ , B_1 = [\mathrm{t}\ , 2\mathrm{t}[\ , .., B_k = [(\mathrm{k}-1)\mathrm{t}\ , k\mathrm{t}[\ , \dots, B_K = [(\mathrm{K}-1)\mathrm{t}\ , K\mathrm{t}[$

Then we count how many measurements have fallen into the boxes $n_1\ pour\ B_1$, $n_K\ pour\ B_K$. We have $n_1 + \dots + n_k$ = N. We obtain a histogram as shown in Figure 17

Therefore, in $\{B_1, \dots, B_K\}$, any sequence equal to or greater than three consecutive zeros $\{n_k, n_{k+1}, n_{k+2}\} = \{0, 0, 0\}$ is called a band of zero probability. This band separates the population from outlier anomalies and sometimes outliers from each other (Figure 17).





Funded by the European Union

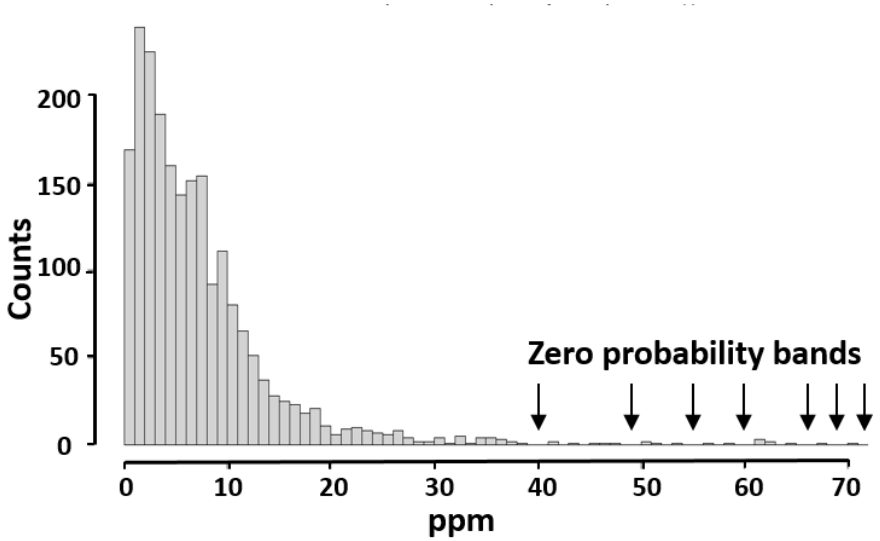


**Figure 17: Histogram of As in European agricultural topsoils, zoom on 0–76 ppm range, Aqua regia extraction of the <2 mm size fraction, chosen  is 2 ppm. Zero probability bands are boxes where three or more 0 counts are found.**

This is a very robust method when the data is dense in the entity under consideration and will flag potential anomalies or outliers that need to be checked one by one. The test conducted on GEMAS data (Figure 18), clearly shows some of the anomalies, including new ones linked to mining waste, but neglects those in certain countries. This seems to be related to sampling density. With few samples, there is less chance of identifying the reference population (background) against which to compare "flatter" anomalies. Applying this method to large-scale data is therefore unsatisfactory. It cannot be applied without modifying large-scale monitoring. In the case of large-scale data, we would need to search not only for sequences with zero probability, but also for areas of low probability density (<$\varepsilon$). With this type of approach, as $\varepsilon$ is chosen by an operator, an epistemic component is added, and all data points are assigned an anomaly score. The maximum score is obtained with zero probability bands. Non-parametric density estimation algorithms make this possible (De Simone and Morandini, 2020) and could be adapted for ISLANDR, but this becomes a clustering problem addressed in the next chapter.





Funded by the European Union

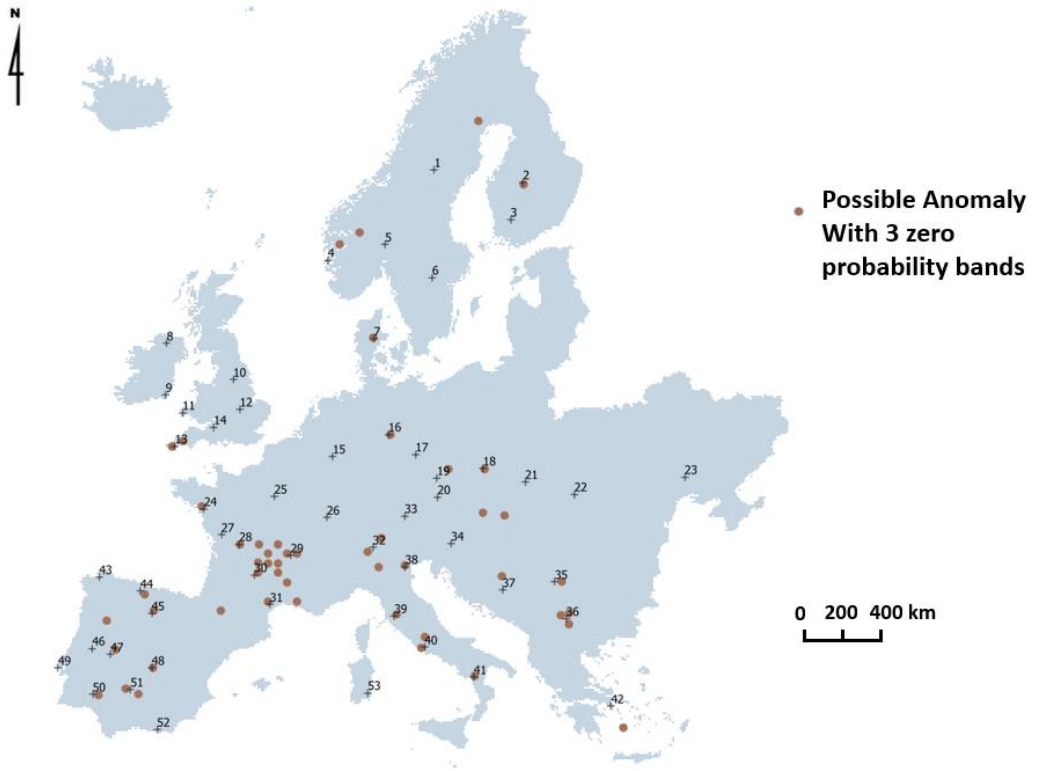


**Figure 18: Anomaly detection on the GEMAS As Ap set by 3 consecutive zero probability bands**

Conducted on the surface soils of the Toulouse ITA, with an average coverage of around one sample per kilometer (Figure 19), it displays 34 anomalies on 10 parameters (Pb, As, Cd, Co, Cr, Cu, Hg, Ni, Zn, and C10C40), and seems more sensitive than the expert (Belbeze et al., 2022), who found only 19 anomalies. Note that all 19 are identified.





Funded by
the European Union

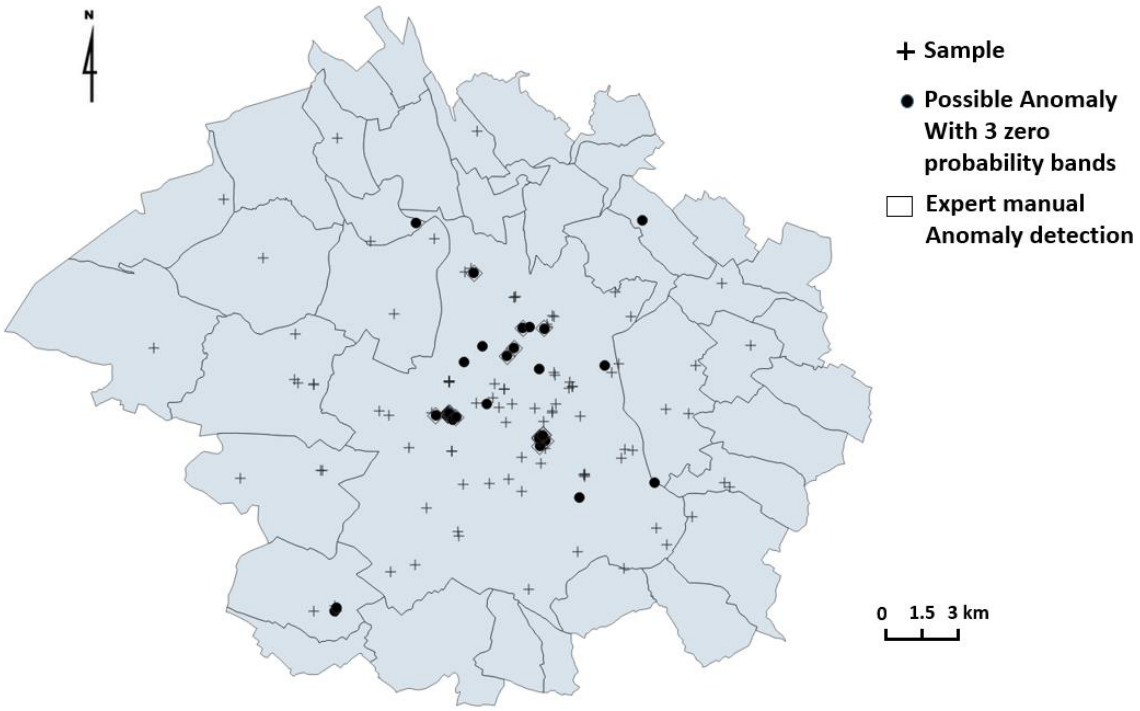


**Figure 19: City of Toulouse ITA, 138 surface samples, 34 possible anomalies separated by 3 zero probability bands, 19 selected by polluted soil expert (Belbeze et al, 2019).**

The city of Toulouse conducted a major historical study of the data in its possession to produce a very precise historical urban inventory (HUI) on a parcel-by-parcel scale. A review of all the city's environmental diagnostics was also conducted to establish which sites had been, or would have to be, remediated, known as SSP zones. Subsequently, the SSP zones are consulted by the city's building permit departments and must be carefully monitored by these departments. This data is not publicly available. However, in ISLANDR project it is possible to use this data to evaluate anomaly detection (Figure 20).





Funded by
the European Union

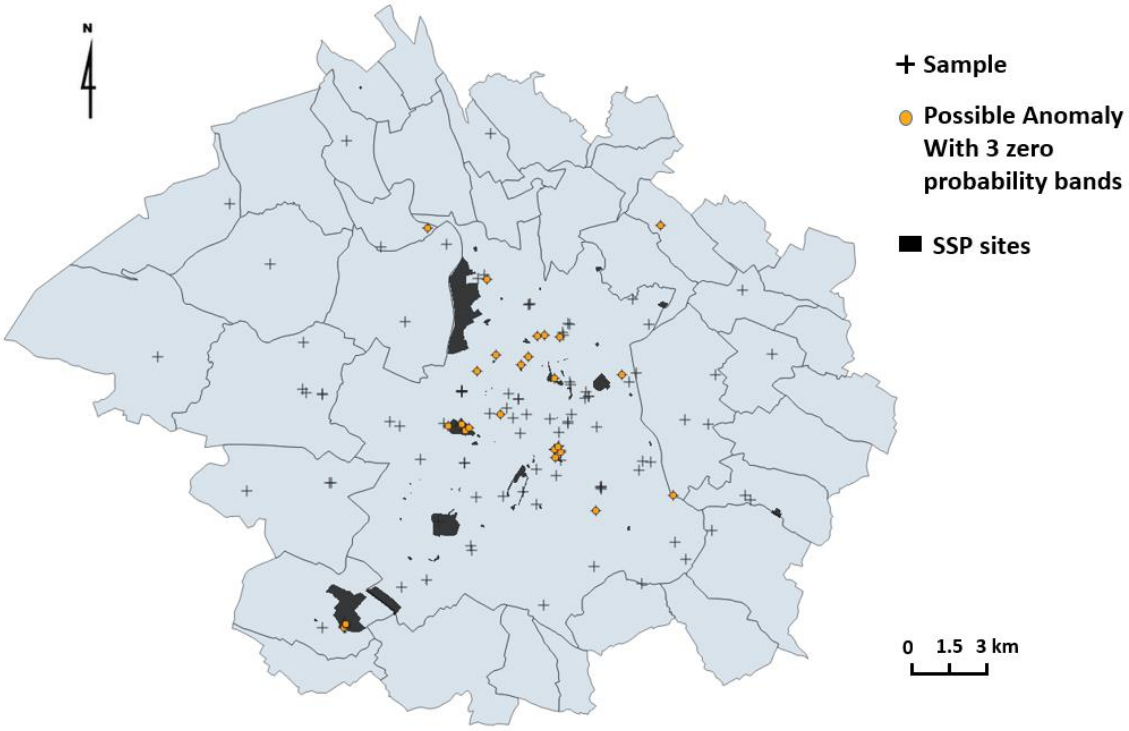


**Figure 20: City of Toulouse ITA, 138 surface samples, 34 possible anomalies separated by 3 zero probability bands, ground truth "SSP" site as referenced by the city (Belbeze et al., 2022).**

Examining this figure shows that:

- When a sample is taken from an SSP zone, it detects it as an anomaly. If the SSP zone is not covered by a sample, it is not detected, as the notion of scale does not allow it to be seen.
- Nevertheless, new zones have been identified on the ITA3 data and merit further verification.

In the end, this detection works better on densely sampled areas and could be implemented on densely sampled ITAs.

## 2.6. Spatial clustering

All clustering methods are sensitive to anomalous values, and those values always occupy most of the identified clusters. Clustering can therefore be performed in such a way that the final result is of little importance, but the initial results indicate anomalies. For ISLANDR, an extremely effective spatial clustering method used in brain imaging has been adopted: SFCM, or spatial fuzzy C-means (Cai et al., 2007; Zhao et al., 2013). The R-based implementation by Gelb and Apparicio (2021), called geocmeans, performs well on large PC configurations and was used on the GEMAS Ap data.





Funded by
the European Union

**With an unambitious target of k=10 clusters, we obtain the already promising result shown in Figure 21. We simply count the number of pixels per cluster and classify them; anomalies will always have small surface areas (**

Table 4).

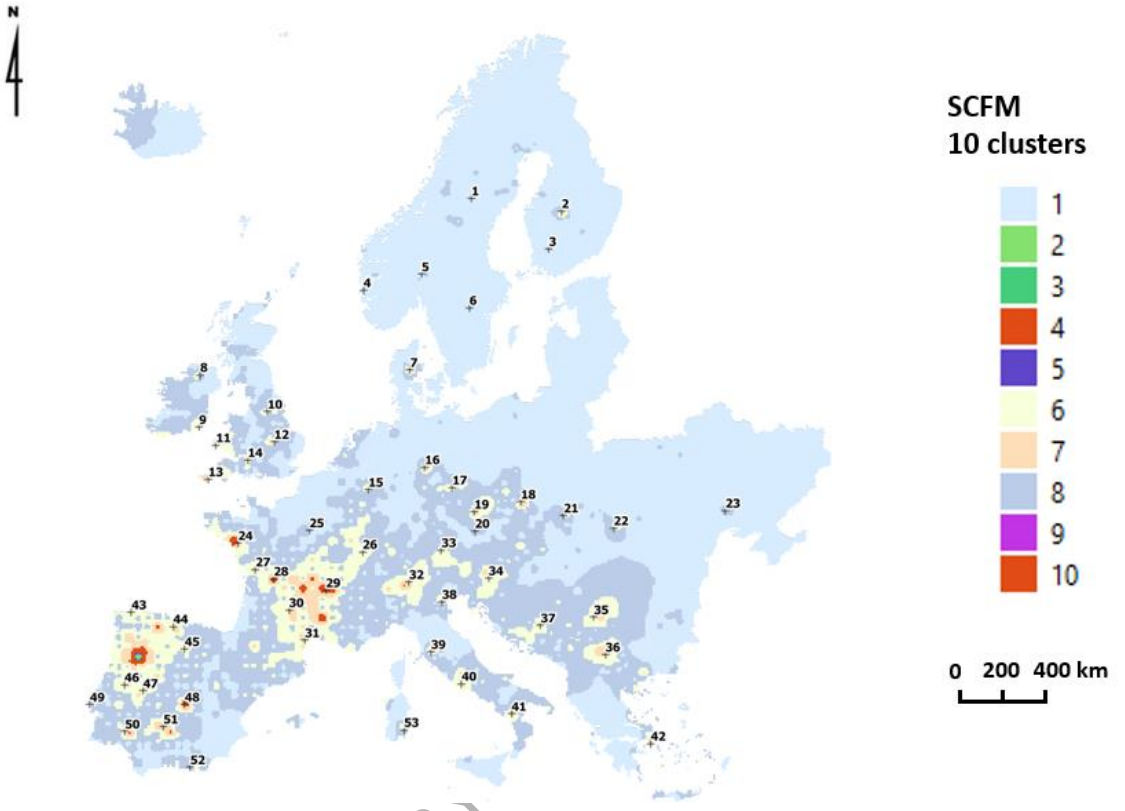


**Figure 21: SFCM clustering of GEMAS As data, k=10**

**Table 4: Cluster areas, GEMAS data, SFCM clustering, k=10**

| Cluster No. | Pixels 10 km x 10 km |
|---|---|
| **2** | **8** |
| **3** | **18** |
| **9** | **18** |
| **5** | **20** |
| **4** | **64** |
| **10** | **1450** |
| **7** | **6818** |
| 6 | 28368 |
| 1 | 34926 |
| 8 | 128632 |

With the objective of correctly imaging anomalies in Nordic countries, the k number was increased to 26 and gives very good results. (Figure 22 and Table 5





Funded by the European Union

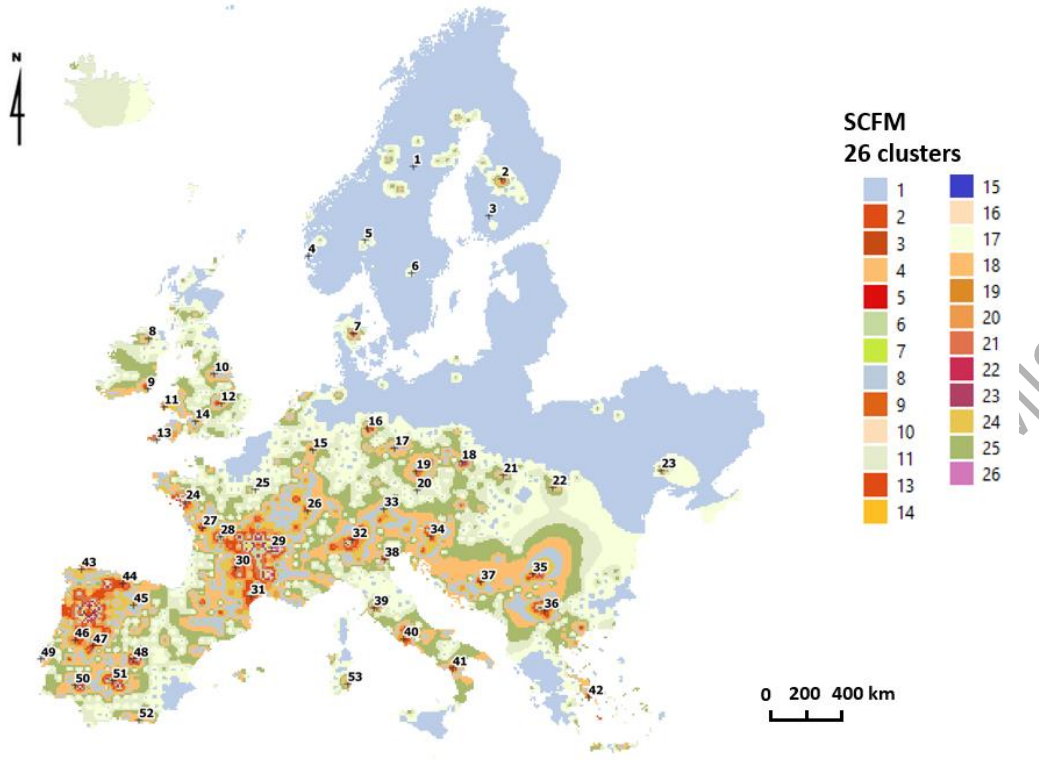


**Figure 22: SFCM clustering of GEMAS As data, k=26**

**Table 5: Cluster areas, GEMAS data, SFCM clustering, k=26**

| Cluster No. | Pixels 10 km x 10 km | Cluster No. | Pixels 10 km x 10 km |
|---|---|---|---|
| **12** | **0** | **16** | **1008** |
| **2** | **8** | **6** | **1488** |
| **5** | **10** | **23** | 3013 |
| **3** | **12** | **21** | 9303 |
| **9** | **18** | **13** | 10231 |
| **14** | **28** | 20 | 22200 |
| **4** | **36** | 8 | 23680 |
| **19** | **38** | 1 | 24052 |
| **15** | **105** | 24 | 43776 |
| **10** | **140** | 11 | 75350 |
| **7** | **182** | 18 | 78732 |
| **22** | **352** | 17 | 134691 |
| **26** | **832** | 25 | 149975 |

Spatialized statistics can then be produced on the clusters (Table 5). These can then be grouped by content, giving a threshold map of content that takes into account their spatial presence. This map then joins the one already obtained by EPH, but with a different anomaly zoning (Figure 23).





Funded by the European Union

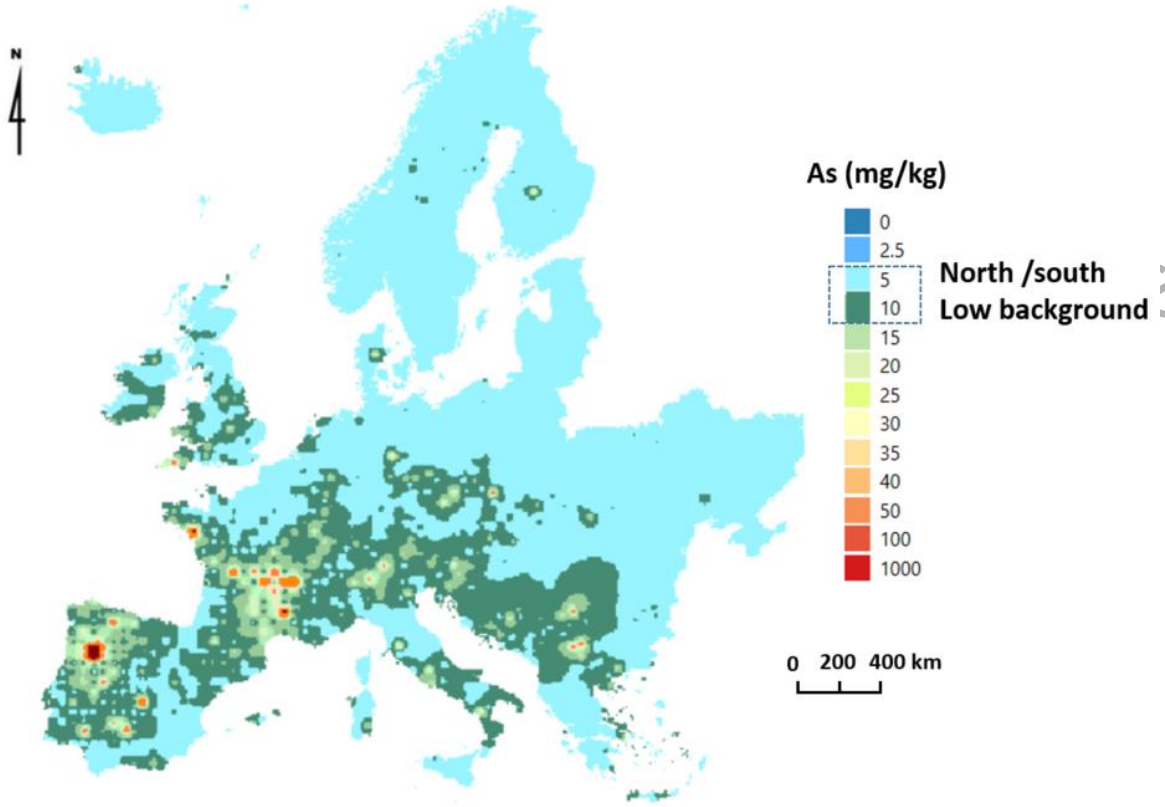


**Figure 23: Cluster surface coloured map, GEMAS data, SFCM clustering, k = 26**



**Table 6: Cluster statistics, GEMAS data, SFCM clustering, k=26**

| Cluster No. | Pixels 10km x 10km | Min. | Q25 | Mean | Med. | Q75 | Q90 | Max. | TIF |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 8 | 516 | 520 | 545 | 544 | 569 | 573 | 575 | 600 |
| 5 | 10 | 229 | 231 | 232 | 232 | 233 | 233 | 234 | 235 |
| 3 | 12 | 329 | 330 | 331 | 331 | 333 | 334 | 335 | 336 |
| 9 | 18 | 282 | 282 | 283 | 283 | 283 | 283 | 283 | 284 |
| 14 | 28 | 314 | 315 | 315 | 315 | 315 | 315 | 315 | 315 |
| 4 | 36 | 122 | 131 | 135 | 137 | 139 | 142 | 147 | 147 |
| 19 | 38 | 206 | 208 | 209 | 209 | 211 | 211 | 212 | 213 |
| 15 | 105 | 150 | 158 | 163 | 167 | 167 | 169 | 170 | 174 |
| 10 | 140 | 88 | 99 | 102 | 103 | 107 | 110 | 113 | 115 |
| 7 | 182 | 62 | 72 | 76 | 77 | 81 | 83 | 85 | 88 |
| 22 | 352 | 76 | 79 | 86 | 82 | 93 | 99 | 100 | 110 |
| 26 | 832 | 57 | 64 | 66 | 66 | 68 | 70 | 75 | 73 |
| 16 | 1008 | 48 | 53 | 55 | 55 | 58 | 60 | 67 | 63 |
| 6 | 1488 | 26 | 33 | 36 | 36 | 38 | 39 | 43 | 41 |
| 23 | 3013 | 37 | 42 | 45 | 45 | 47 | 49 | 56 | 52 |
| 21 | 9303 | 18 | 27 | 29 | 29 | 30 | 32 | 37 | 34 |
| 13 | 10231 | 17 | 22 | 24 | 24 | 25 | 26 | 29 | 27 |
| 20 | 22200 | 13 | 19 | 20 | 20 | 21 | 21 | 23 | 23 |
| 8 | 23680 | 10 | 13 | 14 | 14 | 15 | 15 | 17 | 16 |



Funded by the European Union

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 24052 | 6 | 6 | 6 | 6 | 6 | 6 | 7 | 6 |
| 24 | 43776 | 10 | 16 | 17 | 16 | 17 | 18 | 20 | 19 |
| 11 | 75350 | 6 | 8 | 9 | 9 | 9 | 9 | 11 | 10 |
| 18 | 78732 | 8 | 11 | 12 | 12 | 12 | 13 | 14 | 13 |
| 17 | 134691 | 6 | 7 | 7 | 7 | 8 | 8 | 8 | 8 |
| 25 | 149975 | 7 | 10 | 10 | 10 | 11 | 11 | 12 | 12 |

The SFCM method therefore gives good results in anomaly detection and can be adapted to multi-variables. As such, it should be retained for ISLANDR. Another interesting aspect of SFCM is that it calculates a membership function for each pixel in each cluster, which can be converted into a mass function. This opens up a whole field of potential calculations with a Dempster-Shafer formalism; it has been explored by Hammami (2017) and Haouas (2019). This kind of approach could be useful, as it would make it possible to combine in an original way the uncertainty of the interpolator for the pixel with that of the anomaly search by clustering on the pixel.

## 2.7. ITA3 comparison using the Nemerow index with GEMAS

Nemerow indexes (Nemerow,1985), otherwise known as ratios, enrichment factors, or response factors, are ubiquitous in geochemistry and widely used. Their design is simple, with a minimum of underlying assumptions. They usually consist of searching for a geochemical background B in deep soil or in a nearby control and measuring its enrichment using a ratio. The interpretation criteria vary according to how the index is constructed (Table 7).

**Table 7: Several Nemerow indexes**

| Name | Index name | Formula | Classes | Interpretation | References |
|---|---|---|---|---|---|
| Contamination factor<br>Pollution index<br>Contamination index<br>Anthropogenicity | $P_i$<br>$Cf_i$<br>Tc/Bc<br>Apn% | $P_i = Cf_i = \frac{C_i}{B_i}$ | < 1<br>1–3<br>3–6<br>> 6 | Low<br>Moderate<br>Considerable<br>Very high | Hakanson (1980) |
| Enrichment factor | EF | $EF = \frac{C_i/N_i}{N_i/N_b} = \frac{C_i}{N_i} \cdot \frac{N_b}{C_b}$<br>For example, weighting by iron<br>$EF = \frac{C_i}{C_{Fe}} \cdot \frac{C_b}{C_{bFe}}$ | | | |
| Geoaccumulation index | $I_{geo}$ | $I_{geo} = \ln\left(\frac{C_i}{1.5\,B_i}\right)$ | < 0<br>0–1<br>1–2<br>2–3<br>3–4 | Practically unpolluted<br>Unpolluted to moderately polluted | Mullller (1981) |





Funded by the European Union

| | | | | | |
|---|---|---|---|---|---|
| | | | 4–5<br>> 5 | Moderately polluted<br>Moderately to strongly polluted<br>Strongly polluted<br>Strongly to extremely polluted<br>Extremely polluted | |
| Pollution load index<br>Modified contamination degree | PLI<br>mCd | $PLI = \frac{\prod_{i=1}^{n} Cf_i}{n}$ | < 0<br>> 0 | Non contaminated<br>Contaminated | Thomlinson et al.(1980) |
| Pollution index Nemerow<br>Nemerow integrated pollution index (NIPI) | PIN<br>NIPI | $PI_{Nemerov} = \sqrt{\frac{\left(\frac{1}{n}\sum_{i=1}^{n} P_i\right)^2 + \max\left(P_i^2\right)}{2}}$ | < 0.7<br>0.7–1<br>1–2<br>2–3<br>> 3 | Excellent<br>Clean<br>Slight pollution<br>Moderate pollution<br>Heavy pollution | Nemerow (1985) |
| Pollution index China – Pakistan | PI<br>PLI | $PI_{chinois} = \sqrt[n]{\prod_{i=1}^{n} P_i}$ | | | |
| Degree of contamination | $C_d$ | $C_d = \sum_{i=1}^{n} Cf_i$ | < 8<br>8–16<br>16–32<br>> 32 | Low<br>Moderate<br>Considerable<br>Very high | Hakanson (1980) |
| Ecological risk factor | | $E_i = Tf_i \cdot Cf_i$ | < 40<br>40–80<br>80–160<br>160–320<br>> 320 | Low<br>Moderate<br>Considerable<br>High<br>Dangerous | Hakanson (1980) |
| Potential ecological risk index | RI | $RI = \sum_{i=1}^{n} E_i$ | < 150<br>150–300<br>300–600<br>> 600 | Low<br>Moderate<br>High<br>Very high | Hakanson (1980) |

*With $C_i$ the centration, $B_i$ the geogenic background, and $Tf_i$ the toxicity factor of substance i, n the number of pollutants. $N_i$ reference element or normalization concentration, $C_b$ background concentration, $N_b$ reference element or normalization concentration for the background.*

For ISLANDR, the interpolated raw data should be compared to the GEMAS data extracted at this point. The Toulouse ITA exercise was carried out for lead (Pb). This element is likely to present far more anomalies than arsenic.





Funded by the European Union

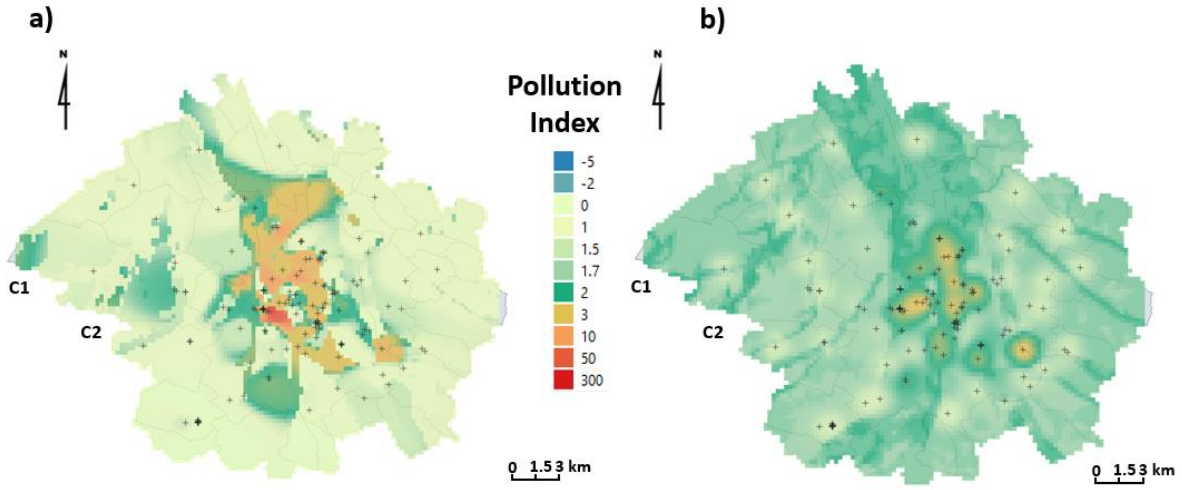


**Figure 24: Two Pollution Index maps calculated for lead (Pb) concentration with 138 surface samples marked with small X's with a) EPH map with geologic covariate and b) dual inequality kriging with geologic covariate (Belbeze et al., 2019).**

The mappings in Figure 24 are of interest, but they are both different. Kriging assigns the entire city a moderate pollution index close to 2, whereas EEPH does not; this is due to kriging smoothing. Two zones in indices C1 and C2 classified as 2 by the mappings raise questions about EEPH interpolation and, to a lesser extent, kriging. C1 is a Bouconne Forest preserved natural zone where no sampling was authorized, and C2 is the valley of the town of Colomiers. After examining the geological data, anomalies C1 and C2 correspond to particular lithologies of the Garonne terraces, which are wrongly assigned a high alluvial value calculated on the same geology but on samples from denser urban areas. This seems to be a common problem with drift kriging, which assigns the same mean value to all alluviums, with the addition of a stochastic component.

However, the pollution index produces interesting land zoning maps (Figure 25), particularly when compared to the map of geographically coherent entities (land units) already produced by Belbeze et al. (2019) using a different spatial clustering technique. They can be performed also to other selected ITAs.





Funded by the European Union

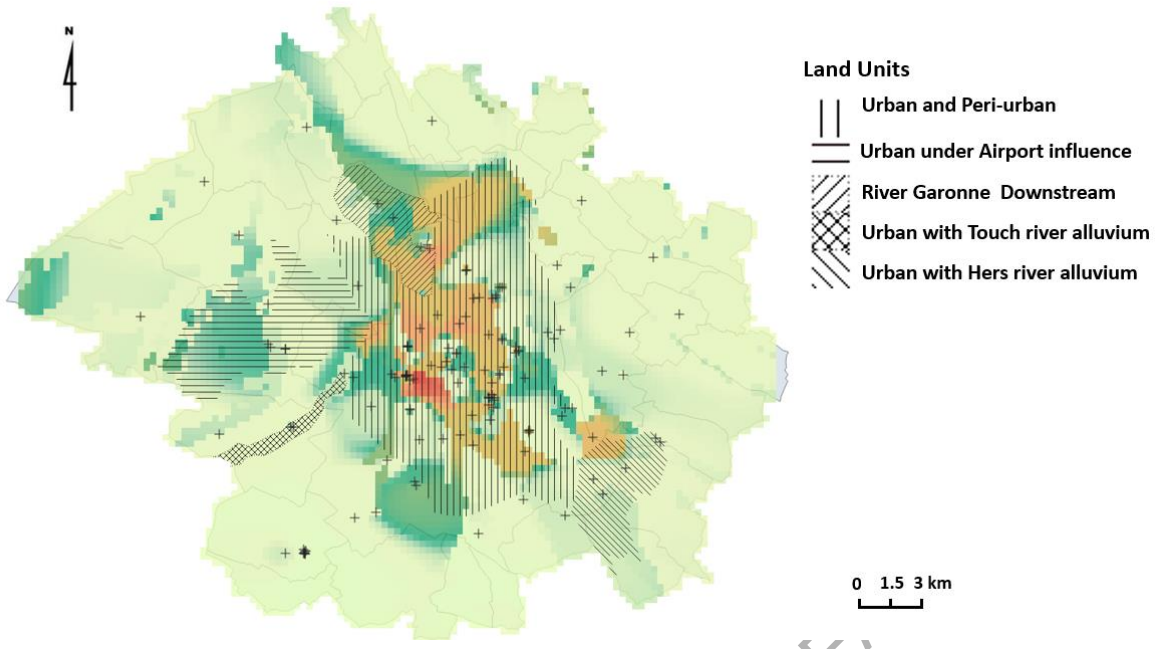


**Figure 25: Land units as established for the city of Toulouse ITA on the EPH Pb continuous map, surface soil, 138 samples, Belbeze et al. (2019)**





Funded by the European Union

# 3. Ensemble Anomaly Detection

## 3.1. Principle

Anomaly detection tests with seven techniques all produced informative maps (Table 8). Those that performed best for our GEMAS test dataset were Reimann statistics (TIF), spatial clustering, and the fractal singularity index (Figure 26).

**Table 8: Anomaly detection techniques adapted to ISLANDR projects and uncertainty propagation possibilities**

| Method | Behavior on large-scale dataset | Type of anomaly detected | Propagation of uncertainties |
|---|---|---|---|
| Reimann statistics (Tukey Inner Fence – TIF) | Sensitive to multi-modality and symmetry of elementary distributions. Powerful method | Range outlier | - |
| Local Moran index | Works well on well-marked anomalies but not particularly sensitive to flat anomalies | Spatial outlier | Monte Carlo |
| C-A fractal method | Sensitive to multi-modality and symmetry of distributions. It gives good results | Spatial outlier | Monte Carlo |
| Fractal singularity index | The fractal singularity method gives excellent results | Spatial outlier | Monte Carlo |
| Zero probability bands | Poor performance in mono-elementary on GEMAS. Good results in multivariate on sites like the Toulouse ITA. | Relationship outlier<br>Range outlier | - |
| Spatial clustering | The spatial fuzzy C-means method gives excellent results. It can also be used in multivariate | Spatial outlier<br>Relationship outlier | Possibilities Dempster-Shafer |
| Pollution index with GEMAS as reference | Highly sensitive Produces useful zoning maps. | Spatial outlier | Monte Carlo |





Funded by the European Union

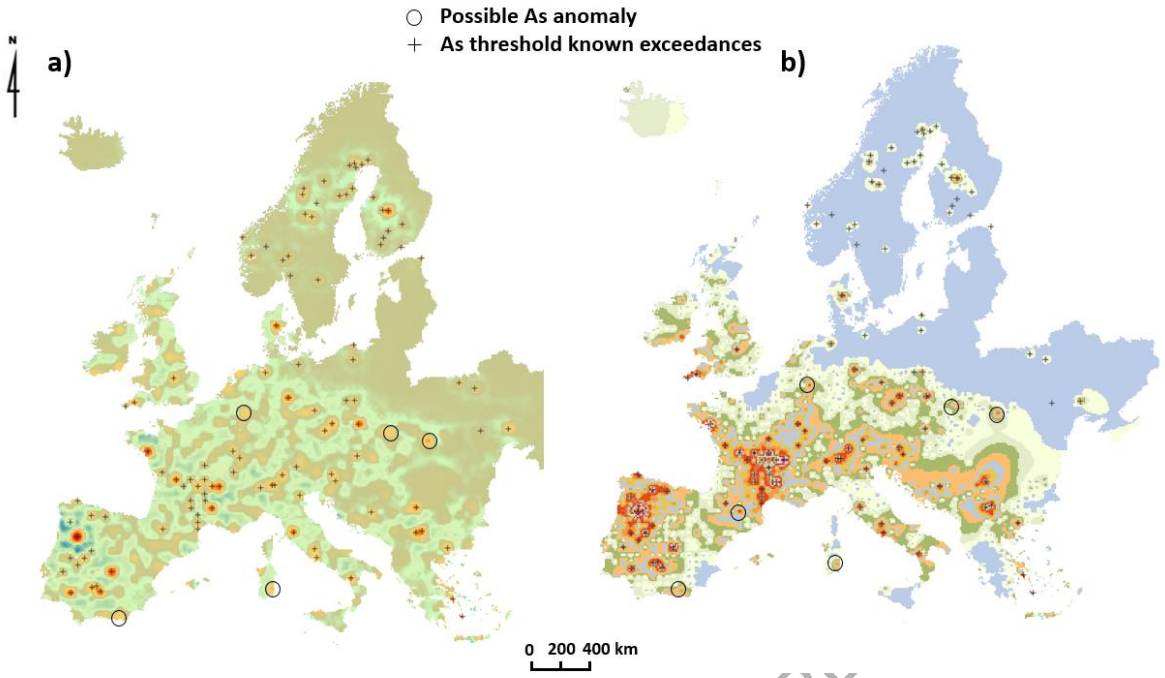


**Figure 26: As known exceedance threshold (Reimann et al., 2018) on the new anomalies detection maps produced with a) Fractal singularity index, and b) SFCM Clustering k=26**

By maintaining and applying these techniques to all our datasets, the detection algorithms will behave like a panel of experts whose consensus can be sought.

- The situation analysis values are calculated on the raw data, then applied to the samples:
    - Anomaly threshold (Q90, Q98, TIF)
    - Probability bands
- On a neutral EEPH interpolation (Annex 1) that magnifies anomalies:
    - C-A fractal (LBKG/HBKG/A threshold)
    - Singularity index
    - Local Moran index
    - Clustering index (single- or multi-element)
    - Nemerow index, if available

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 fuzzification (Table 9).





Funded by the European Union

**Table 9: 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 | BCKG | 0 | 1 | 2 | 1 | 0 | 0 | (0, 22.6, 41.5) |
| Unlikely | BCKG | - | 0 | 2.5 | 0 | - | - | (0, 0, 22.6) |

Tested on real datasets, each technique also has its own advantages and disadvantages that must be considered. This uncertain information can be well modeled using triangular fuzzy numbers, or TFNs (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 10).

**Table 10: Anomaly detection TFN meta-ranking based on GEMAS and ITA3 experiments**

| Anomaly index | Type of outlier | Based on GEMAS experiment | Expert opinion | 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 | 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 | Spatial outlier | A statistical fractal method. Limits are uncertain. Less precise than the window-based method on the GEMAS dataset | Average | (0.05, 0.04, 0.1) |
| Singularity index | Spatial outlier | Window-based statistics. Efficient on GEMAS dataset | Good | (0.17, 0.24, 0.3) |
| Moran index | Spatial outlier | Window-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 Multivariate outlier | Window-based 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 | - for ITA only |





Funded by the European Union

Using fuzzy logic, fuzzy multi-criteria decision-making methods (FuzzyMCDM – FMCDM) are used for classification problems where uncertainty, vagueness, and/or imprecision are present in the decision matrix (Ceballos et al., 2016, 2017).

The FMCDM method takes as input a decision matrix whose rows correspond to the various alternatives to the question (here, these are samples), denoted $A_1$, $A_2$,..., $A_m$, and the columns correspond to n selected criteria, denoted $C_1$, $C_2$, ..., $C_n$. Since one criterion may be more relevant than others, a triangular fuzzy weight $w_1$, $w_2$, ..., $w_n$ is assigned to each criterion. The following table shows the input data for the model, which is a decision matrix with criteria on the horizontal axis and alternatives on the vertical axis:

**Table 11: Decision matrix**

| | $C_1$ | $C_2$ | ... | $C_n$ |
|---|---|---|---|---|
| | $\widetilde{w}_1$ | $\widetilde{w}_2$ | ... | $\widetilde{w}_n$ |
| $A_1$ | $x_{11}$ | $x_{12}$ | ... | $x_{1n}$ |
| $A_2$ | $x_{21}$ | $x_{22}$ | ... | $x_{2n}$ |
| ... | ... | ... | ... | |
| $A_m$ | $x_{m1}$ | $x_{m2}$ | ... | $x_{mn}$ |

The formulas used for sorting are as follows:

$$\widetilde{D} = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_m \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & & C_n \end{array} \\ \begin{bmatrix} \widetilde{x}_{11} & \widetilde{x}_{12} & \cdots & \widetilde{x}_{1n} \\ \widetilde{x}_{21} & \widetilde{x}_{22} & \cdots & \widetilde{x}_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \widetilde{x}_{m1} & \widetilde{x}_{m2} & \cdots & \widetilde{x}_{mn} \end{bmatrix} \end{array}, \quad i = 1,2,\ldots, m; \quad j = 1,2,\ldots, n$$

$$\widetilde{W} = [\widetilde{w}_1, \widetilde{w}_2, \ldots, \widetilde{w}_n], \quad j = 1,2,\ldots n$$

$$P(\widetilde{x}_1 \geq \widetilde{x}_2) = \lambda \max\left\{1 - \max\left[\frac{x_2^M - x_1^L}{x_1^M - x_1^L + x_2^M - x_2^L}, 0\right], 0\right\} +$$

$$(1-\lambda)\max\left\{1 - \max\left[\frac{x_2^U - x_1^M}{x_1^U - x_1^M + x_2^U - x_2^M}, 0\right], 0\right\}$$

There are several ways to apply these multi-criteria sorting equations to fuzzy numbers, including the VIKOR method, the TOPSIS method, and the multi-MOORA method. 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.





Funded by the European Union

With regard to calculating uncertainties on anomalous values, two situations arise:

- Several mappings generated by the interpolator, or a quantile of them, are submitted to the detection algorithm and the impact on the anomaly centroids is determined. This method is comparable to Monte Carlo (this would be the case for the Moran index, fractal singularity, and pollution index with GEMAS).
- For spatial clustering, the SFCM selected makes it possible to assign a mass to each cluster, and thus to merge this information with that of the interpolation, which must first be converted into a possibility. A Dempster-Shafer formalism could be adapted.

## 3.2. Application to GEMAS arsenic data

The detection results can be compiled by aggregating the results of the various algorithms at the level of each sample (Table 12). We can see a degree of consistency between methods, but also differences that become more apparent after partitioning (Table 13). So, a French sample (3586) is classified as anomalous by Reimann statistics, the Moran index, and clustering, while the area concentration, singularity index, and zero probability band methods are inconclusive.

**Table 12: Some anomaly detection algorithms on the GEMAS arsenic dataset Ap sample and EEPH neutral map of the same sample, extract of 10 samples**

| ID | COUNTRY | As in Ap soil | As in moss | Reimann statistics | Zero probability Bands | Index C-A fractal | Singularity index | Moran index | Anomaly cluster Index |
|---|---|---|---|---|---|---|---|---|---|
| 3586 | FRA | 41.1 | 0.36 | TIF | 0 | 1 | 1.92 | 4.37 | 3 |
| 3886 | FIN | 37.1 | 0.17 | TIF | 3 | 4 | 0.98 | 5.57 | 3 |
| 5042 | SPA | 41.1 | 0.32 | TIF | 2 | 1 | 1.90 | 5.05 | 3 |
| 5128 | ITA | 43.2 | 0.31 | TIF | 2 | 1 | 1.35 | 6.68 | 3 |
| 5173 | DEN | 32.7 | 0.21 | TIF | 3 | 4 | 1.24 | 3.85 | 3 |
| 3229 | SRB | 61.6 | 1.02 | Q98 | 3 | 1 | 1.44 | 15.09 | 3 |
| 3289 | FRA | 126.4 | 0.32 | Q98 | 3 | 2 | 1.25 | 75.78 | 3 |
| 3290 | SRB | 51.1 | 0.68 | Q98 | 3 | 1 | 1.34 | 14.04 | 3 |
| 3367 | FRA | 91.8 | 0.37 | Q98 | 3 | 2 | 1.52 | 36.34 | 3 |
| 3493 | FRA | 45.7 | 0.39 | Q98 | 0 | 1 | 1.87 | 9.26 | 3 |
| ... | ... | | | | | | | | |





Funded by the European Union

**Table 13: Partition, GEMAS As dataset, extract of 10 samples**

| ID | COUNTRY | As in Ap soil | Reimann statistics | Zero probability Bands | Index C-A fractal | Singularity index | Moran index | Anomaly cluster index | As in moss | Air dust PM2.5 |
|---|---|---|---|---|---|---|---|---|---|---|
| 3586 | FRA | 41.1 | VC | ALN | ALN | ALN | VL | VC | ML | M |
| 3886 | FIN | 37.1 | VC | VC | VC | VL | VL | VC | L | ML |
| 5042 | SPA | 41.1 | VC | VL | ALN | ALN | VL | VC | ML | ML |
| 5128 | ITA | 43.2 | VC | VL | ALN | L | VL | VC | ML | M |
| 5173 | DEN | 32.7 | VC | VC | VC | L | VL | VC | ML | M |
| 3229 | SRB | 61.6 | VL | VC | ALN | L | VC | VC | H | M |
| 3289 | FRA | 126.4 | VL | VC | L | L | VC | VC | ML | M |
| 3290 | SRB | 51.1 | VL | VC | ALN | L | VC | VC | MH | VH |
| 3367 | FRA | 91.8 | VL | VC | L | ALN | VC | VC | ML | M |
| 3493 | FRA | 45.7 | VL | ALN | ALN | ALN | VL | VC | ML | ML |
| ... | ... | | | | | | | | | |

*With: VC: virtually certain, VL: very likely, L: likely, ALN: as likely as not, U: unlikely and with concentration categories L: low, ML: medium low, M: medium, H: high, MH: medium high, VH: very high*

The sorting algorithms are then applied to the various samples. Each sorting ranks them from most anomalous to least anomalous (Table 14).

**Table 14: Fuzzy multicriteria meta ranking for arsenic, GEMAS As dataset, extract of 10 samples**

| ID | COUNTRY | As in Ap soil | As in moss | Air dust PM2.5 | MMOORA | TOPSIS Vector | TOPSIS Linear | VIKOR | WASPAS | Meta Ranking Sum |
|---|---|---|---|---|---|---|---|---|---|---|
| 3886 | FIN | 37.1 | L | ML | 1 | 1 | 1 | 1 | 1 | 1 |
| 5237 | PTG | 666 | M | ML | 2 | 2 | 2 | 2 | 2 | 2 |
| 5018 | SPA | 104 | ML | M | 3 | 3 | 3 | 3 | 3 | 3 |
| 3289 | FRA | 126 | ML | M | 5 | 4 | 6 | 6 | 5 | 4 |
| 3904 | FRA | 113 | M | M | 6 | 5 | 7 | 7 | 6 | 5 |
| 5173 | DEN | 32.7 | ML | M | 4 | 16 | 4 | 4 | 4 | 6 |
| 4146 | FRA | 120 | ML | ML | 7 | 6 | 8 | 8 | 7 | 7 |
| 5346 | FRA | 92.8 | ML | ML | 12 | 7 | 9 | 9 | 8 | 8 |
| 5128 | ITA | 43.2 | ML | M | 11 | 17 | 5 | 5 | 9 | 9 |
| 3229 | SRB | 61.6 | H | M | 8 | 10 | 10 | 10 | 10 | 10 |
| 3290 | SRB | 51.1 | MH | VH | 9 | 11 | 11 | 11 | 11 | 11 |
| 3594 | SPA | 70.6 | ML | M | 10 | 12 | 12 | 12 | 12 | 12 |
| 4438 | SPA | 50.2 | ML | M | 13 | 13 | 13 | 13 | 13 | 13 |
| 5176 | CZR | 61.3 | ML | M | 14 | 14 | 14 | 14 | 14 | 14 |
| ... | ... | | | | | | | | | |

*With concentration categories L: low, ML: medium low, M: medium, H: high, MH: medium high, VH: very high*

An overall meta-ranking can be assigned to the anomaly using a simple sum or a more complex aggregation system. Contents in mosses and dusts are anthropogenic clues and





Funded by the European Union

allow rapid classification of anomalies. For example, sample 3886 from Finland has a probable geogenic origin, as evidenced by the low content found in the underlying mosses; sample 3229 from Serbia has a high content in mosses and dusts, both indicative of diffuse anthropogenic contamination. Figure 27 shows how GEMAS arsenic data was processed. We can see that the anomalies identified by Tarvainen et al. (2013) are included, as well as new low-signal anomalies, which are one of ISLANDR's objectives and will be closely examined.

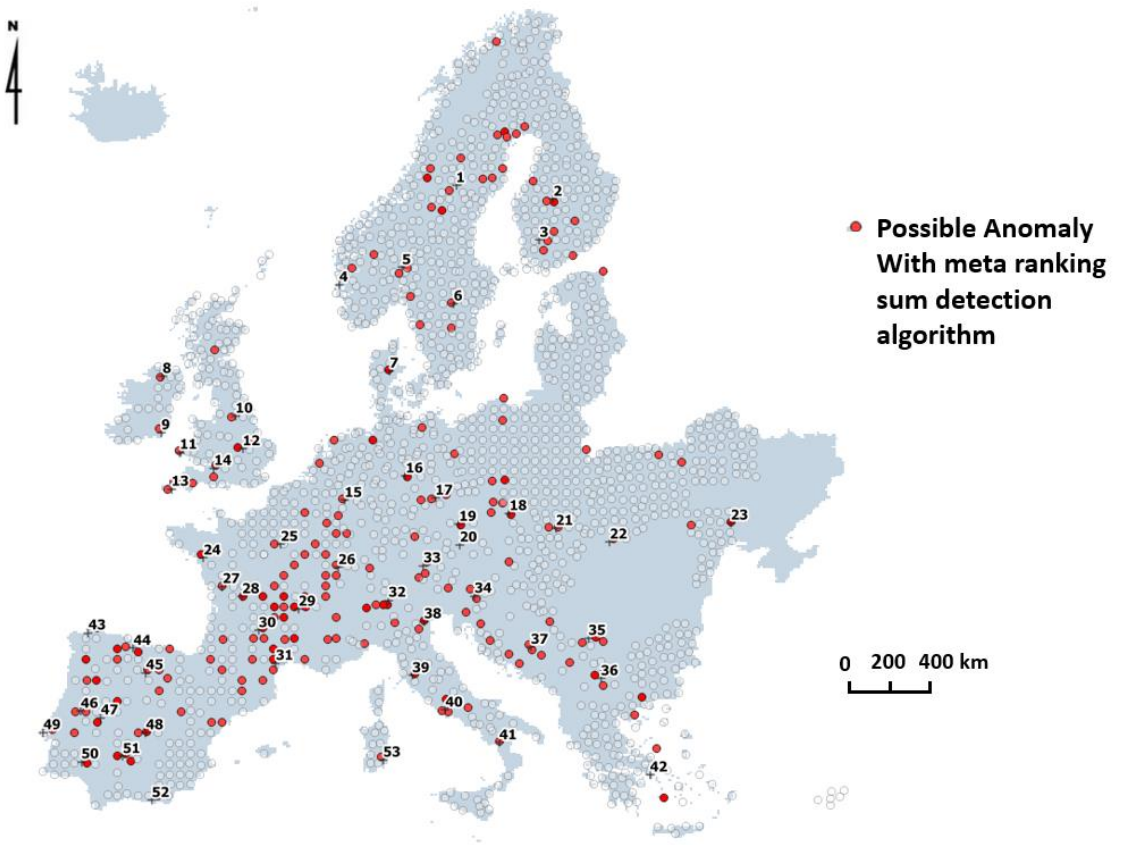


**Figure 27: Anomalies in GEMAS As Ap samples as pinpointed by the ISLANDR detection algorithm**

## 3.3. Application to ITA3 data

The anomalous meta-ranking procedure can be initiated on ITA3 using the raw data and the neutral EEPK obtained with the contents (Figure 28).





Funded by
the European Union

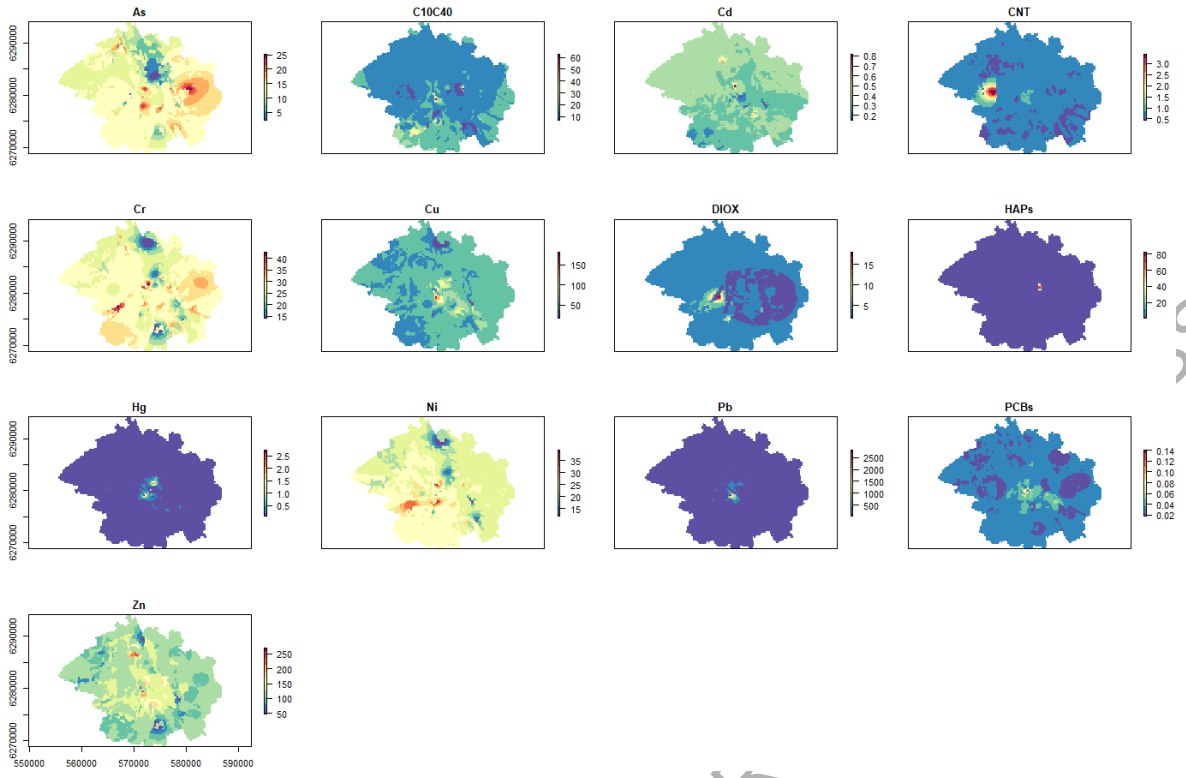


**Figure 28: City of Toulouse ITA, 138 surface samples, EEPH surface baseline maps**

Meta-ranking is quickly obtained for the 8 metal(oid)s (As, Cd, Cr, Cu, Hg, Ni, Pb, and Zn). These show the presence of 41 anomalies out of the 138 samples as researched for ISLANDR (Table 15).

In these samples, and in the rest of the city as a whole, moss samples are anomalous and moderately high levels in dust indicate a diffuse urban influence on these samples.





Funded by the European Union

**Table 15: City of Toulouse ITA, 138 surface samples with their meta-rankings for As, Cd, Cr, Cu, Hg, Ni, Pb, and Zn**

| NAME | Min (7 Meta Rank Sum) | Anomaly no. | As | As Meta Ranking Sum | Cd | Cd Meta Ranking Sum | Cr | Cr Meta Ranking Sum | Cu | Cu Meta Ranking Sum | Hg | Hg Meta Ranking Sum | Ni | Ni Meta Ranking Sum | Pb | Pb Meta Ranking Sum | Zn | Zn Meta Ranking Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0310051M_SLE09 | 1 | 0 | 18 | 1 | 0.1 | 99 | 31 | 2 | 26 | 36 | 0.05 | 133 | 25 | 8 | 33 | 79 | 100 | 30 |
| 0310245Y_SLE02 | 1 | 27 | 11 | 77 | **0.9** | 1 | **43** | 1 | **88** | 10 | **0.26** | 14 | 28 | 2 | 70 | 25 | **150** | 10 |
| P329 | 1 | 32 | 15 | 6 | 0.25 | 120 | 23 | 87 | **110** | 4 | **4.9** | 3 | 30 | 1 | **480** | 2 | **150** | 5 |
| P113 | 1 | 33 | 13 | 60 | 0.31 | 21 | 21 | 100 | **120** | 3 | **4.8** | 1 | 22 | 40 | **4600** | 1 | **150** | 7 |
| P327 | 1 | 40 | **31** | 2 | **0.81** | 35 | 28 | 55 | **210** | 1 | **0.74** | 7 | 28 | 4 | **230** | 9 | **200** | 2 |
| P307 | 1 | 43 | 9.3 | 72 | 0.49 | 103 | 14 | 95 | **220** | 2 | **0.27** | 13 | 13 | 60 | **95** | 10 | **860** | 1 |
| 0310221X_SLU01 | 2 | 36 | 17 | 25 | 0.5 | 62 | 16 | 8 | **95** | 12 | **2.7** | 2 | 14 | 92 | **250** | 5 | **170** | 24 |
| Q16 | 2 | 39 | 12 | 105 | 0.74 | 2 | 14 | 103 | **76** | 27 | 0.15 | 30 | 13 | 114 | **100** | 16 | **190** | 15 |
| Z5T8 | 3 | 0 | 17 | 12 | 0.2 | 33 | 34 | 3 | 30 | 48 | 0.05 | 91 | 18 | 102 | 24 | 74 | 65 | 90 |
| P289 | 3 | 10 | 19 | 3 | 0.51 | 15 | 27 | 78 | 36 | 13 | 0.16 | 6 | 29 | 5 | **230** | 3 | 93 | 9 |
| P348 | 3 | 31 | 16 | 4 | 0.63 | 79 | 30 | 51 | **81** | 7 | **0.33** | 10 | 29 | 3 | **190** | 4 | **150** | 4 |
| 0310211L_SLE06 | 3 | 34 | 16 | 88 | 0.45 | 3 | 29 | 16 | **75** | 9 | **0.2** | 31 | 22 | 45 | 74 | 47 | **160** | 23 |
| P326 | 3 | 38 | 14 | 31 | **0.9** | 5 | 24 | 92 | **98** | 6 | **0.44** | 12 | 26 | 7 | **89** | 12 | **190** | 3 |
| Z15T6 | 4 | 0 | 8.8 | 74 | 0.2 | 30 | 32 | 4 | 13 | 67 | 0.05 | 129 | 29 | 42 | 13 | 118 | 50 | 110 |
| 0312151V_SLU01 | 4 | 14 | 9 | 124 | 0.5 | 72 | 14 | 13 | 38 | 30 | **0.8** | 4 | 12 | 11 | **120** | 17 | 100 | 75 |
| 0311265G_SLU01 | 4 | 22 | 8.2 | 29 | 0.72 | 4 | 19.9 | 117 | 41.9 | 96 | **0.654** | 19 | 15.8 | 124 | **255** | 11 | 123.7 | 87 |
| FGU-TLS-19 | 5 | 0 | 23 | 5 | 0.5 | 82 | 28 | 19 | 24 | 116 | 0.1 | 44 | 26 | 26 | 17 | 121 | 82 | 118 |
| Z5T6 | 5 | 0 | 19 | 7 | 0.2 | 39 | 32 | 5 | 21 | 71 | 0.05 | 93 | 32 | 15 | 18 | 119 | 89 | 80 |
| P184 | 5 | 28 | 13 | 58 | 0.2 | 77 | 19 | 99 | **170** | 5 | **0.21** | 8 | 14 | 70 | 76 | 7 | **150** | 6 |
| 0310245Y_SLE03 | 5 | 30 | 11 | 106 | 0.19 | 125 | 19 | 90 | 71 | 14 | **1** | 5 | 16 | 96 | **140** | 15 | **150** | 84 |
| Z17T3 | 6 | 0 | 10 | 76 | 0.22 | 43 | 32 | 6 | 20 | 76 | 0.05 | 131 | 20 | 107 | 30 | 76 | 69 | 93 |
| 0311316MSLE01 | 6 | 0 | 1 | 33 | 0.4 | 32 | 13.4 | 10 | 35.3 | 39 | 0.1 | 66 | 10.5 | 6 | 68 | 51 | 73.7 | 105 |
| Z31P4 | 6 | 0 | 12 | 65 | 0.26 | 6 | 22 | 44 | 28 | 62 | 0.05 | 130 | 20 | 104 | 58 | 75 | 120 | 73 |
| FGU-TLS-38 | 6 | 42 | 23 | 24 | **0.9** | 9 | 17 | 131 | **150** | 18 | **0.2** | 17 | 15 | 133 | **310** | 6 | **450** | 17 |
| Z17T5 | 7 | 0 | 9.7 | 80 | 0.2 | 47 | 31 | 7 | 15 | 81 | 0.05 | 132 | 19 | 110 | 29 | 77 | 45 | 117 |
| FGU-TLS-1 | 7 | 0 | 17 | 41 | 0.5 | 7 | 29 | 36 | 28 | 111 | 0.1 | 77 | 25 | 36 | 26 | 99 | 87 | 100 |
| FGU-TLS-13 | 8 | 0 | 18 | 8 | 0.5 | 29 | 33 | 25 | 28 | 113 | 0.1 | 80 | 28 | 24 | 34 | 101 | 110 | 98 |
| P328 | 8 | 7 | 13 | 38 | 0.48 | 101 | 21 | 89 | 50 | 8 | **0.24** | 11 | 22 | 35 | **220** | 8 | 91 | 8 |
| 0311883D_SLU01 | 8 | 41 | 10 | 121 | 0.7 | 8 | 15 | 79 | 34 | 83 | **0.4** | 57 | 13 | 91 | **270** | 21 | **360** | 12 |
| FGU-TLS-14 | 9 | 0 | 17 | 9 | 0.5 | 76 | 30 | 9 | 20 | 102 | 0.1 | 42 | 26 | 25 | 16 | 68 | 78 | 103 |
| FGU-TLS-34 | 9 | 0 | 14 | 116 | 0.5 | 126 | 36 | 11 | 22 | 64 | 0.1 | 101 | 30 | 9 | 28 | 105 | 89 | 85 |
| 0310917D_SLU01 | 9 | 3 | 8 | 91 | 0.2 | 65 | 14 | 109 | 33 | 44 | **0.32** | 9 | 11 | 77 | **80** | 14 | 83 | 67 |
| 0310211L_SLE05 | 10 | 0 | 14 | 96 | 0.19 | 10 | 27 | 42 | 26 | 21 | 0.09 | 125 | 22 | 43 | 39 | 81 | 75 | 40 |





Funded by the European Union

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FGU-TLS-28 | 10 | 0 | 18 | 10 | 0.5 | 17 | 30 | 46 | 46 | 49 | 0.1 | 90 | 34 | 10 | 29 | 56 | 100 | 35 |
| P325 | 11 | 0 | 8.4 | 69 | 0.2 | 121 | 20 | 94 | 41 | 11 | 0.05 | 25 | 14 | 59 | 23 | 18 | 70 | 11 |
| FGU-TLS-16 | 11 | 0 | 17 | 11 | 0.5 | 80 | 26 | 52 | 20 | 103 | 0.1 | 43 | 23 | 46 | 19 | 69 | 75 | 114 |
| FGU-TLS-12 | 11 | 9 | 11 | 89 | 0.5 | 11 | 19 | 123 | 39 | 86 | **0.2** | 63 | 20 | 128 | 47 | 100 | 92 | 78 |
| FGU-TLS-17 | 12 | 0 | 10 | 129 | 0.5 | 81 | 13 | 133 | 61 | 23 | 0.1 | 82 | 12 | 12 | 31 | 48 | 49 | 138 |
| FGU-TLS-36 | 12 | 0 | 14 | 75 | 0.6 | 12 | 25 | 96 | 21 | 132 | 0.1 | 103 | 25 | 44 | 49 | 107 | **220** | 27 |
| 0310216S_SLU01 | 12 | 18 | 12 | 113 | 0.29 | 24 | 31 | 12 | 36 | 22 | **0.4** | 32 | 25 | 22 | **87** | 37 | 110 | 38 |
| Z5T5 | 13 | 0 | 15 | 13 | 0.2 | 42 | 26 | 39 | 18 | 74 | 0.05 | 95 | 21 | 95 | 10 | 136 | 44 | 113 |
| 0310993L_SLU01 | 13 | 0 | 13 | 107 | 0.2 | 90 | 25 | 112 | 26 | 82 | 0.07 | 61 | 22 | 53 | 36 | 42 | 84 | 13 |
| TM1 | 13 | 0 | 17 | 20 | 0.21 | 13 | 26 | 45 | 37 | 60 | 0.08 | 20 | 18 | 62 | 38 | 26 | 96 | 50 |
| FGU11 | 13 | 8 | 17 | 93 | 0.5 | 138 | **38** | 14 | 26 | 59 | 0.1 | 117 | 32 | 13 | 34 | 112 | 92 | 104 |
| FGU03 | 13 | 37 | 24 | 14 | 0.6 | 27 | 37 | 17 | 62 | 66 | **0.5** | 38 | **38** | 16 | **200** | 13 | **180** | 18 |
| 0311328A_SLE02 | 14 | 0 | 16 | 32 | 0.12 | 131 | 29 | 28 | 32 | 75 | 0.05 | 102 | 27 | 14 | 21 | 66 | 73 | 72 |
| 0311123C_SLU01 | 14 | 0 | 14 | 43 | 0.2 | 112 | 22 | 113 | 29 | 38 | 0.07 | 127 | 24 | 17 | 25 | 83 | 82 | 14 |
| 0310231H_SLU01 | 14 | 0 | 14 | 98 | 0.2 | 14 | 29 | 18 | 32 | 17 | 0.09 | 126 | 23 | 31 | 49 | 82 | 92 | 44 |
| 0312356TSLE03 | 15 | 0 | 18.4 | 15 | 0.4 | 49 | 20.5 | 110 | 26.8 | 108 | 0.11 | 74 | 19.7 | 127 | 43.2 | 95 | 65 | 111 |
| 0310056T_SLE04 | 15 | 0 | 12 | 110 | 0.1 | 87 | 29 | 15 | 24 | 24 | 0.05 | 134 | 20 | 75 | 31 | 80 | 100 | 63 |
| 0310244X_SLE05 | 15 | 0 | 16 | 26 | 0.1 | 84 | 28 | 21 | 47 | 15 | 0.17 | 33 | 21 | 49 | 43 | 41 | 100 | 66 |
| FGU05 | 15 | 4 | 11 | 123 | 0.5 | 136 | 17 | 121 | **94** | 29 | **0.2** | 15 | 13 | 28 | **100** | 20 | 86 | 128 |
| 0311479P-0311479P_P_SLE05 | 16 | 6 | 13 | 70 | 0.5 | 16 | 17 | 74 | 27 | 105 | **0.3** | 53 | 19 | 84 | 58 | 91 | 91 | 16 |
| 0310056T_SLE03 | 16 | 11 | 8.4 | 114 | 0.1 | 92 | 15 | 64 | 49 | 16 | **0.38** | 47 | 12 | 74 | 75 | 24 | 96 | 60 |
| FGU09 | 16 | 19 | 25 | 16 | 0.5 | 34 | 36 | 26 | 59 | 56 | 0.1 | 115 | **42** | 19 | 46 | 64 | 120 | 42 |
| 0311503R_P_SLU01 | 16 | 20 | 10 | 118 | 0.32 | 26 | 20 | 68 | 50 | 40 | **0.68** | 16 | 16 | 79 | **120** | 30 | 120 | 21 |
| TM2 | 17 | 17 | 16 | 17 | 0.35 | 56 | 27 | 32 | 38 | 57 | **0.25** | 26 | 21 | 37 | 48 | 31 | 110 | 28 |
| 0310244X_SLE04 | 18 | 0 | 18 | 18 | 0.11 | 96 | 22 | 67 | 34 | 20 | 0.1 | 39 | 22 | 47 | 78 | 23 | 98 | 65 |
| FGU-TLS-6 | 18 | 0 | 21 | 53 | 0.5 | 130 | 35 | 20 | 29 | 89 | 0.1 | 107 | 30 | 18 | 26 | 108 | 99 | 91 |
| 0311479P-0311479P_P_SLE04 | 18 | 2 | 10 | 100 | 0.5 | 18 | 15 | 73 | 42 | 101 | **0.7** | 21 | 14 | 83 | **110** | 32 | 78 | 54 |
| 0311202N_SLU01 | 18 | 35 | 16 | 95 | 0.5 | 19 | 25 | 70 | 57 | 42 | **0.6** | 18 | 22 | 54 | **140** | 28 | **160** | 20 |
| 0310047H_SLU01 | 19 | 0 | 5.8 | 94 | 0.22 | 83 | 14 | 62 | 37 | 19 | 0.11 | 37 | 11 | 71 | 47 | 78 | 71 | 58 |
| FGU-TLS-35 | 19 | 23 | 17 | 19 | 0.5 | 105 | 28 | 50 | 37 | 109 | **0.3** | 27 | 24 | 38 | 62 | 106 | 140 | 19 |
| FGU02 | 19 | 25 | 17 | 59 | 0.6 | 61 | 31 | 72 | **78** | 43 | **2.1** | 24 | 24 | 61 | **94** | 19 | 140 | 62 |
| FGU-TLS-4 | 20 | 0 | 13 | 84 | 0.5 | 20 | 17 | 27 | 17 | 133 | 0.1 | 105 | 14 | 23 | 19 | 129 | 52 | 39 |
| 310795000_SLU01 | 20 | 29 | 15 | 101 | 0.26 | 37 | 27 | 60 | 63 | 25 | **0.22** | 62 | 27 | 20 | **85** | 40 | **150** | 43 |
| TM5 | 21 | 0 | 15 | 21 | 0.23 | 107 | 27 | 34 | 25 | 72 | 0.11 | 35 | 20 | 66 | 30 | 36 | 100 | 55 |
| FGU-TLS-37 | 21 | 1 | 10 | 46 | 0.5 | 128 | 20 | 114 | 24 | 93 | **0.3** | 36 | 15 | 21 | **190** | 27 | 77 | 88 |
| 0311171E_SLE04 | 22 | 0 | 7.6 | 119 | 0.22 | 57 | 12 | 97 | 25 | 28 | 0.09 | 22 | 12 | 81 | 23 | 43 | 69 | 82 |
| T15-31031-ST3 | 22 | 0 | 17 | 22 | 0.5 | 41 | 30 | 61 | 22 | 136 | 0.1 | 112 | 25 | 50 | 28 | 60 | 77 | 126 |





Funded by the European Union

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FGU17 | 22 | 0 | 19 | 87 | 0.5 | 22 | 33 | 53 | 48 | 37 | 0.1 | 123 | 28 | 48 | 66 | 55 | 110 | 56 |
| FGU15 | 22 | 0 | 16 | 64 | 0.5 | 118 | 25 | 115 | 36 | 124 | 0.1 | 121 | 24 | 67 | 60 | 115 | **160** | 22 |
| FGU18 | 22 | 13 | 11 | 133 | 0.5 | 67 | 18 | 122 | **75** | 31 | **0.5** | 40 | 15 | 138 | **110** | 22 | 100 | 135 |
| FGU06 | 22 | 26 | 21 | 55 | 0.5 | 137 | 33 | 22 | 46 | 107 | 0.1 | 113 | 32 | 51 | 52 | 110 | **150** | 64 |
| FGU-TLS-5 | 23 | 0 | 7 | 136 | 0.5 | 23 | 14 | 132 | 11 | 137 | 0.1 | 106 | 13 | 134 | 21 | 130 | 49 | 120 |
| FGU19 | 23 | 0 | 17 | 99 | 0.5 | 68 | 28 | 23 | 22 | 119 | 0.1 | 124 | 25 | 56 | 23 | 117 | 82 | 107 |
| TM6 | 23 | 0 | 15 | 23 | 0.25 | 98 | 25 | 59 | 23 | 84 | 0.05 | 99 | 20 | 69 | 24 | 38 | 110 | 34 |
| 0311631E_SLE04 | 23 | 21 | 12 | 120 | 0.37 | 25 | 21 | 76 | 54 | 45 | **0.71** | 23 | 17 | 87 | **130** | 29 | 120 | 26 |
| 0310992K_SLU01 | 24 | 0 | 16 | 54 | 0.2 | 127 | 29 | 24 | 30 | 46 | 0.16 | 52 | 23 | 33 | 74 | 49 | 100 | 47 |
| FGU-TLS-15 | 25 | 0 | 12 | 90 | 0.5 | 78 | 21 | 124 | 15 | 53 | 0.1 | 81 | 16 | 129 | 18 | 120 | 60 | 25 |
| 0311219G_SLU02 | 26 | 0 | 13 | 109 | 0.2 | 59 | 22 | 71 | 36 | 26 | 0.17 | 34 | 20 | 82 | 37 | 85 | 81 | 69 |
| FGU-TLS-10 | 27 | 0 | 15 | 42 | 0.5 | 75 | 28 | 37 | 21 | 50 | 0.1 | 78 | 24 | 27 | 23 | 46 | 76 | 77 |
| T14-31045-SC2 | 27 | 0 | 21 | 27 | 0.5 | 134 | 25 | 106 | 34 | 70 | 0.1 | 111 | 23 | 58 | 73 | 53 | 140 | 41 |
| Z4T1 | 28 | 0 | 10 | 73 | 0.2 | 28 | 25 | 41 | 13 | 65 | 0.05 | 89 | 20 | 99 | 25 | 73 | 51 | 109 |
| FGU-TLS-18 | 28 | 0 | 19 | 28 | 0.5 | 51 | 28 | 40 | 19 | 125 | 0.1 | 83 | 22 | 68 | 23 | 50 | 69 | 127 |
| TM4 | 28 | 15 | 12 | 68 | 0.25 | 95 | 24 | 54 | 26 | 68 | **0.25** | 28 | 20 | 63 | 34 | 33 | 110 | 31 |
| 0311316MSLE02 | 29 | 0 | 6.52 | 36 | 0.4 | 36 | 11.3 | 29 | 24.2 | 41 | 0.1 | 67 | 8.05 | 29 | 49.3 | 86 | 61.2 | 121 |
| FGU04 | 29 | 0 | 15 | 117 | 0.5 | 63 | 24 | 111 | 25 | 92 | 0.1 | 58 | 21 | 122 | 31 | 61 | 83 | 29 |
| TM3 | 29 | 16 | 14 | 52 | 0.28 | 94 | 22 | 58 | 34 | 61 | **0.43** | 29 | 18 | 65 | 46 | 35 | 110 | 33 |
| FGU-TLS-24 | 30 | 0 | 18 | 34 | 0.5 | 122 | 31 | 33 | 54 | 58 | 0.1 | 46 | 27 | 30 | 22 | 70 | 85 | 81 |
| FGU12 | 30 | 0 | 18 | 30 | 0.5 | 115 | 28 | 66 | 44 | 79 | 0.1 | 118 | 28 | 88 | 42 | 113 | 120 | 71 |
| FGU16 | 30 | 0 | 22 | 57 | 0.5 | 54 | 34 | 30 | 42 | 85 | 0.1 | 122 | 29 | 39 | 41 | 116 | 130 | 52 |
| 0311316MSLE03 | 31 | 0 | 1 | 37 | 0.4 | 40 | 11.5 | 31 | 21.7 | 52 | 0.1 | 68 | 9.03 | 32 | 64.4 | 87 | 65.3 | 123 |
| FGU08 | 31 | 0 | 14 | 83 | 0.5 | 31 | 21 | 136 | 22 | 114 | 0.1 | 60 | 20 | 136 | 41 | 63 | 81 | 101 |
| FGU-TLS-31 | 32 | 0 | 10 | 131 | 0.5 | 102 | 21 | 128 | 15 | 54 | 0.1 | 98 | 16 | 112 | 23 | 72 | 54 | 32 |
| 0311719A_SLU01 | 32 | 0 | 7 | 108 | 0.5 | 44 | 16 | 77 | 19 | 32 | 0.1 | 41 | 12 | 89 | 25 | 92 | 96 | 74 |
| 0310171T_0310431A_SLU02 | 33 | 5 | 13 | 63 | 0.42 | 60 | 23 | 105 | 33 | 33 | **0.3** | 49 | 20 | 117 | 37 | 39 | 91 | 36 |
| 0311316MSLE06 | 34 | 0 | 1 | 40 | 0.4 | 66 | 11.5 | 35 | 21.5 | 112 | 0.1 | 70 | 9.24 | 34 | 36.8 | 89 | 56.9 | 99 |
| 310795000_SLU02 | 34 | 0 | 13 | 115 | 0.2 | 71 | 24 | 81 | 21 | 34 | 0.05 | 136 | 21 | 97 | 31 | 97 | 110 | 53 |
| 0311631E_SLE03 | 34 | 12 | 7.1 | 128 | 0.28 | 52 | 15 | 75 | 38 | 80 | **0.4** | 55 | 11 | 86 | **110** | 34 | 96 | 57 |
| 0312356TSLE04 | 35 | 0 | 10.2 | 127 | 0.4 | 119 | 12.5 | 38 | 12.6 | 35 | 0.11 | 75 | 10.6 | 132 | 42 | 96 | 42.3 | 133 |
| FGU14 | 35 | 0 | 21 | 35 | 0.5 | 117 | 27 | 116 | 22 | 87 | 0.1 | 120 | 28 | 90 | 24 | 114 | 70 | 119 |
| FGU-TLS-8 | 37 | 0 | 13 | 126 | 0.5 | 132 | 21 | 134 | 18 | 106 | 0.1 | 109 | 18 | 135 | 21 | 131 | 66 | 37 |
| FGU10 | 38 | 0 | 13 | 92 | 0.5 | 38 | 23 | 137 | 44 | 110 | 0.1 | 116 | 23 | 125 | 44 | 111 | 130 | 70 |
| 0311316MSLE05 | 39 | 0 | 6.71 | 39 | 0.4 | 114 | 21 | 118 | 48.5 | 90 | 0.1 | 69 | 23.8 | 94 | 49.6 | 88 | 101 | 61 |
| FGU01 | 41 | 24 | 20 | 49 | 0.5 | 135 | 31 | 43 | 35 | 73 | **0.3** | 65 | 27 | 41 | **85** | 44 | 140 | 46 |
| 0311964SSLE04 | 44 | 0 | 1 | 44 | 0.4 | 48 | 16.9 | 119 | 17.2 | 115 | 0.1 | 71 | 17.2 | 126 | 20.5 | 67 | 56.2 | 125 |
| SLU_0311008C_01 | 45 | 0 | 5.9 | 138 | 0.1 | 111 | 13 | 57 | 7.5 | 97 | 0.05 | 138 | 9 | 52 | 12 | 134 | 32 | 45 |
| SLU_0310950P_01 | 45 | 0 | 7.6 | 137 | 0.11 | 45 | 20 | 102 | 9.8 | 47 | 0.05 | 137 | 11 | 113 | 17 | 133 | 38 | 108 |





Funded by the European Union

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0311964SSLE05 | 45 | 0 | 1 | 130 | 0.4 | 69 | 17.5 | 80 | 18 | 94 | 0.1 | 72 | 17.3 | 116 | 27 | 45 | 62.1 | 86 |
| FGU-TLS-22 | 45 | 0 | 11 | 125 | 0.5 | 89 | 22 | 84 | 51 | 55 | 0.1 | 45 | 20 | 103 | 26 | 52 | 79 | 79 |
| FGU-TLS-7 | 45 | 0 | 19 | 45 | 0.5 | 58 | 30 | 56 | 22 | 134 | 0.1 | 108 | 24 | 55 | 29 | 109 | 85 | 92 |
| Z5T2 | 46 | 0 | 8.1 | 78 | 0.2 | 46 | 19 | 48 | 12 | 78 | 0.05 | 97 | 16 | 109 | 10 | 137 | 41 | 115 |
| FGU-TLS-30 | 47 | 0 | 5 | 134 | 0.5 | 123 | 10 | 47 | 10 | 95 | 0.1 | 96 | 8 | 130 | 13 | 126 | 34 | 112 |
| 0312178Z_0310227D_SLU02 | 47 | 0 | 8 | 47 | 0.2 | 133 | 14 | 101 | 22 | 120 | 0.18 | 59 | 11 | 119 | 37 | 94 | 65 | 89 |
| FGU-TLS-26 | 48 | 0 | 15 | 50 | 0.5 | 93 | 26 | 107 | 23 | 104 | 0.1 | 48 | 22 | 120 | 21 | 124 | 60 | 132 |
| FGU-TLS-20 | 48 | 0 | 16 | 48 | 0.5 | 86 | 24 | 125 | 20 | 63 | 0.1 | 85 | 24 | 101 | 20 | 122 | 68 | 95 |
| 0311503R_P_SLU02 | 48 | 0 | 6.2 | 102 | 0.2 | 74 | 14 | 69 | 38 | 98 | 0.07 | 128 | 11 | 80 | 37 | 84 | 100 | 48 |
| FGU07 | 49 | 0 | 21 | 56 | 0.5 | 64 | 33 | 49 | 34 | 121 | 0.1 | 114 | 28 | 85 | 51 | 62 | 83 | 130 |
| 0312004KSLE05 | 49 | 0 | 13.1 | 61 | 0.4 | 70 | 19.9 | 120 | 20.6 | 117 | 0.1 | 73 | 18.7 | 93 | 28.2 | 93 | 128 | 49 |
| FGU-TLS-11 | 50 | 0 | 5 | 51 | 0.5 | 50 | 10 | 130 | 9 | 131 | 0.1 | 79 | 9 | 121 | 10 | 138 | 30 | 136 |
| FGU-TLS-27 | 50 | 0 | 8 | 111 | 0.5 | 55 | 17 | 127 | 15 | 51 | 0.1 | 50 | 15 | 72 | 29 | 54 | 55 | 97 |
| FGU-TLS-32 | 51 | 0 | 12 | 97 | 0.5 | 104 | 23 | 93 | 19 | 129 | 0.1 | 51 | 22 | 123 | 17 | 127 | 70 | 137 |
| 0311328A_SLE03 | 51 | 0 | 14 | 104 | 0.12 | 106 | 26 | 63 | 38 | 77 | 0.06 | 135 | 22 | 57 | 30 | 90 | 72 | 51 |
| FGU-TLS-25 | 53 | 0 | 14 | 67 | 0.5 | 53 | 25 | 86 | 42 | 69 | 0.1 | 88 | 23 | 106 | 29 | 104 | 98 | 83 |
| T14-31046-S1 | 54 | 0 | 12 | 103 | 0.5 | 109 | 16 | 108 | 31 | 91 | 0.1 | 54 | 16 | 118 | 42 | 59 | 84 | 94 |
| T14-31047-SC3 | 56 | 0 | 14 | 81 | 0.5 | 110 | 23 | 135 | 27 | 99 | 0.1 | 56 | 29 | 76 | 14 | 135 | 75 | 96 |
| FGU-TLS-39 | 57 | 0 | 13 | 122 | 0.5 | 129 | 22 | 98 | 25 | 88 | 0.1 | 104 | 20 | 111 | 35 | 57 | 93 | 68 |
| SLU_310785878_01 | 58 | 0 | 8.8 | 135 | 0.18 | 113 | 23 | 104 | 19 | 100 | 0.15 | 64 | 14 | 115 | 40 | 58 | 75 | 59 |
| FGU-TLS-21 | 62 | 0 | 15 | 62 | 0.5 | 88 | 22 | 126 | 16 | 126 | 0.1 | 86 | 17 | 100 | 18 | 123 | 62 | 131 |
| FGU-TLS-0 | 64 | 0 | 11 | 86 | 0.5 | 73 | 23 | 82 | 22 | 123 | 0.1 | 76 | 20 | 64 | 40 | 98 | 84 | 76 |
| FGU13 | 65 | 0 | 12 | 112 | 0.5 | 116 | 18 | 138 | 16 | 138 | 0.1 | 119 | 18 | 137 | 31 | 65 | 53 | 116 |
| FGU-TLS-9 | 65 | 0 | 17 | 85 | 0.5 | 108 | 27 | 65 | 18 | 135 | 0.1 | 110 | 21 | 78 | 18 | 132 | 68 | 106 |
| FGU-TLS-23 | 66 | 0 | 14 | 66 | 0.5 | 91 | 23 | 85 | 25 | 122 | 0.1 | 87 | 20 | 105 | 28 | 103 | 90 | 122 |
| FGU-TLS-29 | 71 | 0 | 13 | 82 | 0.5 | 97 | 25 | 88 | 19 | 127 | 0.1 | 92 | 20 | 73 | 16 | 71 | 63 | 134 |
| FGU-TLS-3 | 71 | 0 | 14 | 71 | 0.5 | 100 | 25 | 91 | 18 | 128 | 0.1 | 94 | 20 | 108 | 20 | 125 | 72 | 124 |
| FGU-TLS-2 | 79 | 0 | 13 | 79 | 0.5 | 85 | 19 | 83 | 26 | 118 | 0.1 | 84 | 18 | 98 | 24 | 102 | 67 | 129 |
| FGU-TLS-33 | 100 | 0 | 9 | 132 | 0.5 | 124 | 19 | 129 | 15 | 130 | 0.1 | 100 | 16 | 131 | 17 | 128 | 56 | 102 |

*Note: bolded and grayed samples indicate an anomalous samples*





Funded by the European Union

Lastly, it is possible to display these anomalies on a map (Figure 29) and remove them from the set to gain access to the diffuse part of the measured contents.

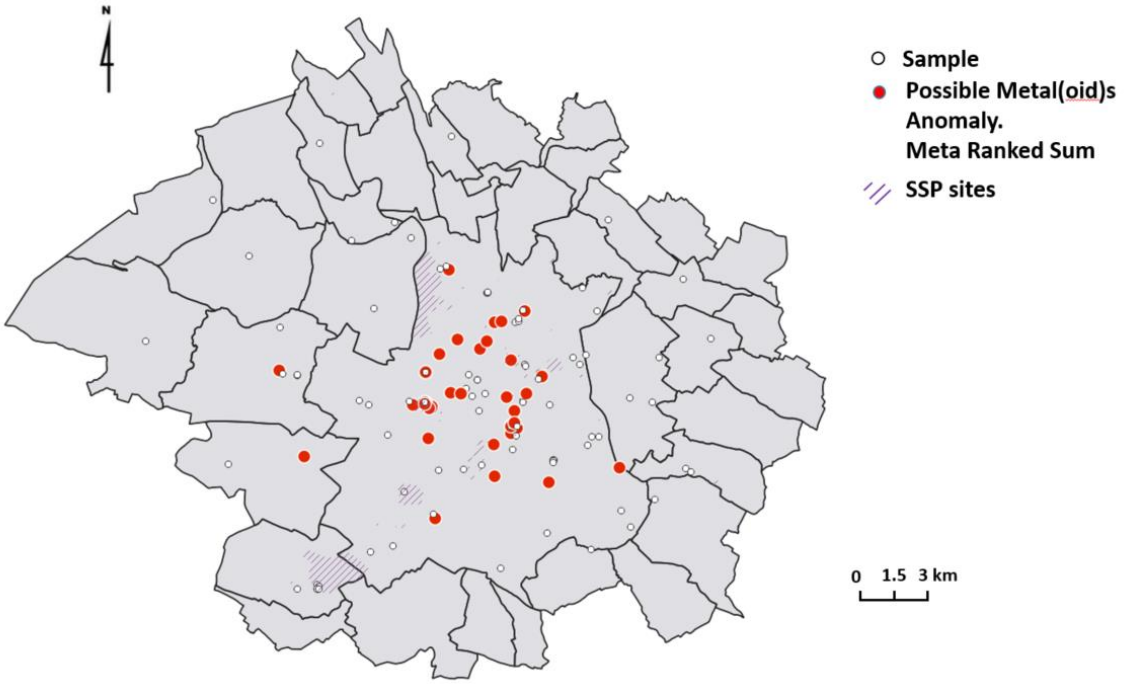


**Figure 29: Metal(oid) anomalies in ITA3, 138 surface samples, as pinpointed by the ISLANDR detection algorithm**





Funded by the European Union

# 4. From Hot spot identification to diffuse contamination map

If we look at how geochemists interpret data (Salminen, 2005; Demetriades, 2018), we can see in Figure 30 that we need to separate the baseline—the part that is the sum of a background and diffuse urban pollution—from anomalies and the part that comes from point sources.

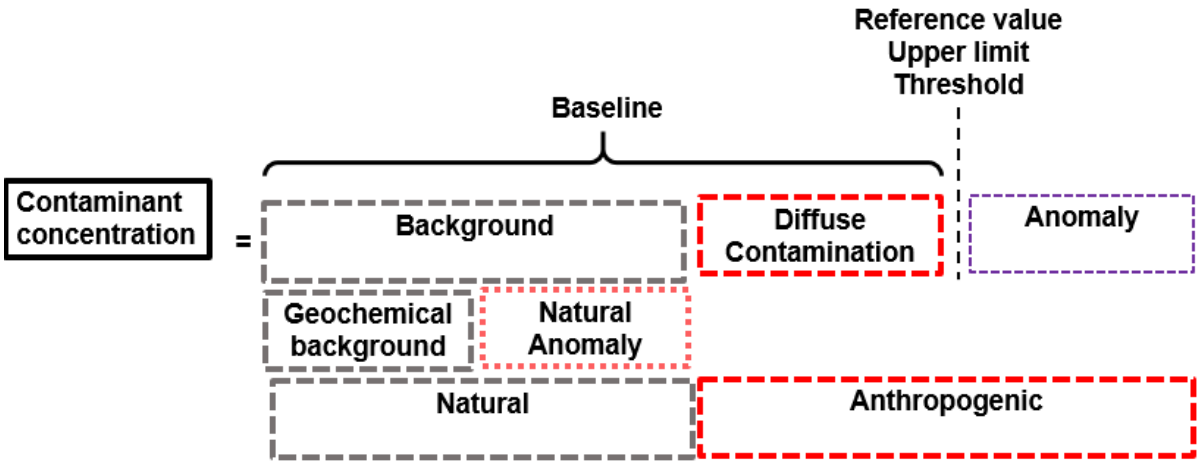


**Figure 30: Conceptual models of urban soil contaminant concentrations as used by Salminen (2005) and Demetriades (2018)**

This is an intrinsically additive model. The approach adopted for ISLANDR involves separating anomalies from the contaminant concentration to reach the baseline, which includes the targeted diffuse contamination. If the natural background is also available (either through a specific campaign or by typological sorting of available measurements), it will then be possible to filter out the diffuse contamination (Figure 31).





Funded by the European Union

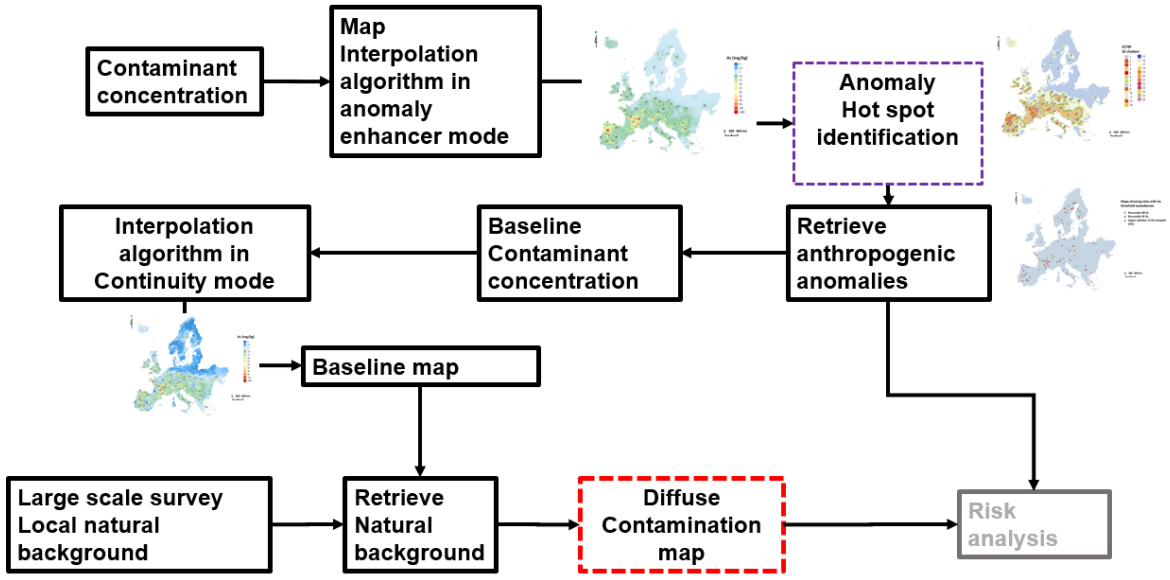


**Figure 31: Objectives of this WP: Algorithm for hot spot identification and interpolation of any data given for risk analysis purpose.**

# Conclusion

Our research has led us to propose an innovative anomaly detection algorithm especially adapted to cases where data is sparse (< 30), clustered and uncertain. Unlike detections made on kriged maps and other deep learner algorithms, our algorithm has a higher detection rate because it has his own interpolator without smoothing behavior. Our algorithm can work with small sets as less than 10 data and is now ready to process any ISLANDR ITA or Large Scale European survey as LUCAS.





Funded by the European Union

# Bibliography

**Agterberg, F. P. (2012).** Sampling and analysis of chemical element concentration distribution in rock units and orebodies. Nonlinear Processes in Geophysics, 19(1), 23–44. https://doi.org/10.5194/npg-19-23-2012

**Aitchison, J. (1986).** The statistical analysis of compositional data (1. publ). Chapman and Hall.

**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). Journal of Geochemical Exploration, 93(1), 21–34. https://doi.org/10.1016/j.gexplo.2006.07.006

**Anselin, L. (1995).** Local Indicators of Spatial Association—LISA. Geographical Analysis, 27(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

**Auffret, A. G., Kimberley, A., Plue, J., Skånes, H., Jakobsson, S., Waldén, E., Wennbom, M., Wood, H., Bullock, J. M., Cousins, S. A. O., Gartz, M., Hooftman, D. A. P., & Tränk, L. (2017).** HistMapR: Rapid digitization of historical land-use maps in R. Methods in Ecology and Evolution, 8(11), 1453–1457. https://doi.org/10.1111/2041-210X.12788

**Belbeze, S., Assy, Y., Le Cointe, P., & Rame, E. (2022).** CAPacité d'Infiltration des eaux pluviales du territoire de TOULouse Métropole (CAPITOUL) (Rapport Final Nos. RP71904-FR; p. 72). BRGM.

**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; p. 347). BRGM.

**Bouchon-Meunier, B., & Marsala, C. (2003).** Logique floue, principes, aide à la décision. Hermès Science publications.

**Cai, W., Chen, S., & Zhang, D. (2007).** Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. Pattern Recognition, 40(3), 825–838. https://doi.org/10.1016/j.patcog.2006.07.011

**Caritat, P. de, & Cooper, M. (with National Geochemical Survey of Australia). (2011).** The geochemical atlas of Australia. Geoscience Australia.

**Carranza, E. J. M. (2009).** Geochemical anomaly and mineral prospectivity mapping in GIS (1st ed). Elsevier.

**Ceballos, B., Lamata, M. T., & Pelta, D. A. (2016).** A comparative analysis of multi-criteria decision-making methods. Progress in Artificial Intelligence, 5(4), 315–322. https://doi.org/10.1007/s13748-016-0093-1





Funded by the European Union

**Ceballos, B., Pelta, D. A., & Lamata, M. T. (2018).** Rank Reversal and the VIKOR Method: An Empirical Evaluation. International Journal of Information Technology & Decision Making, 17(02), 513–525. https://doi.org/10.1142/S0219622017500237

**Cheng, Q. (1999a).** Markov Processes and Discrete Multifractals. Mathematical Geology, 31(4), 455–469. https://doi.org/10.1023/A:1007594709250

**Cheng, Q. (1999b).** Multifractality and spatial statistics. Computers & Geosciences, 25(9), 949–961. https://doi.org/10.1016/S0098-3004(99)00060-6

**Cheng, Q. (2007a).** Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geology Reviews, 32(1–2), 314–324. https://doi.org/10.1016/j.oregeorev.2006.10.002

**Cheng, Q. (2007b).** Multifractal imaging filtering and decomposition methods in space, Fourier frequency, and eigen domains. Nonlinear Processes in Geophysics, 14(3), 293–303. https://doi.org/10.5194/npg-14-293-2007

**Cheng, Q. (2014a).** Generalized binomial multiplicative cascade processes and asymmetrical multifractal distributions. Nonlinear Processes in Geophysics, 21(2), 477–487. https://doi.org/10.5194/npg-21-477-2014

**Cheng, Q. (2014b).** Vertical distribution of elements in regolith over mineral deposits and implications for mapping geochemical weak anomalies in covered areas. Geochemistry: Exploration, Environment, Analysis, 14(3), 277–289. https://doi.org/10.1144/geochem2012-174

**Cicchella, D., De Vivo, B., Lima, A., Albanese, S., & Fedele, L. (2008).** Urban geochemical mapping in the Campania region (Italy). Geochemistry: Exploration, Environment, Analysis, 8(1), 19–29. https://doi.org/10.1144/1467-7873/07-147

**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). Environmental Earth Sciences, 75(19), 1325. https://doi.org/10.1007/s12665-016-6093-4

**Frontasyeva, M., Harmens, H., & Uzhinskiy, A. (2020).** Mosses as biomonitors of air pollution: 2015/2016 survey on heavy metals, nitrogen and POPs in Europe and beyond. (Report of the ICP Vegetation No. ISBN: 978-5-9530-0508-1; p. 136). Joint Institute for Nuclear Research. http://www1.jinr.ru/Books/REPORT-Frontasyeva_sait.pdf

**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





Funded by the European Union

**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, p. 26). Environment Centre Wales. https://icpvegetation.ceh.ac.uk/sites/default/files/Mosses%20as%20biomonitors%20of%20atmospheric%20POPs%20pollution.pdf

**Koenderink, J. (2021).** The structure of images: 1984–2021. Biological Cybernetics, 115(2), 117–120. https://doi.org/10.1007/s00422-021-00870-0

**Lindeberg, T. (1994).** Scale-Space Theory in Computer Vision. Springer. https://doi.org/10.1007/978-1-4757-6465-9

**Pawlowsky-Glahn, V., & Buccianti, A. (Eds.). (2011).** Compositional data analysis: Theory and applications. Wiley.

**Pawlowsky-Glahn, V. (with Egozcue, J. J., & Tolosana-Delgado, R.). (2015).** Modeling and analysis of compositional data. Wiley.

**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

**Pitard, F. F. (2019).** Sampling theory and sampling practice: Heterogeneity, sampling correctness, and statistical process control (Third edition). Taylor and Francis.

**Reimann, C., Birke, M., & Demetriades, A. (2014).** Chemistry of Europe's agricultural soils: Part A : Methodology and interpretation of the GEMAS data set. Bundesanstalt für Geowissenschaftem und Rohstoffe.

**Reimann, C., Birke, M., Filzmoser, P., & O'Connor, P. (2014).** Chemistry of Europe's Agricultural Soils, Part B : General Background Information and Further Analysis of the GEMAS Data Set. Bundesanstalt für Geowissenschaften und Rohstoffe.

**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-lan, A., Cicchella, D., De Vivo, B., De Vos, W., Dinelli, E., Ďuriš, M., … Sadeghi, M. (2018).** GEMAS: Establishing geochemical background and threshold for 53 chemical elements in European agricultural soil. Applied Geochemistry, 88, 302–318. https://doi.org/10.1016/j.apgeochem.2017.01.021

**Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., & Ladenberger, A. (2012).** The concept of compositional data analysis in practice—Total major element concentrations in agricultural and grazing land soils of Europe. Science of The Total Environment, 426, 196–210. https://doi.org/10.1016/j.scitotenv.2012.02.032





Funded by the European Union

**Targa, J., Ripoll, A., Banyuls, L., Gonzalez Ortiz,A, & Soares, J. (2023).** Status report of air quality in Europe for year 2021, using validated data (No. ETC-HE Report 2023/1). ETC-HE. https://www.eionet.europa.eu/etcs/all-etc-reports

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

**Tukey, John W. (1977).** Exploratory Data Analysis. Repr. Reading, Mass.: Addison-Wesley.

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

**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. Computers & Geosciences, 90, 189–201. https://doi.org/10.1016/j.cageo.2015.11.001

**Zhao, F., Jiao, L., & Liu, H. (2013).** Kernel generalized fuzzy c-means clustering with spatial information for image segmentation. Digital Signal Processing, 23(1), 184–199. https://doi.org/10.1016/j.dsp.2012.09.016





Funded by the European Union

# Appendix 1. Interpolation Algorithm

## Abbreviations and acronyms

| Acronym | Description |
|---|---|
| WP | Work Package |
| C&D | Communication & Dissemination |
| A | Anomaly |
| ANN | Artificial neural network interpolation technique |
| ANR | French national research agency |
| AT | Averaging Time is used in health and safety risk assessment |
| BA | Bioaccessibility is used in health and safety risk assessment |
| BASIAS | Potentially polluted site for French authorities |
| BASOL | Known polluted or remediated site for French authorities |
| BRGM | French Geologic Survey |
| BW | Body Weight is used in health and safety risk assessment |
| C-A | Model of concentration variation by area |
| CAX | Learning set for the EEPH algorithm (C are observations, AX is the knowledge base of covariable of arbitrary dimension) |
| CDF | Cumulative distribution function |
| CLORPT | CLimate, Organisms, Relief, Parent material, and Time |
| CODA | COmpositional Data Analysis |
| CoOK | Co ordinari kriging is a multivariate geostatistical technique |
| cov.PI | Coverage of the 90% confidence interval is used in interpolation inter comparison test |
| CS | Concentration in Soil is used in health and safety risk assessment |
| D | A Dose as used in risk assessment |
| DUK | Detrended universal kriging is a geostatistical technique |
| ED | Exposure Duration is used in health and safety risk assessment |
| EDC | Empirical diffusion coefficient |
| EEPH | Enhanced Experimental Probabilistic Hypersurface as created for ISLANDR |
| EF | Exposure Frequency used in health and safety risk assessment |
| ELN | Elastic net regression is a non geostatistical interpolator technique |





Funded by the European Union

| Acronym | Description |
|---|---|
| EMEP | Co-operative programme for monitoring and evaluation of the long-range transmission of air pollutants in Europe |
| EML | Ensemble of machine learning models is an hybrid interpolator |
| EPH | Experimental Probabilistic Hypersurface |
| FMCDM | Fuzzy multi-criteria decision-making is a kind of classifier where uncertainty, vagueness, and/or imprecision are present in the decision matrix |
| FOREGS | The Forum of the European Geological Surveys Directors (FOREGS) was an informal group that provided the directors of geological surveys with a platform for exchanging ideas on the status of each national institute. FOREGS ceased its activities in September 2005, but was taken over by EuroGeoSurveys.FOREGS has carried out a Europe-wide Geochemical Baseline Mapping Programme. |
| FPGN | National French top soil geochemical Background study |
| GEMAS | GEochemical MApping of Agricultural Soil. The GEMAS project collected 2108 Ap horizon soil samples from regularly ploughed fields in 33 European countries, |
| GP | Gaussian processes are a generic non geostatistical interpolator technique |
| GPR | Gaussian process regression is a non geostatistical interpolator technique |
| GRNN | General Regression neural-networks is a non geostatistical interpolator technique |
| GSLIB | Geostatistical Software Library from Deutsch and Journel (1997) |
| GWR | Geographically weighted regression |
| HBKG | High background of geogenic origin used in fractal study |
| HH, LL, HL, LH | Spatial cluster type HH is a strong value in a strong neighborhood, LL is a weak value in a weak neighborhood, HL is a strong value within weak values and LH a weak value within strong values |
| HOUSES | Harmonized Operation of Uncertainties in Spatialized Environmental Systems. It is a French research project. |
| HUI | Historical urban Inventory is a comprehensive urban historical study |
| IDS | Inverse distance square weighting is a popular non geostatistical interpolator technique |
| IDW | Inverse distance weighting is a popular non geostatistical interpolator technique |
| IER | Individual Excess Risk are used in risk assessments |
| ILR | Isometric log ratio is a transformation used for multivariate data |





Funded by the European Union

| Acronym | Description |
|---|---|
| IRSN | French Nuclear & Safety agency |
| ISLANDR | Information-based Strategies for Land Remediation: Our project. |
| ITAs | ISLANDR Test Areas. |
| K | Simple Kriging is a geostatistical interpolator technique |
| KDE | Kernel density estimation is a non geostatistical interpolator technique |
| KED | Kriging with external drift is a geostatistical interpolator technique |
| LBKG | Low geogenic Background |
| LEPS | Linear error in probability space is a statistical measure |
| LOQ | Limit of Quantification of a measurment or a laboratory analysis |
| LANU | LAnd Use |
| LUCAS | Land use and land cover survey. The data collected by LUCAS provides harmonised and comparable statistics on land use and land cover across the whole of the EU's territory. |
| MAE | Mean absolute error used in interpolation inter comparison test |
| MAF | Maximum autocorrelation function is a multivariate and spatial dimension reduction technique |
| MAXE | Maximum error used in interpolation inter comparison test |
| MBS | Multilevel B-Spline is a non geostatistical technique |
| MBSDE | Multilevel B-splines with external drift is a hybrid interpolator technique |
| Mcov.PI | Mean absolute deviation of the accuracy plot is used in interpolation inter comparison test |
| MCRPS | Mean continuous ranked probability score is used in interpolation inter comparison test |
| MDS | Multidimentional scaling is a method of dimension reduction |
| ME | Mean error is used in interpolation inter comparison test |
| MIDW | Multifractal Inverse Distance Weighting interpolation is a non geostatistical interpolator technique |
| MISE | Mean integrated squared error (MISE) is used in interpolation inter comparison test |
| MLP | Multi layer preceptron a kind of artificial intelligence and is a non geostatistical interpolator technique |
| MSE | Mean squared error used in interpolation inter comparison test |
| MULTI-MOORA | The MULTIMOORA is a ranking obtained by aggregating the results of the ternary ranking methods Ratio System, |





Funded by the European Union

| Acronym | Description |
|---|---|
| | Reference Point Approach, and Full Multiplicative Form. It is a technique used in Multicriteria decision-making. |
| NN | The Nearest Neighbor algorithm (NN) os |
| NNW | Refers to neural networks in a general sense and are non geostatistical interpolator technique |
| OCS | Organic Carbon Survey in HOUSES Dataset |
| ODC | Optimal diffusion coefficient |
| OK | Ordinary kriging is a geostatistical interpolator technique |
| OLS | Ordinary least squares regression is a non geostatistical interpolator technique |
| PC | Principal component from a Principal Component Analysis |
| PCA | Principal Component Analysis is a statistical methode is a method for Dimensionality Reduction |
| PDS | Exceedance probability is the probability that a certain threshold will be exceeded (PDS for the french "Probabilitée de Dépassement de Deuil") |
| PN | Probability of necessity in Pearl causal inference theory |
| PNS | Probability of necessity and sufficiency in Pearl causal inference theory |
| PS | Probability of sufficiency in Pearl causal inference theory |
| PSO | Particle swarm optimization |
| Q90, Q98 | 90 percentile, 98 percentile |
| QRF | Quantile Regression Forest is a non geostatistical interpolator technique |
| R | R is a scientific and statistical programming language |
| RBF | Radial basis function is a non geostatistical interpolator technique |
| RMSE | Root mean squared error is used in interpolation inter comparison test |
| RTOP | Interpolation of Data with Variable Spatial Support is a geostatistical interpolator technique |
| SCORPAN | Soil, Climate, Organisms, Relief, Parent material, Age, N is for space, spatial or geographic position |
| SFCM | Spatial fuzzy C-means a clustering algorithm |
| SI | Soil Ingestion (kg/d) used in health and safety risk assessment |
| SIC | Sparse Imprecise Clustered datas |
| SIC2004 | The Spatial InterComparison test for emergency mapping of radiological incidents |
| SSP | Polluted Sites (From the french "Sites et Sols Pollués") |
| SVM | Support vector machine is a non geostatistical interpolator technique |





Funded by the European Union

| Acronym | Description |
|---|---|
| TFN | Triangular fuzzy numbers are used in Fuzzy logics mathematics |
| TGK | Trans Gaussian Kriging is a geostatistical interpolator technique |
| TIF | Tukey Inner Fence is statistical term |
| TIN-PPV | Nearest neighbour algorithm and triangulation interpolator is a non geostatistical interpolator |
| TOPSIS | A Technique for Order Preference by Similarity to Ideal Solution used in Multicriteria decision-making |
| TPH | Total Petroleum Hydrocarbon |
| T-SNE | T-distributed stochastic neighbor embedding is a method of dimension reduction |
| UER | Unit Excess Risk are used in risk assessments |
| UMAP | Uniform Manifold Approximation and Projection is a method for Dimensionality Reduction |
| VIKOR | VIKOR is a technique for Multicriteria Optimization and Compromise Solution used in Multicriteria decision-making. The name VIKOR is Serbian "VIseKriterijumska Optimizacija I Kompromisno Resenje" |
| w.PI | The width of the 90% confidence interval is used in interpolation inter comparison test |
| WASS | The WASSerstein distance is a statistical distance measure between observations |
| XGBTREES | A modified random forest algorithm is a non geostatistical interpolator technique |





Funded by the European Union

# Introduction

The approach adopted for ISLANDR involves separating anomalies from the contaminant concentration to reach the baseline, which includes the targeted diffuse contamination. To do so, the interpolation model selected should maintain and facilitate anomaly detection. To provide a solid basis for risk studies, it must be equipped with a highly advanced uncertainty management system to be compatible with the future risk calculation module, which at this stage of WP1report production is not yet defined. The algorithm is designed to spatially generate several probability densities that can then be converted into the uncertainty management format of other sets (Figure 32).

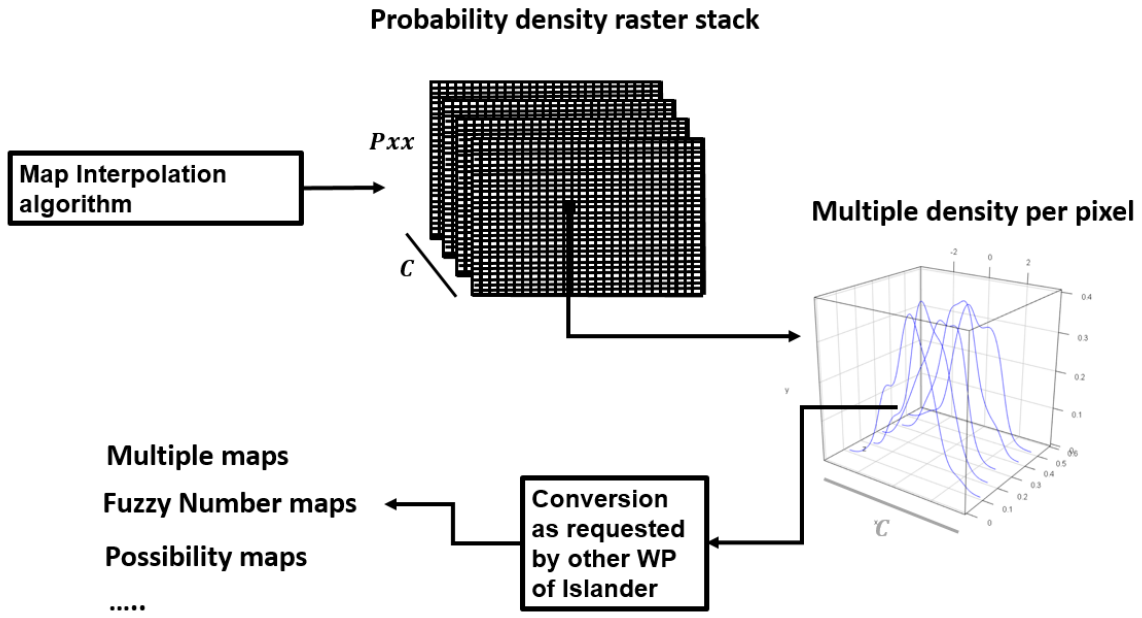


**Figure 32: Proposed uncertainty output of the interpolation model**

Finally, given that the interpolator will need to handle disparate data, various countermeasures will have to be implemented, in particular:

- Data clustering.
- Censoring of certain data (< LOQ).
- Multi-support (size) of samples.
- The addition of expert knowledge such as ranges and directions of phenomena, geology of the area, and propagation of uncertainties.

After a quick review of the interpolation techniques usually used for European mapping, this report will focus on the interest of diffusive phenomena, the resulting information diffusion mapping, and the algorithmic developments involved in obtaining a totally new and flexible interpolator capable of producing informative maps, despite their SIC status.





Funded by the European Union

# 1. Interpolating European data

Soil content maps are regularly generated for the European Community. The techniques used vary depending on the author and the project, for example, moving median (MM) (Salminen et al., 2005), kriging (K) (Tarvainen et al., 2013), multilevel B-splines (MBS) (Panagos et al., 2014), kriging with external drift (KED) (Toth et al., 2013, 2016; Heuvelink et al., 2016), geographically weighted regression (GWR) (Xu and Zhang, 2021), 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 is in no way an exhaustive review, but rather a few examples from selected well-known authors to illustrate the practice. For more information, a more comprehensive review of interpolation techniques used by geochemists to establish urban backgrounds can be found in Belbeze et al., 2023.





Funded by the European Union

**Table 16: Characteristics of key methods used to map European soils**

| Encountered approaches | Description | Example |
|---|---|---|
| No maps – dot plot 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 (Reiman, 2008; Rhind et al., 2013). | Reimann et al., 2018; Negrel 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. | Negrel et al., 2015 |
| The nearest neighbor algorithm (NN) and TIN-PPV triangulation | 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 are quick to perform for densely sampled areas. This type of interpolation is useful because it does not smooth out the data and allows for quick 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 corresponds to a vision of the physical phenomenon considered as consisting of linear trends to which a fluctuating error of small amplitude is added. | Jordan, 2018 |
| Simple kriging (K) | The kriging interpolation (Chiles and Delfiner, 2012) is similar in its general form to the IDW, but differs in the way the weights are calculated. Kriging uses an experimentally determined covariance function. Kriging assumes that the data is regionalized (strong hypothesis) 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., 2014 |
| 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 where the geostatistical methods of the French mining school are used: 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 and al., 2017) is a reference under R platform for the implementation of the KED method. | Salminen et al., 2005; Lado et al., 2008; Tóth et al., 2013; Tóth et al., 2016; Pereira et al., 2012; Ducarme et al., 2003; 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 et al., 2020). Some pending questions are possible regarding over- and underestimations induced by heterogeneous populations. | Van Eynde et al., 2023; Xiao et al., 2023; Hengl et al., 2021; Wadoux et al., 2020 |





Funded by the European Union

| | | |
|---|---|---|
| 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. 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, and 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 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. | Ballabio et al., 2024 |





Funded by the European Union

The process used for all these mapping techniques is as follows.

- Structural analysis:
  - Data verification, at which point certain data outliers are removed based on expert opinion.
  - The applicability of methodological choices (stationarity, support, additivity, accumulation) is verified, along with possible data transformation (log translation, anamorphosis, Box-Cox, etc.), and the potential construction of an external drift using various models, including machine-learning algorithms.
- **Spatial statistics calculation**, mainly variogram calculation.
- **Model calibration** on the variogram (K, KED, MBS, GWR). It should be noted that only certain functions with suitable derivation properties (exponential, spherical, Mattern) are used by the profession. The model can be calibrated automatically (using least square or maximum likelihood, for example), but the author controls and sets the shape beforehand.
- **Hyper-parameter calibration**: in the case of machine learning algorithms (QRF) or Ensemble Machine Learning models, it is necessary to calibrate the learning parameters through numerous runs or anneals (such as the number of random trees or resampling). ;
- **Estimating** or simulating content.

We also note that these models use correlations between low-density measured content and a densely sampled variable (like geology or usage) to increase the resolution of their output map (commonly known as drift). This is the CLORPT (Jenny, 1941) or SCORPAN (McBratney et al., 2003) concept. As a reminder:

- CLORPT = CLimate, Organisms, Relief, Parent material, and Time.
- SCORPAN = S (soil or measured attributes of the soil at a point), C (climate), O (organisms, including land cover and natural vegetation) R (relief; topography including terrain attributes and classes), P (parent material, including lithology), A (age, the time factor) N (space, spatial or geographic position).

The idea is to co-model the results of a small number of costly surveys with measurements of covariates (geology, land use, photographs, etc.) evenly distributed over the area.
Given these observed elements, the methods used suffer from uncertainties related to model calibration and, for those mentioned above, to the cascade of covariate scales, which can create a false reality. These are epistemic uncertainties. For more information on this topic, please refer to Loquin and Dubois (2010), who have detailed the epistemic uncertainty of kriging.

In short, these various model mappings are based on major epistemic choices:

- The phenomenon to be interpolated is continuous between two measurements.
- Data transformation does not lead to misinterpretation
- Experts choose two models:
  - One for the spatial response of the variable.





Funded by the European Union

- One for its links to the explanatory parameters (covariates), often with the assumption that the greater the number of covariates, the better the result.
- Values that are considered anomalous are removed from the calculation and the data may be transformed (log, normal score, anamorphosis).

When these expert models have sufficient data and are well calibrated, they are effective, validated, and published, but if the data is sparse and difficult (Sparse, Imprecise, Clustered – SIC), 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 significant and the spatial structuring is correctly modeled.

Naturally, for ISLANDR's ITAs if the data allows it, advanced geostatistical techniques will be applied. Otherwise, our SIC algorithm will be used, which enables us to cover all possible scenarios and, above all, the diversity of measurement sources that other methods cannot always take into account.

# 2. Mapping scale and anomaly

The levels measured on the ITAs are measured on a given scale. As the propagation of pollutants or reclamation is by nature discontinuous, it does not seem feasible in the SIC context to produce a fine-scale pollution map from coarse-scale sampled data. Only a transition from fine to coarse scale can be envisaged (Figure 33), using data blurring (Koenderink, 1984; Lindeberg, 1994).

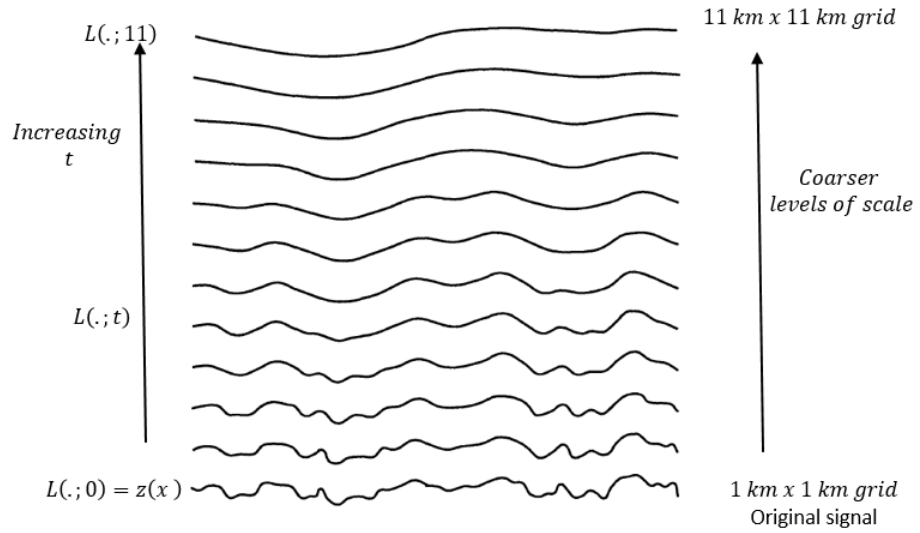


**Figure 33: From the original 1km² grid signal to that obtained by 11km² grid measurements, adapted from Lindeberg (1994).**





Funded by the European Union

Such a smoothing effect can be observed on multi-scale data from the Toulouse ITA (Figure 34 - Table 17) where we have geochemical background data on a large scale of 50km x 50km, FGPN data on a 15km x 15km grid, and urban background data on a 1km x 1km grid obtained through KED kriging and the new EEPH (Enhanced Experimental Probabilistic Hypersurface) technique, which will be described in the following paragraphs.

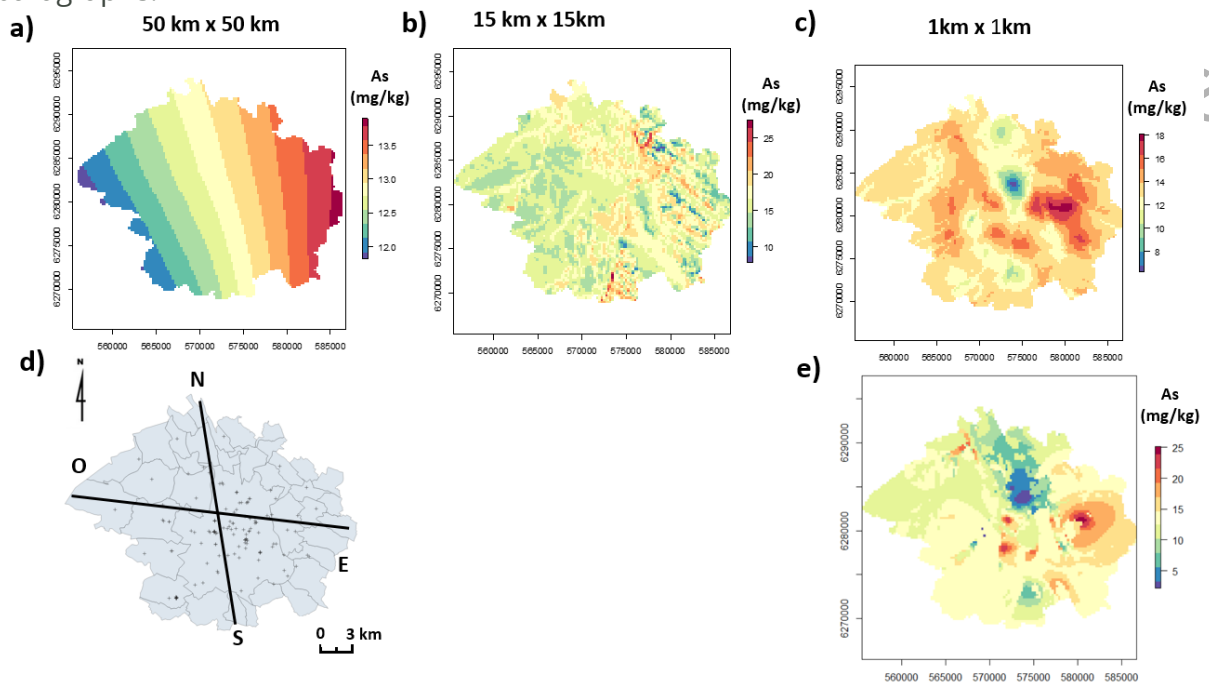


**Figure 34: Four maps of various As concentration surveys of surface soil resolutions in the Toulouse metropolitan area**

*with a) GEMAS Ap sample 50 km x 50 km (kriged and zoomed), b) French Natural Background 15 km x 15 km (KED), c) urban geochemical background (KED) 1 km x 1 km, d) cross-section localization, and e) urban geochemical baseline (EPH, Experimental Probabilistic Hypersurface) of 1 km x 1 km.*

Examining a north-south cross-section of these maps (Figure 35) using Lindberg's technique (1994) shows that the level of detail changes with resolution, through a form of complex, diffusive smoothing.





Funded by the European Union

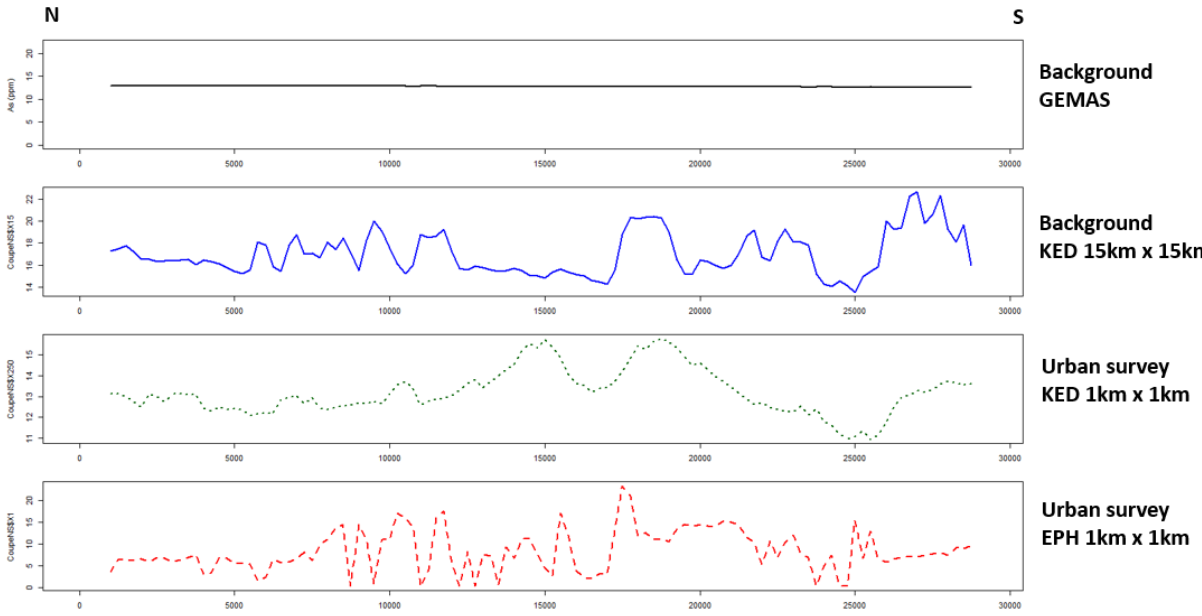


**Figure 35: North-–South cross-section of the above maps**

**Table 17: Main statistics of the above maps**

| | GEMAS 50 km x 50 km | FPGN 15 km x 15 km | EPH 1 km x 1 km | KED 1 km x 1 km |
|---|---|---|---|---|
| Purpose | Large-scale background tendency | State-level background tendency | Urban geochemical study | Urban geochemical study |
| Min | 11.8 | 7.7 | 2.1 | 6.3 |
| 1st quantile | 12.4 | 14.9 | 11.8 | 12.7 |
| Median | 12.8 | 16 | 13.3 | 13.5 |
| 3rd quantile | 13.2 | 17 | 14.1 | 14 |
| Max | 13.8 | 27.4 | 25 | 18 |

This paragraph is not intended to compare the merits of the various methods used to map content in the city of Toulouse. None of them can claim to represent the exact content at every point (Belbeze et al, 2023).

In practice, the phenomena studied are often characterized as non-stationary, non-isotropic, heterogeneous, which means that theoretical changes of scale through Green's function (i.e., diffusion) are only true for a restricted domain. This is why the author is proposing a local information diffusion algorithm.

Lastly, we can mention the field of artificial intelligence (Figure 36), where neural networks are trained to find diffusion coefficients that have altered a given image. The system is trained on millions of images, which it broadcasts and then reconstructs, providing the kind of high-performance multiscale artificial vision theorized by Koenderink (1984). However, when it comes to pollution, we never have the millions of data that would be needed.





Funded by the European Union

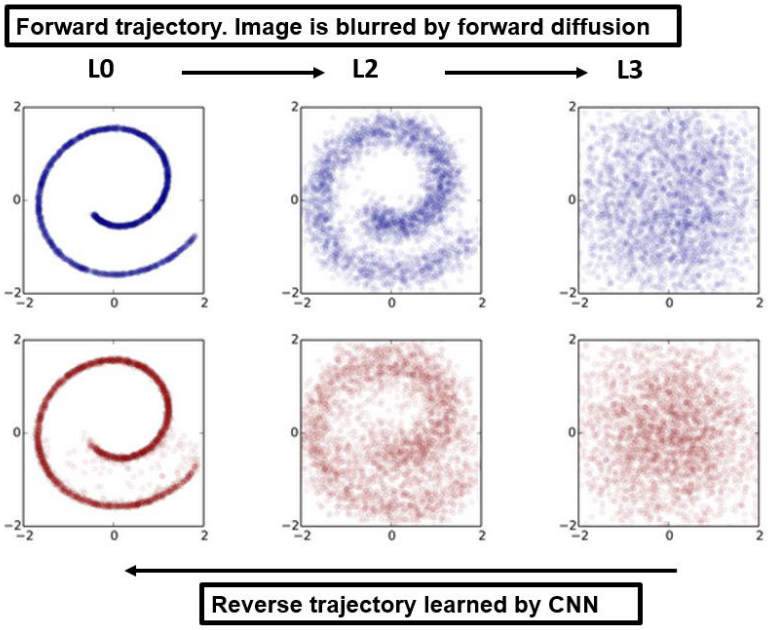


**Figure 36: An example of training a diffusion model for modeling 2D swiss roll data. (Sohl-Dickstein et al., 2015)**

## 2.1. The investigation grid and the anomaly scale

It is easy to understand why the survey scale does not allow us to see the complete reality of the underlying scales. Since the size of the object observed is smaller than the grid, the probability of intercepting it, and therefore of seeing it, will vary. This problem has been widely studied in mining research, and as an example, we have taken Singer's formula (1972), modified by Sego and Wilson (2007) to allow for detection denials (false negatives). Figure 37 shows the output of such a model for circular anomalies and a square sampling grid. This means that, without analysis errors (η=0), the probability that a 50 km x 50 km large-scale monitoring will intercept a kilometer-sized circular anomaly identified by local monitoring is only 0.0003; if the anomaly is about ten kilometers wide (national monitoring), it is 0.031; then the probability increases very rapidly, and with a radius of 25 km, the probability is already 0.78.





Funded by the European Union

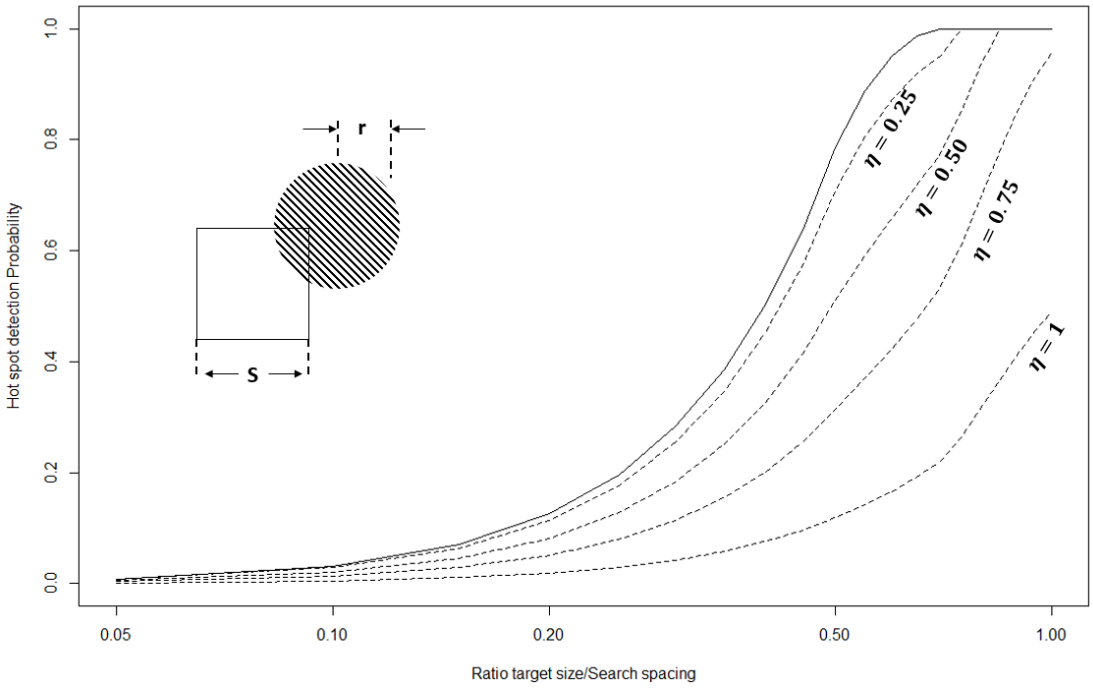


**Figure 37: Hot spot detection probability with various false negative rates. Algorithm from Sego and Wilson (2007)**

*With r, radius of circular target, S: searching square grid, η: false negative rate*

Knowing the total number of samples in a large-scale monitoring, we can estimate the average number of small anomalies detected using a binomial distribution (Table 18).

$$P_{k,n} = \binom{n}{k} \lambda^k (1-\lambda)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$





Funded by the European Union

**Table 18: Number of small-scale anomalies detected by 2018 monitoring of 50x50 km² grid samples**

| Survey | Probability detection by 50x50 km survey ($\lambda$) | Number of anomalies detectable by monitoring of N = 2018 ech |
|---|---|---|
| Survey anomalies 2 km x 2km ( r = 500) | 0.000314 | 1 at p = 0.33 |
| Survey anomalies 2 km x 2 km (r = 1000) | 0.00126 | 2 at p = 0.25 |
| Survey anomalies 20 km x 20 km (r = 10000) | 0.126 | 199 |
| Survey anomalies 25 km x 25 km (r = 25000) | 0.785 | Nearly all |
| Survey anomalies 30 km x 30 km (r = 30000) | 0.951 | Nearly all |

For a 50 km x 50 km survey, the vast majority of anomalies detected will be larger than 25 km x 25 km, while local anomalies detected will represent only a few points.

The observation grid of a phenomenon probabilistically sorts out some of the phenomena we want to see. Even large-scale monitoring can still intercept a small anomaly with probability. Anomaly occurrence would then follow a binomial distribution or its Poisson distribution limit.

## 2.2. Advanced geostatistics taking into account physical laws

Kriging or Gaussian process regression and their variants are used in more than half of all geochemical maps (Belbeze et al., 2023) and, like cartographic AI, are the subject of ongoing research (Chiles and Delfiner, 2011). The author would have been surprised if a kriging variant solution did not exist for the diffusive phenomena of interest to us. In 2019, Albert and Rath demonstrated that the covariance function that is so prized by geostatisticians is simply a Gaussian process (GP) kernel. Van der Boogaart (2001) then developed the relationship between physical phenomena and kriging. He shows that the choice of a particular variogram can be used to solve physical differential equations. Ranftl (2022) has extended this research to neural networks, so that they can also solve physical equations by choosing the activation kernel between neurons.

i.e. g is a Gaussian process $GP$ with a mean of zero,

$$g(x) \sim GP\big(0, k(x, x')\big)$$

with $k(x, x')$ as the covariance function or, in Gaussian formalism, a kernel.

Moreover, if g(x) is governed by a physical equation represented by a linear operator $\widehat{OL_x}$, we have





Funded by the European Union

$$\widehat{OL_x}(g(x)) = 0$$

as a physical operator, of which there are several, such as the Helmholtz operator for waves ($\widehat{OL_x} = \nabla^2 + \upsilon$ ) or, in our case, the Laplace operator ($\widehat{OL_x} = \nabla^2$) for diffusion.

For a Gaussian process, Van der Boogaart (2001) shows that if K is such that

$$\widehat{OL_x}\widehat{OL_{x'}}\, k(x,x')|_{x=x'} = 0$$

then it obeys the physical equation of the operator and vice versa. To solve this equation, the chosen approach is to decompose it into a sum of products of basic $\phi$ functions (wavelets). Albert and Rath (2019) demonstrate that the basic functions are the fundamental solutions of the physical equations

$$k(x,x') = \lim_{n\to\infty} \sum_{i,j}^{P} \phi_i(x) M_{ij}\, \phi_j(x')$$

The Van der Boogaart equation then becomes

$$\lim_{n\to\infty} \sum_{i,j}^{N} (\widehat{OL_x}\phi_i(x)) M_{ij}\, (\widehat{OL_{x'}}\phi_j(x')) = 0$$

This is then solved on a case-by-case basis using the available data. It should be noted that all these constructions and derivations are only valid for Gaussian processes (GP) and linear operators. In the case of a non-linear operator applied to Gaussian processes, the resultant is not a Gaussian process. The preferred approach is to linearize the differential equation beforehand, if possible (Chen et al., 2022).

Under certain conditions, Gaussian field regression can therefore reproduce physical phenomena (Chen et al., 2022) and is widely used, especially when there is a lot of data (Hensman et al., 2013). However, like many geostatistical methods, fine-tuning a covariance model and applying Gaussian field interpolation to a neighborhood of points assuming they have the same mean is highly complex. More often than not, the "Gaussianity" of data must first be restored using logarithmization, Box-Cox, or anamorphosis (Belbeze et al., 2023). This ensures the best possible linear interpolation (Chiles et Delfiner,1999). This is an interpolation with a strong continuity assumption.

For a SIC case, this kind of model calibration and expert sample selection is not possible. Moreover, the assumption that soil content is a Gaussian field seems unlikely in environmental data (Riemann and Filzmoser, 2000).





Funded by the European Union

## 2.3. Developing specifications for our mapping

In 2018, Belbeze (2018) carried out a project on this subject, examining mapping methods used by expert hydrogeologists and, among other things, comparing them with existing geostatistical techniques. This made it possible to detail expert action for mapping, therefore allowing us to plan for the inclusion of expert opinion in our mapping for ISLANDR.

For example, for a groundwater map, this involves gathering all the information sources needed to plot iso-concentrations in a single GIS layer:

- Locally refined mean water piezometry.
- Surface water/groundwater exchanges.
- Geology and lithology.
- Past and present land use.
- Analysis quality.
- Contaminant concentrations.

Based on their knowledge of hydrogeological processes, an expert manually draw plumes, iso-concentrations. In practice, they visualize the path of a contaminant from well to well, in line with their knowledge (Figure 38):

- In keeping with the physical phenomenon, paths traced are always perpendicular to piezometers.
- If there is no change in lithology, the interpolation of concentrations is linear, proceeding from upstream to downstream and from the highest concentration to the lowest.
- If there is a change in lithology, the angle of the iso line profile should mark the jump.
- If there is an edge (plane or imposed flow), that is the starting point.

In the case of different points that are very close to each other, the nearest neighbor value is applied, or else the highest value.

- If the concentration is near a threshold, it changes if the nearest neighbor value is at the threshold.
- For outliers, the expert checks analysis quality to determine whether to exclude the point or assign the nearest neighbor.





Funded by the European Union

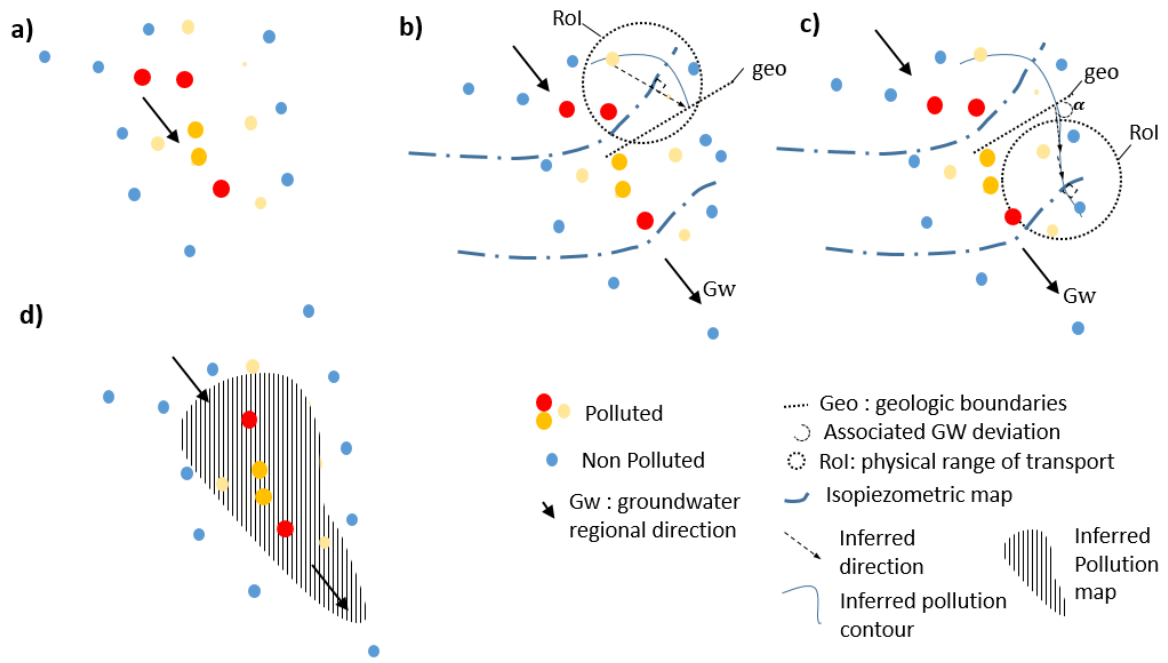


**Figure 38: Groundwater example of expert-driven map production (adapted from Belbeze, 2018)**

*with a) initial well results, and b) expert knowledge added; contouring starts from the top and goes down, c) the flow is diffracted by geologic structure, and d) final contouring*

The expert applies the same type of reasoning to create soil plumes for polluted sites and soils. The expert delineates zones based on analyses and their knowledge of backfilling methods, former buildings, and the extent and shape of pollution plumes. Figure 39 shows that this expert knowledge makes it possible to give anthropogenic pollution the realistic angular contours of historical landfill pits, and that the information to be processed may prove to be dissonant.

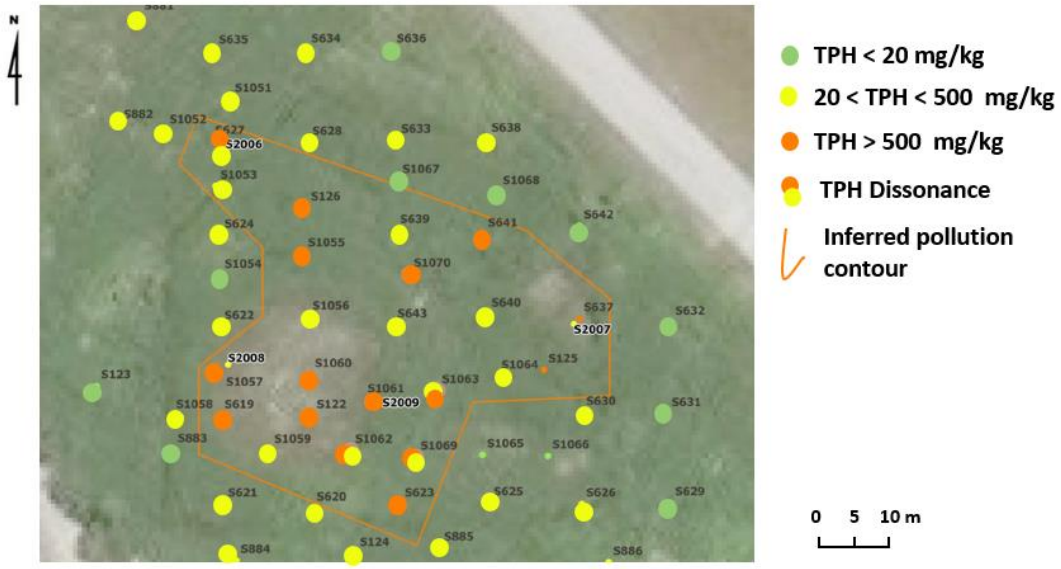


**Figure 39: Polluted soil example of expert-driven map (adapted from Belbeze, 2023)**



Funded by the European Union

In mining geology, experts enter their knowledge of the deposit's origin, faults, and ore continuity on the map, as well as the mining company's extraction possibilities, to produce a map that incorporates as many uncertainties as possible. This approach is comparable to that used for contaminated sites and soils. An example taken from Bardossy and Fodor (2004) is presented in Figure 40.

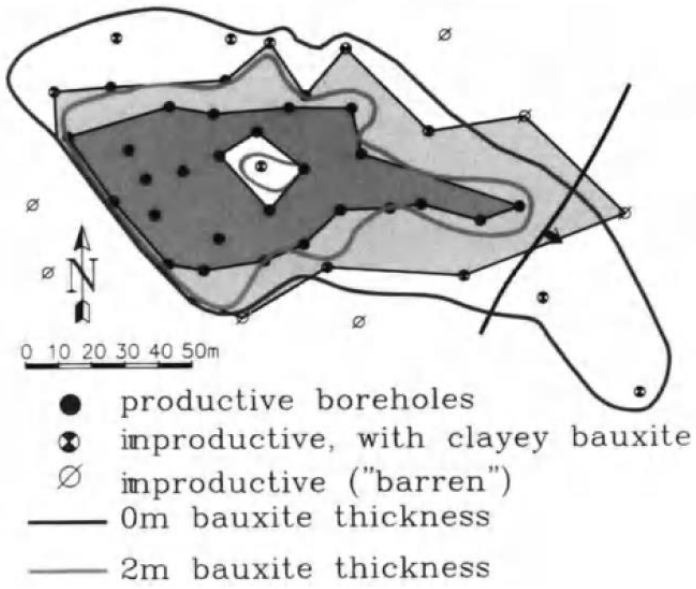


**Figure 40: The Bakonyoszlop XIII. deposit map from Bardossy and Fordor (2004)**

For our ISLANDR interpolator, the expert can add knowledge of physical phenomena, directions, new information, and ranges of phenomena to mapping for our ITAs, and even make decisions when there is dissonance (Figure 41). As a result, the ISLANDR SIC interpolation algorithm must be able to integrate a diffusive physical phenomenon and covariates that the algorithm will have to take into account, depending on their relevance to the expert, ranges, directions, poor sample quality, various sample sizes, advanced uncertainty management and more.





Funded by the European Union

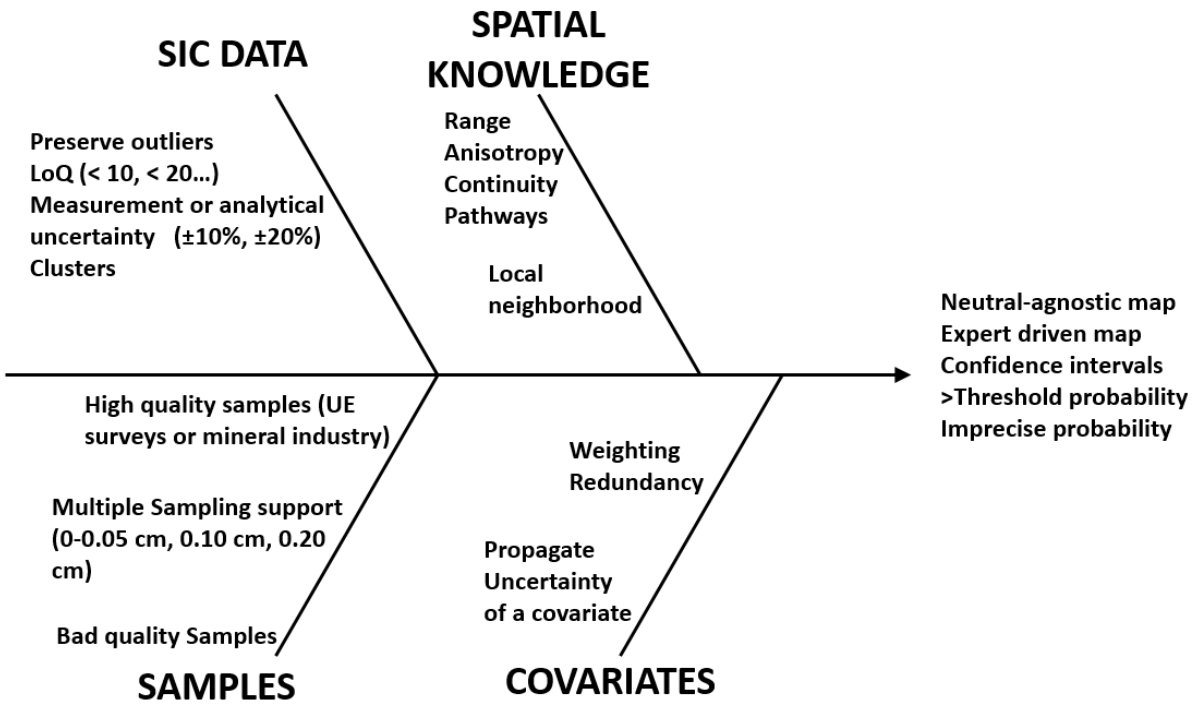


**Figure 41: Fishbone diagram for our algorithm**

Advanced geostatistical methods have already been developed to handle large amounts of data, particularly when physical phenomena are taken into account (Pannecoucke, 2020). Other multi-agent research (MMA) developed for geological modeling could also be adapted to geochemical mapping problems (De Kemp, 2021), but that would be a departure from the SIC aspect.





Funded by the European Union

# 3. Information diffusion mapping

Information diffusion mapping is proving particularly successful where geostatistics lacks the points to obtain a robust variogram (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, 2020). Two mathematical approaches led to this theoretical development:

- An approach developed in China that is based on potential and the analogy between information transfer and the natural phenomena of diffusion or vibration (Huang, 2002; Huang and Shi, 2002). This technique is simply called Information Diffusion; the authors apply it to flood hazard mapping (Huang et al., 1998; Zhou et al., 2000; Chang et al., 2007 etc.), seismicity (Bai et al., 2015), and more recently precipitation (Huang et al., 2019) as illustrated in Figure 42.

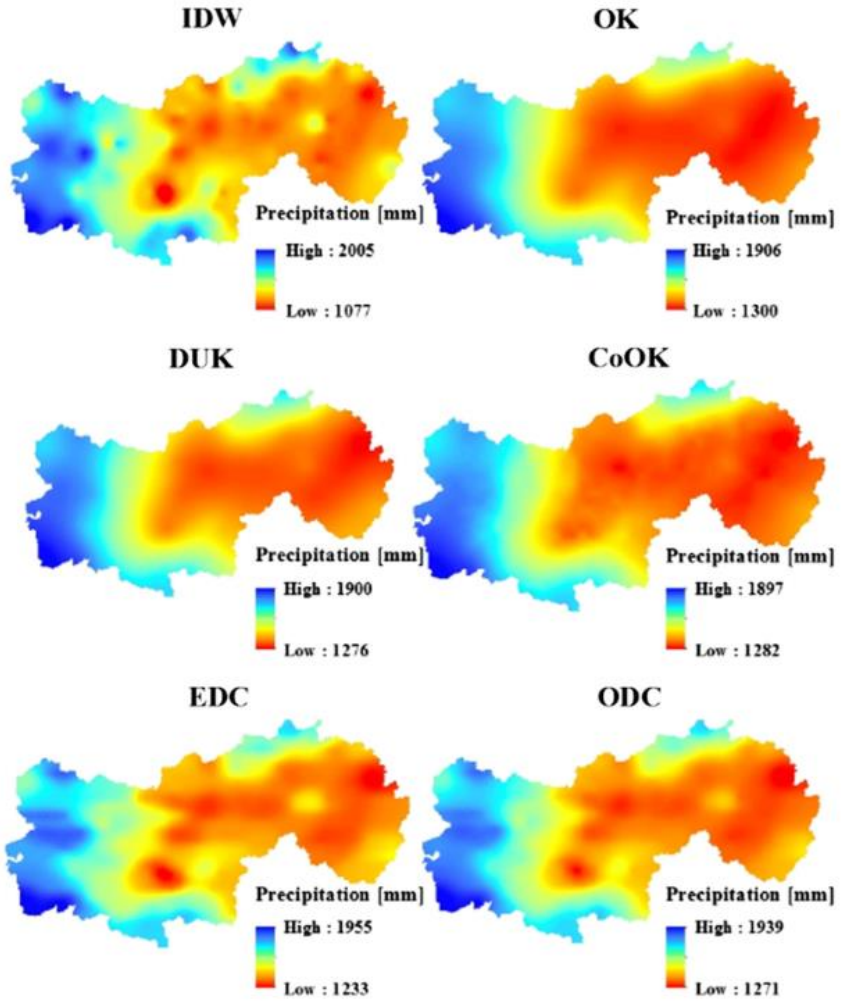


**Figure 42: 2013–2017 precipitation predictions for the city of Linan by IDW, OK, DUK, CoOK and information dissemination by ODC (Gaussian) and EDC (non-Gaussian) from Huang et al. (2019).**





Funded by the European Union

- Beauzamy (2004) developed an entirely probabilistic approach called Experimental Probabilistic Hypersurface, which is based on the propagation of information entropy (Zeydina and Beauzamy, 2013). The IRSN uses this approach for nuclear safety calculations (Godan et al., 2015), various models of neutron sensor networks for nuclear reactor operation, and territory monitoring for radioactive plumes (Khalipova et al., 2017). This method is designed to be as neutral as possible in terms of assumptions. In its initial version, it does not include spatial covariance and is immune to the outliers it magnifies, which is one of ISLANDR's goals.

The analogy with the diffusion or vibration phenomena of the Chinese possibilist approach seems too academic for our very discontinuous pollution. As the probabilistic approach makes no such assumptions and is already spatially weighted, it will be given preference and enhanced.

## 3.1. Possibilistic approach

For Huang (2002), the calculation procedure is based on the possibilities introduced by Zadeh (1978) and developed by Dubois and Prade (87), and is as follows:

If we have n observation points, noted $A_i$, i =1, ..., n with $C_i$ the resulting value from the i-th measurement, We then have a sample $L = \{C_1, C_2, \dots, C_n\}$ of a possible $U = [min, max]$.

If the bijection $\mu : L \times U \rightarrow [0,1]$ is decreasing (convex in u), this is called the diffusion of information from L to U.

Given $L = \{l_1, l_2, \dots, l_n\}$ as our ensemble of probability density measures $f(l)$, its unbiased estimator is

$$\hat{f}(l) = \frac{1}{nd} \sum_{i=1}^{n} \mu\left(\frac{l - l_i}{d}\right)$$

Where d is a constant,

the diffusion function may be linear. If $L = \{l_1, l_2, \dots, l_n\}$, and U is split into m intervals of length Δ, $U = [u_1, u_2, \dots, u_{nm}]$:

$$\mu(l,u) = \begin{cases} 1 - \frac{|l-u|}{\Delta} \; si \; |l-u| \; \leq \Delta \\ 0 \; sinon \end{cases}$$

The U elements are also called control points. We also note that $q_{ij} = \mu(l_i, u_j)$. $q_{ij}$ is the information at the control point $u_j$ given by $l_i$ .

By analogy with the physical processes of diffusion in chemistry, the normal diffusion function (NDF) is defined by

$$\mu(l,u) = \frac{1}{h\sqrt{2\pi}} exp\left[-\frac{(l-u)^2}{2h^2}\right]$$





Funded by the European Union

With the empirical diffusion coefficient (EDC)

$$h = \begin{cases} 0.6841(b-a)\ pour\ n = 5 \\ 0.5404(b-a)\ pour\ n = 6 \\ 0.4482(b-a)\ pour\ n = 7 \\ 0.3839(b-a)\ pour\ n = 8 \\ 2.6851\dfrac{(b-a)}{n-1}\ pour\ n \geq 9 \end{cases}$$

with $b = max\{l_i\}$ and $a = min\{l_i\}$

The optimized diffusion coefficient (ODC) ($h$) is obtained by minimizing an objective function:

$$s(h) = min\left|\frac{\hat{f}(l)}{2n\sqrt{\pi}\left[\hat{f}(l+h) - 2\hat{f}(l) + \hat{f}(l-h)\right]^2}\right|$$

Similar to the physical process of vibrating a tuning fork, a vibrating string diffusion function (VSDF) is established. It is used to assess earthquake risks (Figure 43).

$$\text{VSDF} = \mu(l,u) = \frac{1}{2}\exp\left(-\frac{|l-u|}{d}\right)$$

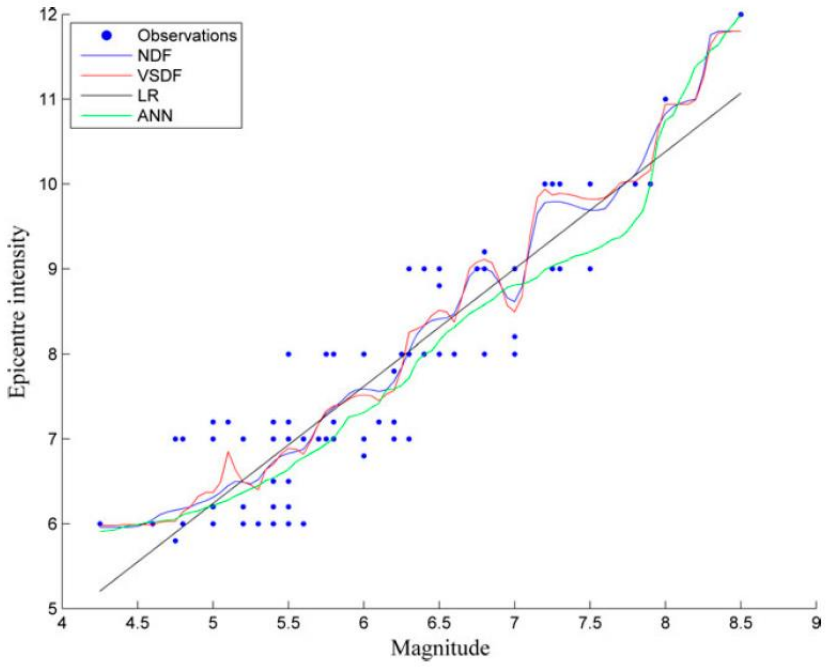


**Figure 43: Calculation of intensity at the epicenter of an earthquake as a function of magnitude from Bai et al. (2015)**

*with NDF: normal diffusion function, VSDF: vibrating string diffusion function, LR: linear regression, ANN: neural networks*

Lastly, using the intrinsic properties of possibilities, it is possible to show that the possibilistic approach approximates reality more quickly than conventional interpolation but mathematicians argue that it is not proven.





Funded by the European Union

## 3.2. Weighted probabilistic approach

This method of transferring information by diffusive processes is called Experimental Probabilistic Hypersurface or EPH.(Figure 44).

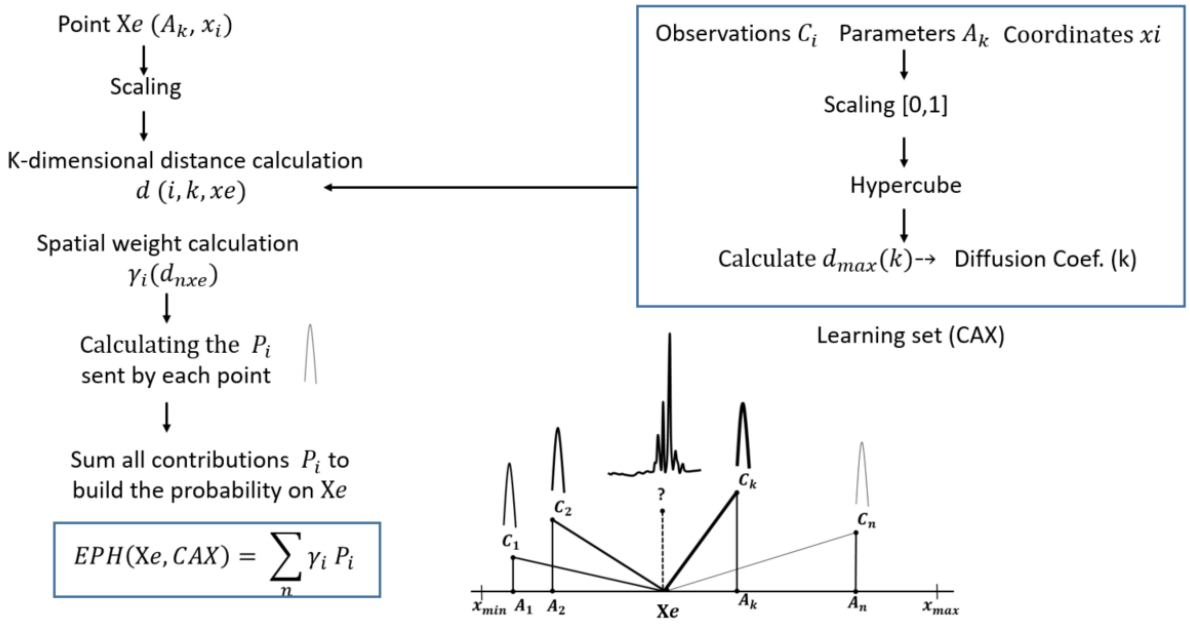


**Figure 44: Building the Experimental Probabilistic Hypersurface (EPH), adapted from Zeydina and Beauzamy (2013)**

For 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 (Figure 45).

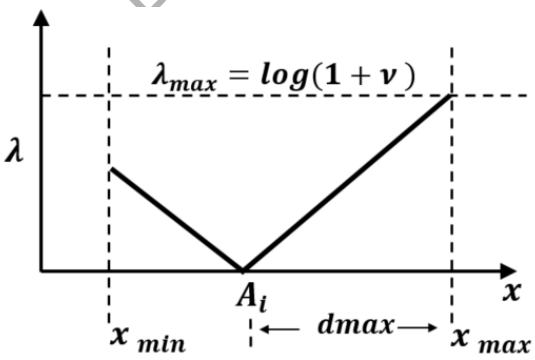


**Figure 45: Entropy variation with the distance to point K to be evaluated**

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 limits of the modeled phenomenon and discretization of this interval (τ steps, υ intervals).

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



Funded by the European Union

precision required (ppm, ten ppm, etc.). A small change in the min-max limits has little impact on the final result.
We generate $c_j$ , the discretization of result range C (with the τ step in υ intervals), and λ, a parameter linked to entropy that is calculated in a way that retains only the minimum information at each point.

$$\lambda = \frac{Log(\mathrm{u}+1)}{d_{max}}$$

With $d_{max}$, the maximum possible distance between the measurement points and the min-max limits of the domain.

The EPH construction is based on the notations shown (Figure 46).This calculates the k-dimensional distance between each measurement point $A_i$ and the point to be reconstructed X, then the contribution of each $A_i$ at the density of X.

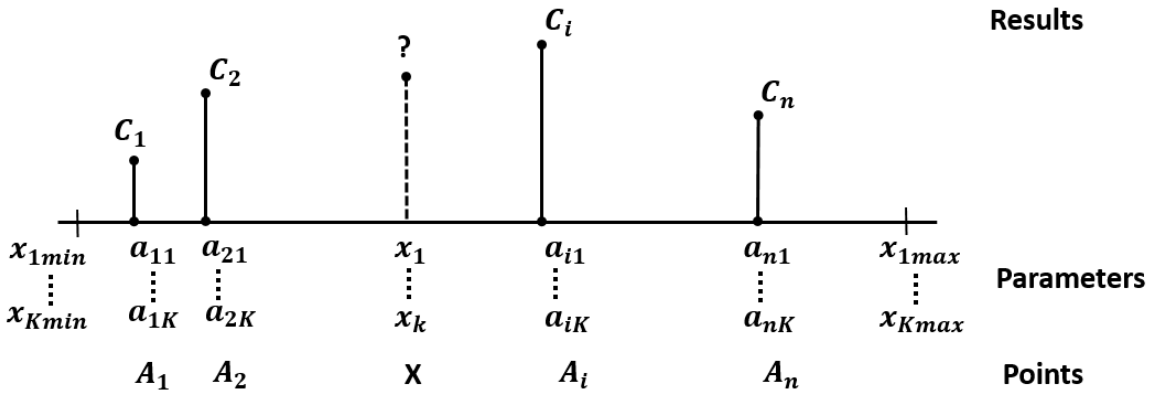


**Figure 46: Notations for the diffusion of information from points A to X where concentration C is unknown**

In detail, with n observation points, denoted by $A_i$, i =1, …, 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 $A_i(a_1, \dots, a_K)$ and $X(x_1, \dots, x_K)$.

We calculate $d_i = d(A_i, X)$, the distance between the point to be reconstructed and the i-th measurement.

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

$$\sigma = \frac{\tau e^{\lambda d_i}}{\sqrt{2\pi e}} \quad (eq.2)$$

Each n-point $A_i$ contributes to the final result of the density of X:

$$P_{A_i,j}(X) = \frac{\tau}{\sigma\sqrt{2\pi}} exp\left[-\frac{(c_j - C_i)^2}{2\sigma^2}\right] \quad (eq.3)$$





Funded by the European Union

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 (Figure 47). 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:

$$P_{xj}(X) = \gamma_1 P_{1,j}(X) + \cdots + \gamma_n P_{n,j}(X) = \sum_{i=1}^{n} \gamma_i \, P_{A_i,j} \quad (eq.4)$$

Where $\gamma_i = \frac{d_i^{-1}}{\sum_{i=1}^{n} d_i^{-1}}$ in dimension 1 and

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

In fact, each $P_{xj}(X)$ is a conditional probability knowing the other measurements made:

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

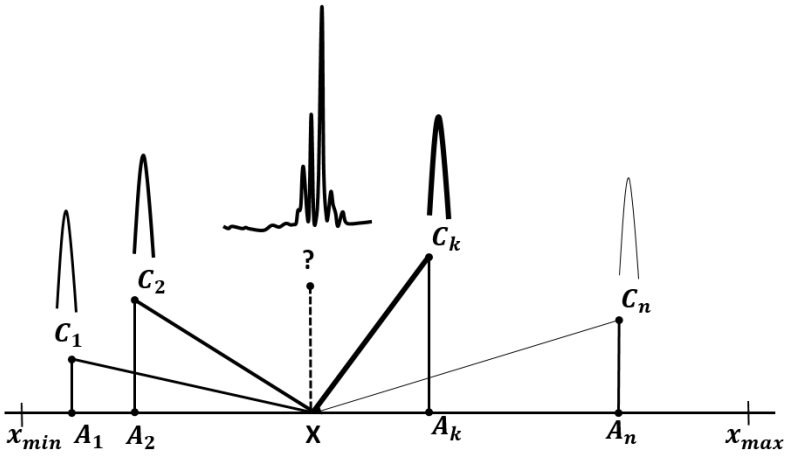


**Figure 47: Information diffusion process attenuated by neighbour distance to X**

Diffusion coefficients 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 distribution (Zeydina et Beauzamy, 2013).

From this probability distribution in X, we can extract a confidence interval $[l_1, l_2]$ at a confidence $\propto$ . For example, $\propto = 0.95$ will give us the interval [Q5, Q95] for our expected value.

$$\sum_{j=0}^{l_1-1} P_{xj} \leq \frac{1-\propto}{2} \quad (eq.5)$$

$$\sum_{j=l_2+1}^{u} P_{xj} \leq \frac{1-\propto}{2} \quad (eq.6)$$





Funded by the European Union

You can also easily modify the distribution of coefficient calculation $\gamma_i$ or $P_{A_i,j}(X)$ to add uncertainties to the calculations. It is also possible to calculate several $P_{xj}(X)$ by $m$ Monte Carlo simulation of a parameter and combine these EPHs to obtain the resultant.

$$P_{xj}(X) = \sum_{k=1}^{m} q_k {P_{xj}}^k(X) \qquad (eq.7)$$

With $q_k$ probability of scenario k

However, if we want to examine the algorithm's response to the whole range of parameter variations and their interactions, the previous method quickly becomes too complicated and costly, so it is better to use a global Monte-Carlo approach based on a probability of exceeding a threshold with several possible situations:
We have $T_0$ if the threshold is considered:

$$P_x(X > T_0) = \sum_{c_j=T_0}^{u} \sum_{i=1}^{n} \gamma_i \, P_{A_i,j} \qquad (eq.8)$$

If we know nothing about the parameters $a_1,..a_k$. We assume that they all follow a uniform distribution on [0,1] and that they are independent:

$$P_{global}(X > T_0) \approx \int_0^1 \dots \int_0^1 P_x(X > T_0)\, da_1..da_k \qquad (eq.9)$$

If, like the in Bayesian approach, we have distributions $h_k$ for the parameters $a_1,..a_k$. If they are considered independent:

$$P_{global}(X > T_0) \approx \int_0^1 \dots \int_0^1 P_x(X > T_0)\, h_1(a_1) \dots h_k(a_k)\, da_1..da_k \qquad (eq.10)$$

If the parameters are not independent and we have the joint distribution $h$ for the parameters $a_1,..,a_k$. We obtain:

$$P_{global}(X > T_0) \approx \int_0^1 \dots \int_0^1 P_x(X > T_0)\, h(a_1,..,a_k)\, da_1..da_k \qquad (eq.11)$$

With SIC data, and barring expert opinion, we have little chance of knowing the joint distributions. We therefore need to be cautious about the parameters we introduce into the model. Compared to kriging (Berton, 2018), EPH is superior in the case of sparsely sampled phenomena (typically n <30). Kriging has been proven to work better when there is a lot of data (generally speaking, any weighted sum of neighboring content tends on average towards the true value when n is large—the so-called law of large numbers), and the dependency between data depends only on the distance between points. The best method for mapping sites will therefore be to study the data and the variogram to see if kriging is possible; if not, EPH will be the best solution. In view of these definitions, it should be noted that, unlike Kriging, EPH makes no assumptions about data variability, particularly in terms of continuity. EPH is sensitive to outliers, which it magnifies. This is a desired effect for ISLANDR, but it remains too sensitive to data clusters, which needs to





Funded by the European Union

be corrected. The possibilities of inserting uncertainties and optimization into EPH are promising and will be pursued.

## 3.3. Links to KDE theories

The KDE method (Parzen, 1962) is recognized as one of the most efficient for calculating a variable's density function. Its estimator has the following form:

$$\hat{p}_n(x) = \frac{1}{n\Delta}\sum_{i=1}^{n} K\left(\frac{X_i - x}{\Delta}\right)$$

With $X_1$, .., $X_n$ observations, kernel function K, and Δ, the smoothing step (smoothing bandwidth). With a Gaussian kernel, we would obtain:

$$\hat{p}_n(x) = \frac{1}{n\Delta\sqrt{2\pi}}\sum_{i=1}^{n} exp\left(\frac{-(X_i - x)^2}{2\,\Delta^2}\right)$$

Since the Δ step is constant, by specifying $y = \frac{X_i - x}{\Delta}$ it is possible to calculate the KDE bias algebraically (Whittle, 1958):

$$E\left(\hat{p}_n(x_0) - p(x_0)\right) = \frac{1}{2}\Delta^2 p''(x_0)\int y^2 K(y)dy + o(\Delta^2)$$

Its variance

$$Var\left(\hat{p}_n(x_0)\right) = \frac{1}{n\Delta}p(x_0)\int K^2(y)dy + o\left(\frac{1}{n\Delta}\right)$$

And other indicators (Wand and Jones, 1994)

MSE$\left(\hat{p}_n(x_0)\right)$ = MISE$(\hat{p}_n)$= $o(\Delta^4)$ + $o\left(\frac{1}{n\Delta}\right)$

Several spatializations of the kernel method have been proposed, so that the neighbor effect decreases with distance as it does in reality (Levine, 2004; Gibin et al., 2007). In 2008, Xie and Jan suggested a formulation adopted by commercial GIS (kernel interpolation):

$$\hat{p}_n(x) = \frac{1}{nr^2}\sum_{i=1}^{n} K\left(\frac{X_i - x}{r}\right)$$

With r the search radius and n the number of neighbors in r.

The most commonly used kernels are, of course, Gaussian kernels and certain quadratic forms. The spatial formulation for the Gaussian kernel is as follows:

$$K\left(\frac{X_i - x}{r}\right) = \frac{1}{\sqrt{2\pi}}exp\left[-\frac{(X_i - x)^2}{2r^2}\right] \; avec\; X_i - x \le r \; et \; 0 \; sinon$$

$$\hat{p}_n(x) = \frac{1}{nr^2}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}exp\left[-\frac{(X_i - x)^2}{2r^2}\right]$$

Authors (Silverman, 1986; O'Sullivan and Unwin, 2002; O' Sullivan and Wang, 2007) have studied the choice of K and the optimal choice of its parameter r, and concluded:





Funded by the European Union

- Parameter r is more important than the choice of k. The best r is usually chosen by cross-validation.
- The search bandwidth r determines the smoothness of the estimated spatial density.

The spatial consideration of 2D KDE is therefore inverse to the distance weighted on the search radius. In the case of multiple dimensions (k-dimension), a pragmatic approach involves approximating the true multi-dimensional kernel by multiplying several kernels together (Scott, 1992).

$$\hat{p}_n(x) = \frac{1}{nh_1 \times \ldots \times h_k} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_1}\right) \times \ldots \times K\left(\frac{X_i - x}{h_k}\right)$$

In parallel, 2D KDE techniques lend themselves remarkably well to the resolution of physical differential equations (Buhmann, 2004). In such cases, the kernels used are called radial basis functions or RBFs. Their growth in physical modeling in recent years has been considerable, and their rapid, accurate results have made commercial physical modeling software like ANSYS and COMSOL a success. For recent work in this field, see Roux et al., 2020 and Shi et al., 2021. The latest developments combine KDE and neural networks (Webster, 2023). For pollution mapping, RBF techniques are used when it is not possible to apply other geostatistical techniques (presence of outliers, non-convergence of variography algorithms, etc.). Buhmann (2004) presents a mathematically detailed study of RBF interpolation. Finally, the use of 2D KDE for hot spot detection (Lin et al., 2010) may be of interest to ISLANDR.

Compared to EPH, we see that $\gamma_i$ is an inverse power function of the distance to point x, while $\sigma_i$ also varies exponentially in inverse as a function of distance.

$$\hat{p}_n(x) = \sum_{i=1}^{n} \gamma_i \frac{\tau}{\sigma_i\sqrt{2\pi}} exp\left[-\frac{(X_i - x)^2}{2{\sigma_i}^2}\right], \sigma = \frac{\tau e^{\lambda d_i}}{\sqrt{2\pi e}}, \gamma_i = \frac{d_i^{-k}}{\sum_{i=1}^{n} d_i^{-k}}$$

Hence, using the KDE formalism

$\hat{p}_n(x) = \sum_{i=1}^{n} \gamma_i \frac{\tau}{\sigma_i\sqrt{2\pi}} K\left(\frac{X_i - x}{\sigma_i}\right)$ avec $K(x) = \frac{1}{\sqrt{2\pi}} exp\left(\frac{-x^2}{2}\right)$

EPH can therefore be seen as a non-stationary, non-linear case of 2D KDE or RBF whose coefficients depend on distance. This dependence of $\sigma_i$ and $\gamma_i$ on distance means that unlike $\Delta$ or $h_k$, they cannot be taken out of sums, which compromises algebraic calculations of bias, variance, MSE, and MISE. However, many of the theoretical mathematical demonstrations developed for KDEs seem transposable to EPH. It can be noted that in the big data context the success of kernel methods stems from the fact that the sum of N kernels weighted to estimate a quantity always tends toward this quantity when N is large (law of convergence of large numbers). This law also applies to EPH, kriging, and other GRNN, making them the methods of choice for big data physics-based interpolation.





Funded by the European Union

# 4. Improved information diffusion for the ISLANDR projet

## 4.1. Objective

This involves introducing new knowledge into the calculation of the "neutral" EPH as it was developed by its creators (Zeydina and Beauzamy, 2013), such as including the scope of phenomena, the weight of explanatory variables, uncertainty on variables, declustering, anisotropy, etc. These modifications have necessitated writing specific codes, modifying the EPH equations or encapsulating the EPH in a double global Monte Carlo system (Figure 48). The system was coded in R (R Core Team, 2022) and can be called EEPH for Enhanced Experimental Probabilistic Surface.

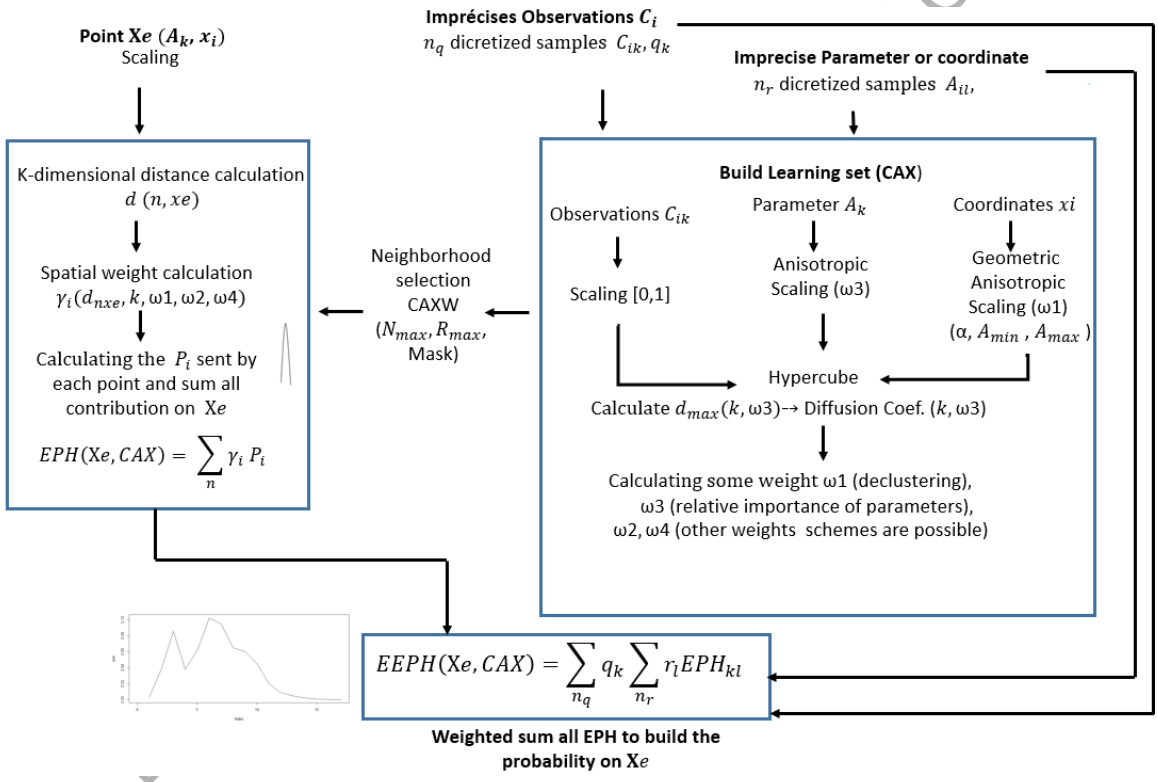


**Figure 48: Building the Enhanced Experimental Probabilistic Hypersurface (EEPH), as the authors have proposed for ISLANDR**

To refine these algorithms and illustrate their effects, a core sample from Dahlberg (1975) with only 13 sparse data points was adapted as a basis for working with small numbers of imprecise data points (Figure 49). With this type of dataset, it is impossible to construct a variogram and therefore to undertake any kind of kriging, which is the desired complexity.





Funded by the European Union

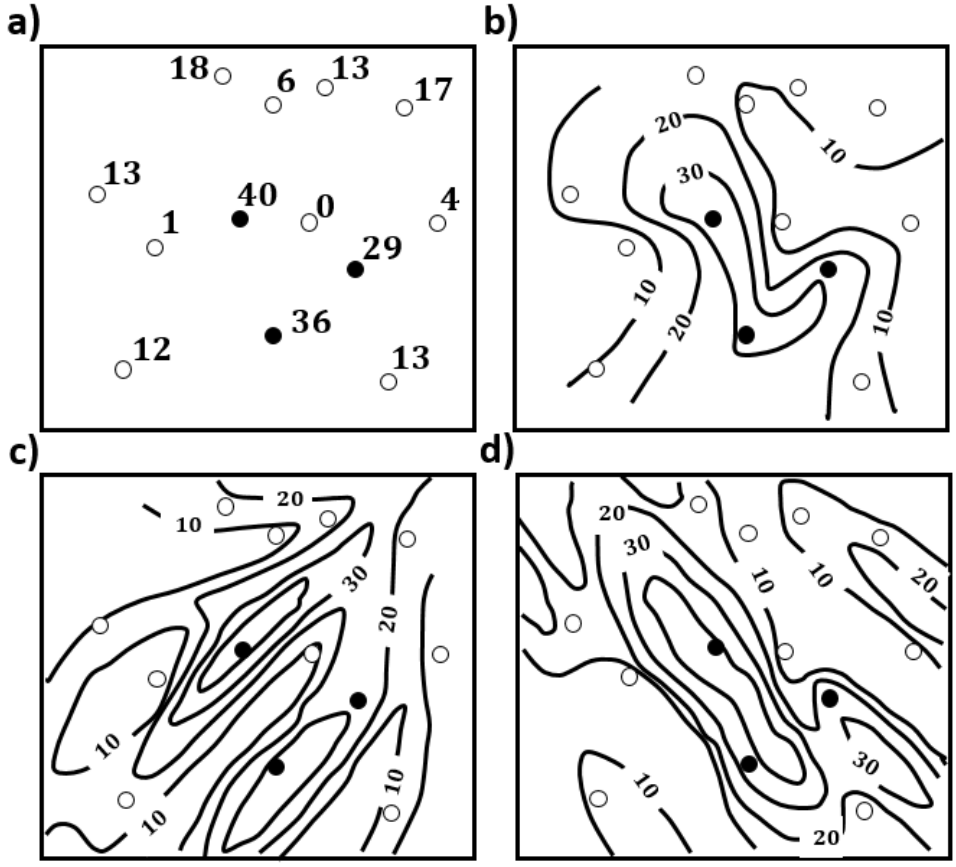


**Figure 49: Four different SIC dataset interpretations adapted from Dahlberg (1975)**

*with a) 13 sand thickness data points in m, b) manual triangulation, c) geologist interpreting profiles as regional northwest strike and southwest paleoslope fluvial deposits, and d) geologist interpreting profiles as Channel sand deposits.*

## 4.2. EPH in the Zeydina and Beauzamy (2013) version

The EPH algorithm derives a probability density at any point (Figure 50), which makes it possible to calculate the 95% or 90% confidence interval around the expected value (Figure 51).





Funded by the European Union

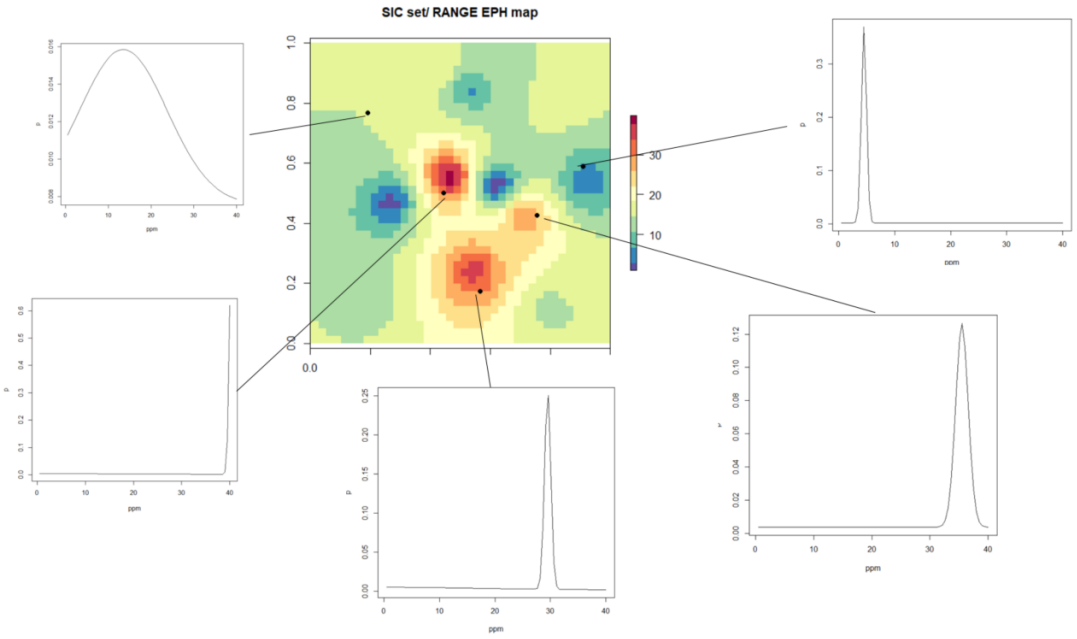


**Figure 50: 13 data points with no uncertainty from Dahlberg (1975), corresponding EPH-generated map with calculated probability density for 5 selected points.**

It is also possible to determine the probability of exceeding a threshold (**Error! Reference source not found.** d).

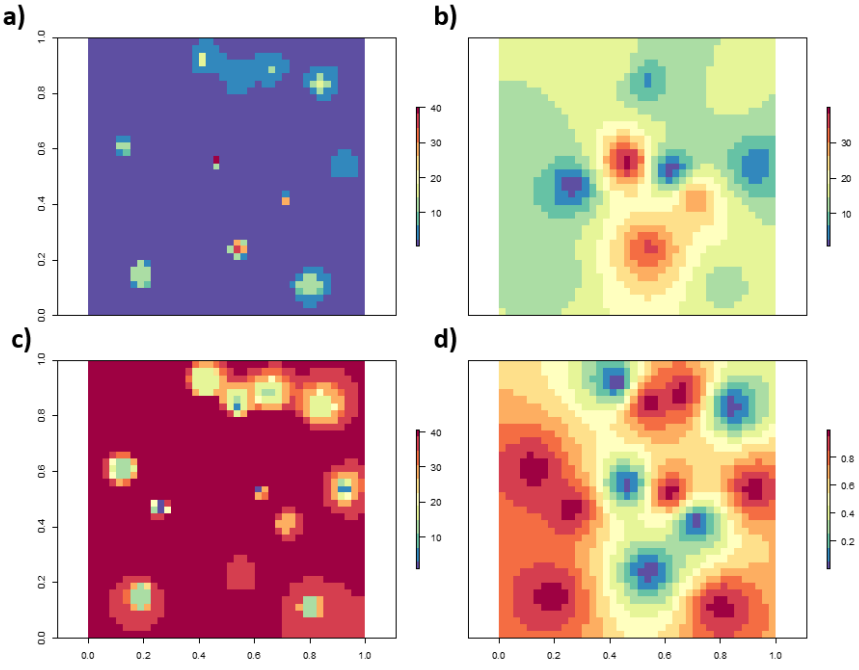


**Figure 51: 13 data points with no uncertainty from Dahlberg (1975) and corresponding EPH-generated map**

*with a) low limit l_1 at 95% confidence such that ⟦E ∈[l⟧ _1,l_2] , b) E, probabilistic expected value by EPH, c) High limit l_2 at 95% confidence such that ⟦E ∈[l⟧ _1,l_2], and d) probability of sand thickness < 15 m.*



Funded by the European Union

This neutral calculation for a given confidence interval is easily converted into a triangular or trapezoidal fuzzy number for each point on the plane (Bardossy and Fodor, 2013). These numbers (Figure 52) can thus be inserted as trapezoidal (Figure 53) in Hyrisk calculations (Guyonnet et al., 2005). This last calculation would be easy to encapsulate and modify to process spatial data (rasters). This approach could be developed within the framework of other ISLANDR WPs, if they wanted to do so.

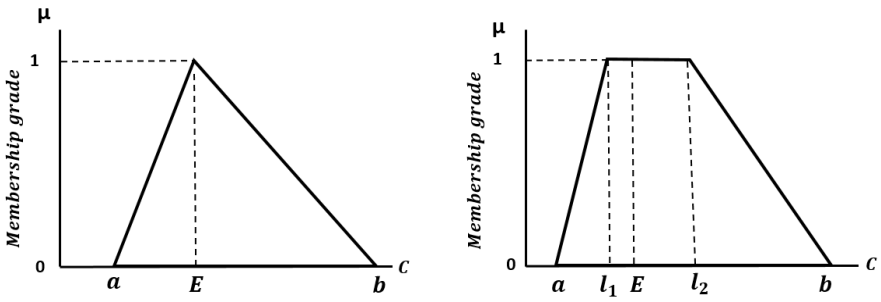


**Figure 52: Neutral EPH results maps as converted to fuzzy number per point.**

With [a, b] range of possible concentrations, E expected value as calculated by EPH, [l_1, l_2] 95%, 90%, or 70% confidence interval as required

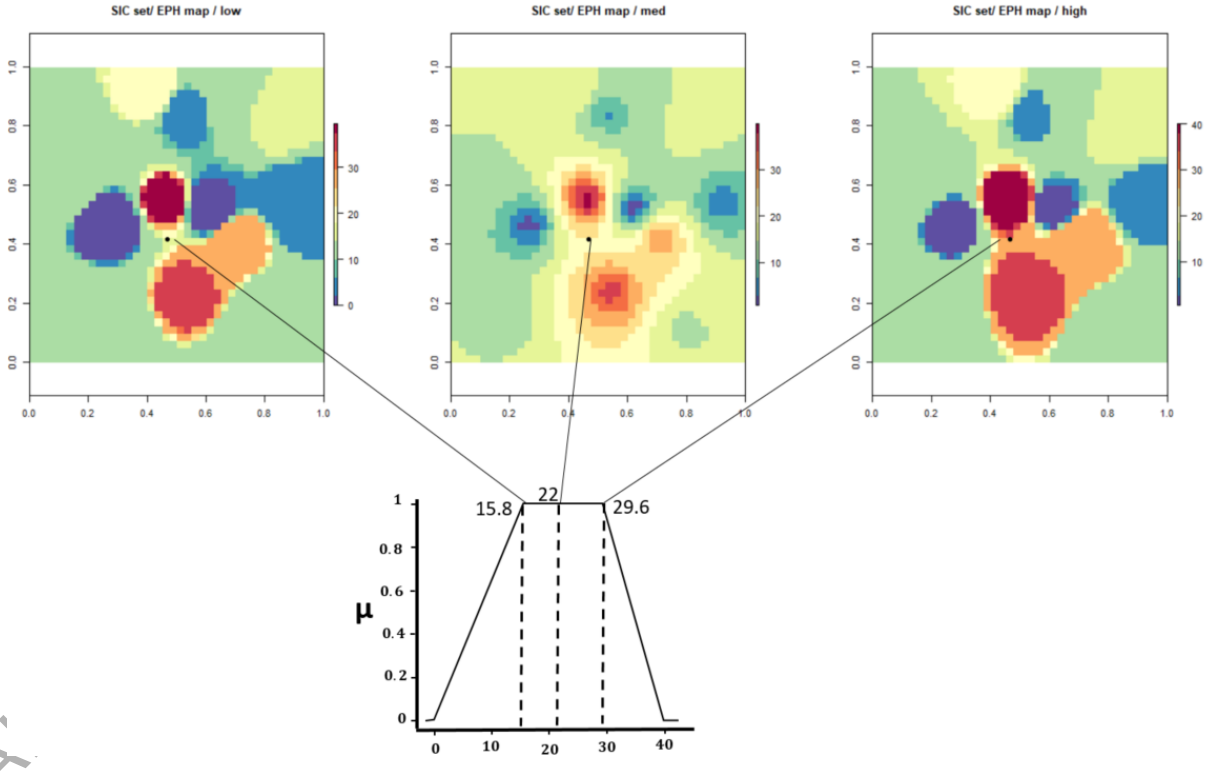


**Figure 53: 13 data points with no uncertainty from Dahlberg (1975) and corresponding EEPH-generated map at 70% confidence interval as converted to trapezoidal fuzzy number per point.**

Note the support of the fuzzy number is min/max of the sand survey from Dahlberg (1975). The core is the 70% confidence limit around the expected value.



Funded by the European Union

## 4.3. Thoughts about factoring uncertainty into our calculations

Uncertainty will always be present in our pollution measurements for the ISLANDR project. Uncertainty may affect our data or our interpolation parameters, for example, due to imprecision in the geolocation system, sampling error, or analysis error.
There are various types and representations of uncertainty (Ferson and Ginzburg, 1996):

- Stochastic uncertainty affecting the measurement, which is usually represented by a probability distribution or a known distribution such as uniform, Gaussian, binomial, Poisson, or hypergeometric distribution. In Figure 54 a, an expert would say that Cd concentration follows a Gaussian distribution with mean 1.5 ppm and standard deviation 1. Without any knowledge of the distribution, we would tend to use a known probability distribution. So, in Figure 54 b, for example, if a triangular probability is choosen the measured Cd content is estimated at 1 mg/kg and ranges from 1 to 2 ppm.

- Uncertainty in the form of a min-max interval for an unknown measurement or parameter. In Figure 54 c, an expert would say that the Cd concentration is between 0.5 and 2.5 ppm.

- There are uncertainties in the form of min-max intervals, as before, but with expert-selected preferences for some that they consider more plausible than others. We then use fuzzy sets. In Figure 54 d, an expert would say that although concentrations may vary from 0.5 to 2.5, the most plausible level would be between 1 and 1.5 ppm.

- The possibilities are flexible in use, and it is also possible for the expert to know the shape of the distribution of the variable, such as binomial, but to be unaware of all the parameters. This is referred to as imprecise probability or possibility distribution. In Figure 54 e, the most probable level is 1 mg/kg, the distribution is binomial, but the number of times it has been observed over the total number of trials is unknown.

- For uncertainties between two minimum and maximum probability distributions, this is a $[\overline{P}, \underline{P}]$ pair. This is called plausibility-credibility. In Figure 54 f, an expert estimates a 50% chance of our content falling between 0.7 and 1.5. If experts know the quantile limits, it is possible to create a P-box, which is a $[\overline{F}, \underline{F}]$ pair of cumulative distributions. In Figure 54 g, an expert would say that it has a 50% chance of falling between 1.7 and 2.5 ppm.





Funded by the European Union

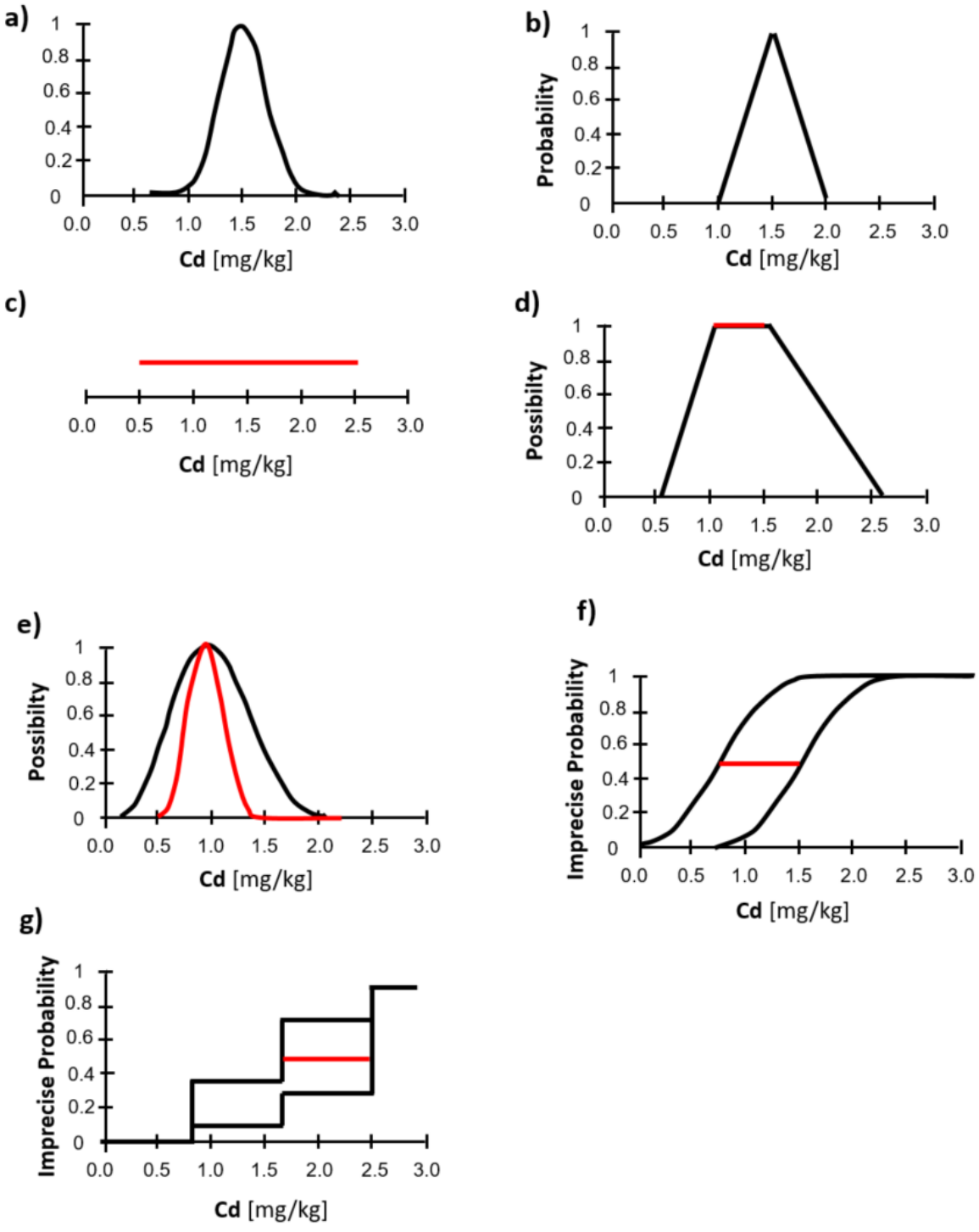


**Figure 54** : Some well-known representations of uncertainty that could be handled by our algorithm.

Subsequently, in our spatial calculations, we can incorporate this uncertainty into our calculations by sampling the probability distributions, the intervals that we inject into our calculations using a Monte Carlo procedure to obtain the distribution function of the



Funded by
the European Union

result (Figure 55). Imprecise possibilities and probabilities can also be sampled, but will give results expressed also in imprecise probabilities (Figure 56).

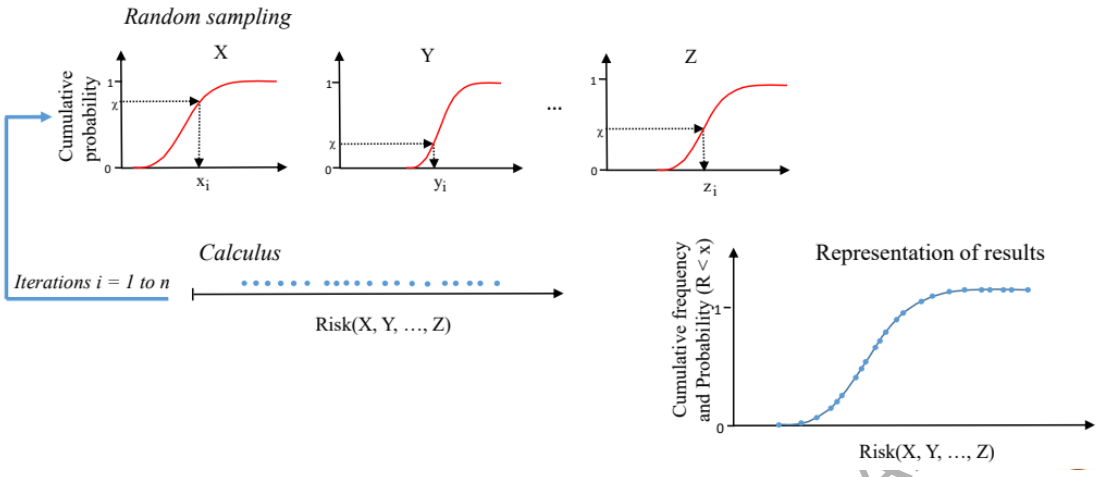


**Figure 55: The Monte Carlo method of propagating stochastic uncertainty from Baudrit et al., 2006.**

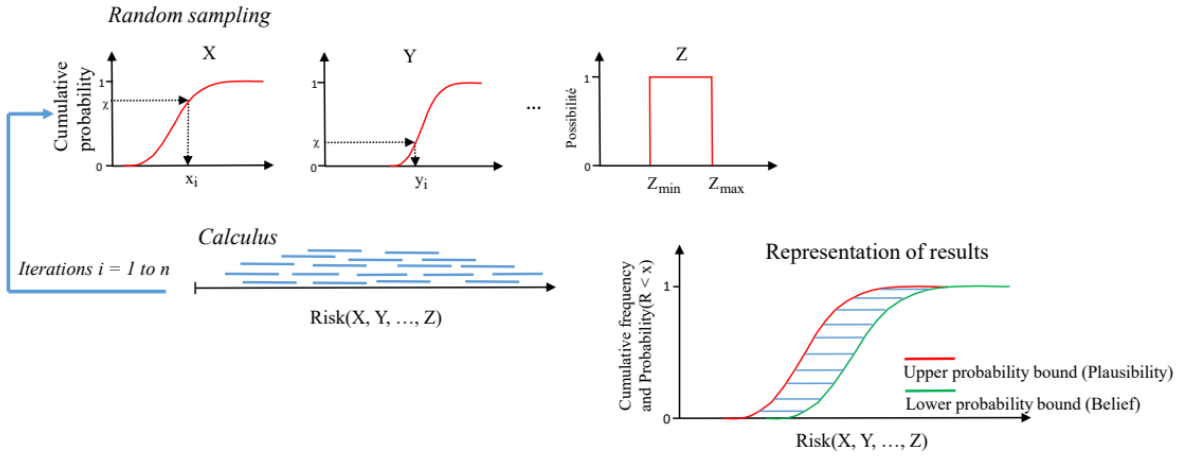


**Figure 56: The hybrid approach: associating stochastic and epistemic uncertainty in propagation, from Baudrit et al., 2006.**

## 4.4. Values below the detection limit and measurement uncertainties

### 4.4.1. EEPH calculation by interval

The presence of what are known as censored values corresponds to values below the limit of quantification. 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. We replace our unquantified values with discretized (and unsimulated) $m$ values of index k



Funded by the European Union

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, and the LOQ distribution could just as well be a possibility instead of a uniform probability. With SIC data, there is no particular reason to prefer one distribution over another.

The calculation protocol is taken from the previous chapter. If $q_k$ is the probability of occurrence of the discretized observation of index k, we will have

$$P_{xj}(X) = \sum_{k=1}^{m} q_k {P_{xj}}^k(X) \quad (eq.7)$$
$$\text{Or } {P_{xj}}^k(X) = \sum_{i=1}^{n} \gamma_i \, P_{A_i,j,k} \quad (eq.4)$$

Therefore, $P_{xj}(X) = \sum_{k=1}^{m} q_k \sum_{i=1}^{n} \gamma_i \, P_{A_i,j,k}$

And by rearranging the sums,

$$P_{xj}(X) = \sum_{i=1}^{n} \gamma_i \sum_{k=1}^{m} q_k \, P_{A_i,j,k} \quad (eq.12)$$

The calculation on the 13 datasets (Dahlberg ,1975), four of which have been censored at 7 m thickness (LOQ), is shown in Table 19 and the calculated expected value in Figure 57. This figure shows the powerful effect of <LOQ values on mapping. The expected value calculated by EPH decreases, reflecting the wide spread of the probability density.

**Table 19: Input data for calculations on the 13-point dataset (Dahlberg,1975) with both datasets modified by censoring and/or applying uncertainty to some.**

| Name | X | Y | Original sand thickness (m) | Censored sand thickness (m) | Is censored indicator | Uncertainty as known on censored sand thickness (Mean %) |
|---|---|---|---|---|---|---|
| D1 | 0.13 | 0.61 | 13 | 13 | 0 | 20 |
| D2 | 0.42 | 0.93 | 18 | 18 | 0 | 30 |
| D3 | 0.26 | 0.47 | 1 | 7 | 1 | 0 |
| D4 | 0.54 | 0.85 | 6 | 7 | 1 | 0 |
| D5 | 0.19 | 0.15 | 12 | 12 | 0 | 20 |
| D6 | 0.46 | 0.56 | 40 | 40 | 0 | 20 |
| D7 | 0.66 | 0.88 | 13 | 13 | 0 | 20 |
| D8 | 0.62 | 0.53 | 0 | 7 | 1 | 0 |
| D9 | 0.84 | 0.84 | 17 | 17 | 0 | 20 |
| D10 | 0.54 | 0.24 | 35 | 35 | 0 | 20 |
| D11 | 0.71 | 0.42 | 29 | 29 | 0 | 20 |
| D12 | 0.92 | 0.54 | 4 | 7 | 1 | 0 |
| D13 | 0.80 | 0.11 | 13 | 13 | 0 | 20 |

Similarly, our content may be affected by uncertainty, usually expressed as a relative percentage ±E, for example a TPH content of 50 ppm ± 10%. The values therefore fall





Funded by the European Union

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, like with the LOQ problem, we propose a discretization approach. We replace our concentration range with $m$ discretized values (non-random draw) between high and low range, each yielding an EPH 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 our uncertainty any particular shape. Nevertheless, the algorithm is still compatible with a possibilistic approach, and the error distribution could take many shapes. With SIC data, there is no particular reason to prefer one distribution over another. As uncertainty increases, the expected value calculated by the EPH decreases, reflecting the spread of the probability density under the effect of uncertainty (Figure 57 c).

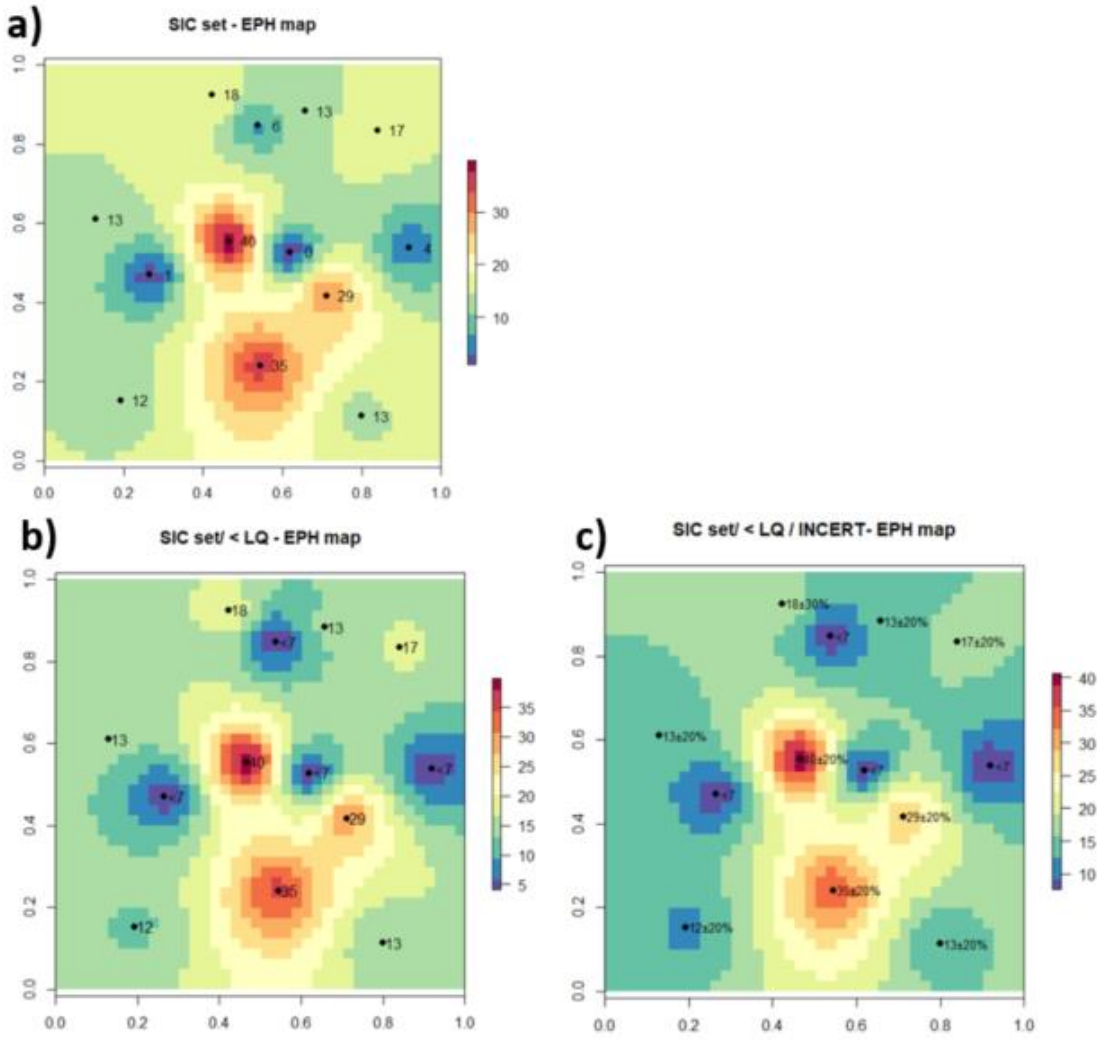


**Figure 57: 13 data points from Dahlberg (1975) and corresponding EPH-generated map**

*with a) expected value map, b) expected value with censored 4/13 <LOQ (7 m), and c) expected value with 4/13 <LOQ and 8/13 with analytical uncertainty of 20% and 1/13 of 30%.*





Funded by the European Union

### 4.4.2. Content uncertainties expressed in Hyrisk format for a spatial EEPH

To open the algorithm to any uncertainty experts know about content or parameters, a version of EPH has been designed for Hyrisk formalism (Dubois and Guyonnet, 2011), and it is now possible to use non-uniform shapes for uncertainty, but this is a "strong" expert choice. With this particular version of EEPH, for each measurement point or parameter, the user must set a type of uncertainty (possibility, probability, imprecise probability) and specify its various distribution parameters in Hyrisk format (Figure 58)

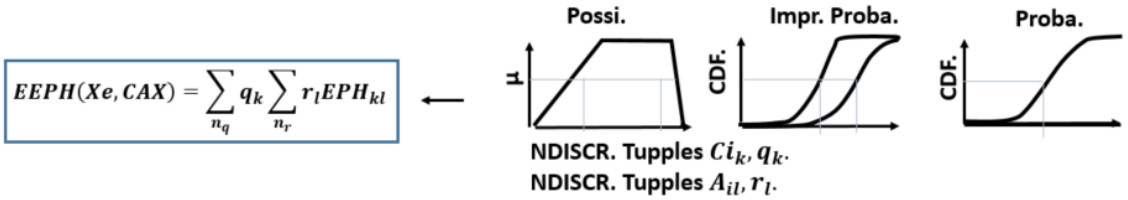


**Figure 58: Feeding the Enhanced Experimental Probabilistic Hypersurface (EEPH) with several representations of an uncertain quantity**

To this end, the measurement data format must specify the type of uncertainty representation associated with the measurement ("possi", "proba", "impr proba"), its distribution type, and its possible parameters in user mode: its sampling function and quantile function. The distribution is then sampled and the $q_k$ weight is calculated. Table 20 shows a dataset of Hyrisk inputs for our dataset of thirteen core samples. It is purely hypothetical and has been created to demonstrate that, through expertise, core loss measurement, and representativeness, the measurement distributions on each are known or fixed as prior in a Bayesian approach

**Table 20: Input data for calculations on the 13-point dataset (Dahlberg ,1975) with Hyrisk input of the various measurements.**

| Name | X | Y | HYRISK | Mode | Para1 | Para2 | Para3 | Para4 |
|---|---|---|---|---|---|---|---|---|
| D1 | 0.13 | 0.61 | 1 | imp proba-normal | 13 | 3 | 17 | 5 |
| D2 | 0.42 | 0.93 | 1 | proba-triangle | 12 | 24 | 18 | |
| D3 | 0.26 | 0.47 | 1 | possi-interval | 0 | 7 | | |
| D4 | 0.54 | 0.85 | 1 | proba-lognormal | 1.8 | 0.4 | | |
| D5 | 0.19 | 0.15 | 1 | proba-lognormal | 2.4 | 0.1 | | |
| D6 | 0.46 | 0.56 | 1 | possi-trapezoid | 30 | 40 | 35 | 38 |
| D7 | 0.66 | 0.88 | 1 | proba-lognormal | 2.5 | 0.2 | | |
| D8 | 0.62 | 0.53 | 1 | possi-interval | 0 | 7 | | |
| D9 | 0.84 | 0.84 | 1 | proba-lognormal | 2.8 | 0.2 | | |
| D10 | 0.54 | 0.24 | 1 | proba-lognormal | 3.5 | 0.1 | | |
| D11 | 0.71 | 0.42 | 1 | proba-lognormal | 3.36 | 0.4 | | |
| D12 | 0.92 | 0.54 | 1 | proba-lognormal | 1.38 | 0.2 | | |
| D13 | 0.80 | 0.11 | 1 | proba-lognormal | 2.5 | 0.4 | | |





Funded by the European Union

Once the data has been imported, the Hyrisk version of the EPH calculates the expectation or probability distribution, based on the probability-possibility distribution of the content at each point (Figure 59). When compared with previous figures for the same site, this figure shows the powerful effect of uncertainty on the deposit's overall structure.

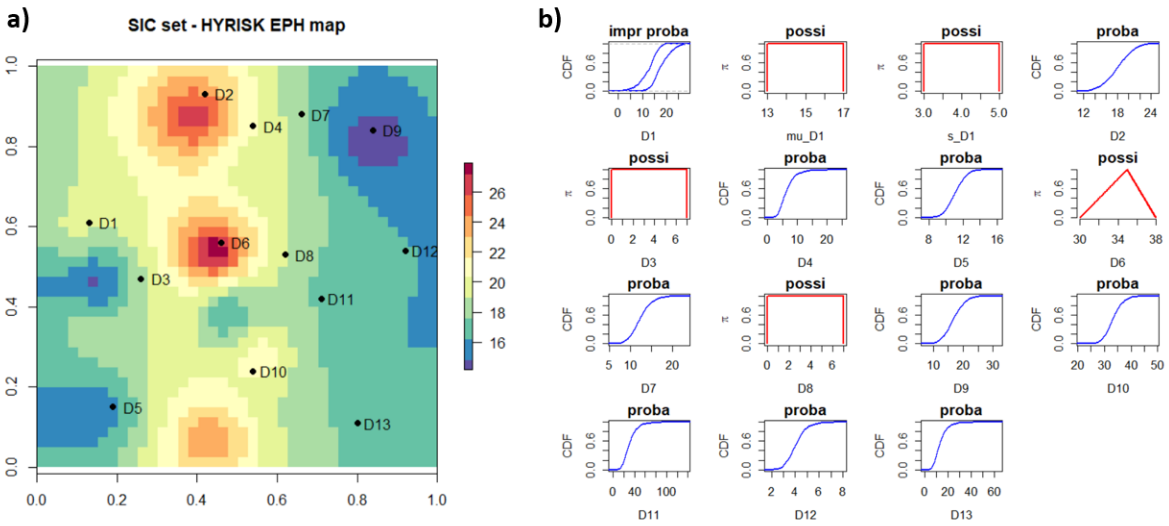


**Figure 59: 13 data points from Dahlberg (1975) with uncertainty in Hyrisk formalism and the corresponding EEPH maps generated.**

*With a) expected value map of the SIC set and b) SIC set in Hyrisk format.*

For a more complex imprecise probability calculation, the distribution type is fixed, but the distribution parameters themselves have a mode and four possible parameters. In this way, the Hyrisk table can be populated with up to 12 parameter fields. In site studies, this type of configuration will rarely be included, due to the uncertainties and SIC characteristics involved, but it has been coded for the sake of theoretical academic completeness.

### 4.4.3. From spatial EEPH to Hyrisk calculation

The EEPH calculation is then run in full tensor mode of probability, which will generate either a "proba" or a "dual proba" for Hyrisk at each point (Figure 60).





Funded by the European Union

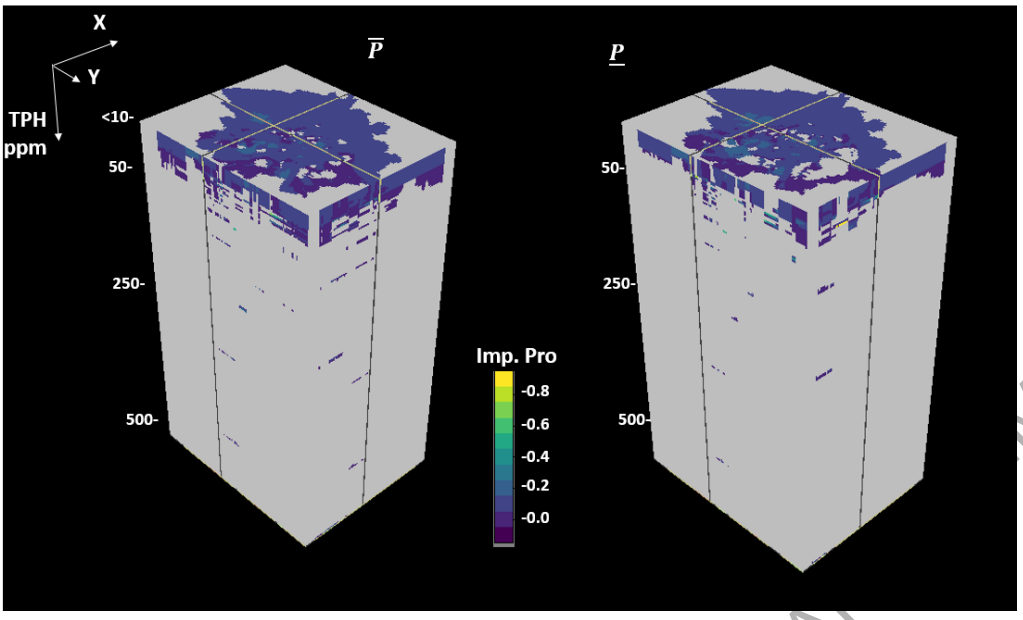


**Figure 60: Tensor cubes of imprecise probability as produced by our EEPH algorithm.**

Topsoil samples from Belbeze et al. (2019), n=139, 0-10cm, TPH analysis by multiple laboratories, LOQ: 10 mg/kg (8 samples), 20 mg/kg (57 samples).

To switch from this tensor representation to a Hyrisk format, a user mode is created in Hyrisk. This user mode includes a specific proba type, a discretization universe, and quantile and random sampling functions.

- An "EEPH with vector field parameter _pxx" (experimental density of EEPH at a point) user type.
- The "_universe" of the EEPH discretization.
- The "_quser" functions: a quantile extraction and random sampling function _qrand specific to each point must be created (Figure 61)





Funded by the European Union

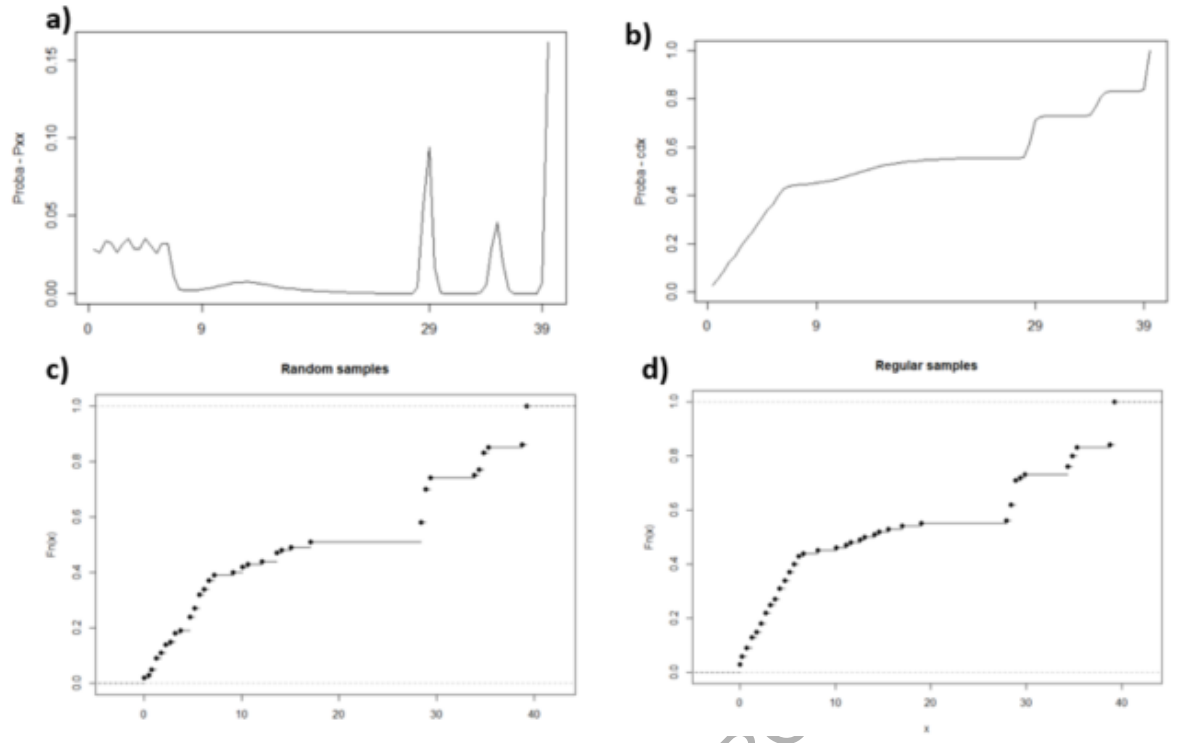


**Figure 61: EEPH probability density exports for Hyrisk calculus**

*with a) an EEPH probability density of a random point at coordinate (23, 23) on the Dahlberg set grid, b) generated continuous cdf, c) experimental cdf of 100 random samples, d) experimental cdf of 100 regular samples.*

In this way, it is possible to create a Hyrisk level for each pixel from the cube tensors (Figure 62) and thus perform a typical chronic risk calculation by applying the following formulas and assumptions (Dubois and Guyonnet, 2011).

$$IER = D \times UER$$

With
IER = Individual Excess Risk (expected excess cancers resulting from dose D)
D = Dose absorbed (mg/kg-d)
UER = Unit Excess Risk (expected excess cancer per unit dose; [mg/kg-d]-1). For arsenic, UER = (1.5 mg/Kg-d)-1.

And

$$D = \frac{SI \times CS \times BA \times EF \times ED}{BW \times AT}$$

With
SI = Soil Ingestion (kg/d). Preferred value is 70 mg/d on a 0-200 support.
CS = Concentration in Soil (mg/kg) taken as a result of the full tensor probability EEPH.
BA = As Bioaccessibility (unitless) EF = Exposure Frequency (days/yr). Preferred value is 10% on a 0–52% support.
EF = Exposure Frequency (d/yr). Preferred value is 78–156 on a 00–365 support.
ED = Exposure Duration (yrs). A child scenario is 6 yrs.





Funded by the European Union

BW = Body Weight (kg). Normal of mean 15.5 and sd = 5.4 kg.
AT = Averaging Time (yrs). 70 yrs is standard risk procedure.

In this way, from an uncertain arsenic map such as those of ITA3, it is possible to run a Hyrisk calculation (Figure 63), which will take 2.21 days.

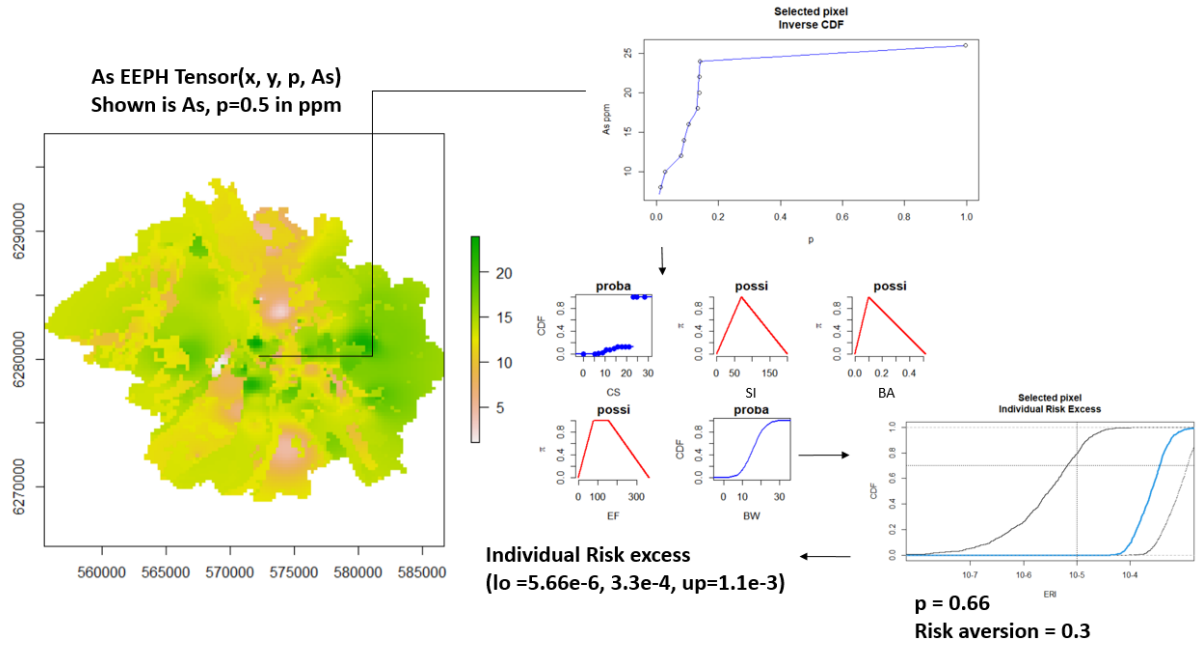


**Figure 62: Hyrisk analysis calculus from EEPH maps.**

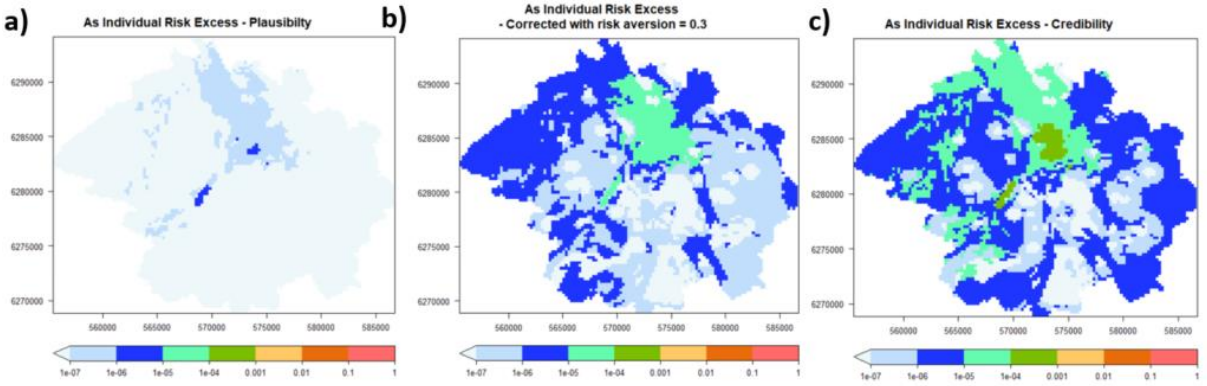


**Figure 63: Spatial individual excess risk estimates resulting from a Hyrisk/EEPH. As concentration in ETA3.**

An area to the north of the metropolitan area has a potentially unacceptable individual risk for arsenic (Figure 63). This calculation and our knowledge of this area must be refined before we alert the authorities. It should also be pointed out that this calculation can only be performed on SIC data. The memory and time requirements for proba-possi hybrid spatialized risk reasoning will be very high, as each parameter of the content possibility functions or parameters at any point will have to be stored in memory or downloaded to the hard drive.



Funded by
the European Union

## 4.5. Uncertainties regarding an explanatory variable or parameter

When the uncertainty of a parameter is known in discrete, interval, fuzzy number, distribution interval form, we can consider various values for it and, as we have seen, launch as many EPHs as necessary, which will be recombined to give the final result (Figure 64). This system is particularly suitable for uncertainties that cannot be incorporated directly into the EPH calculation, such as those relating to content. This would be the case for X and Y positioning errors or other parameters.

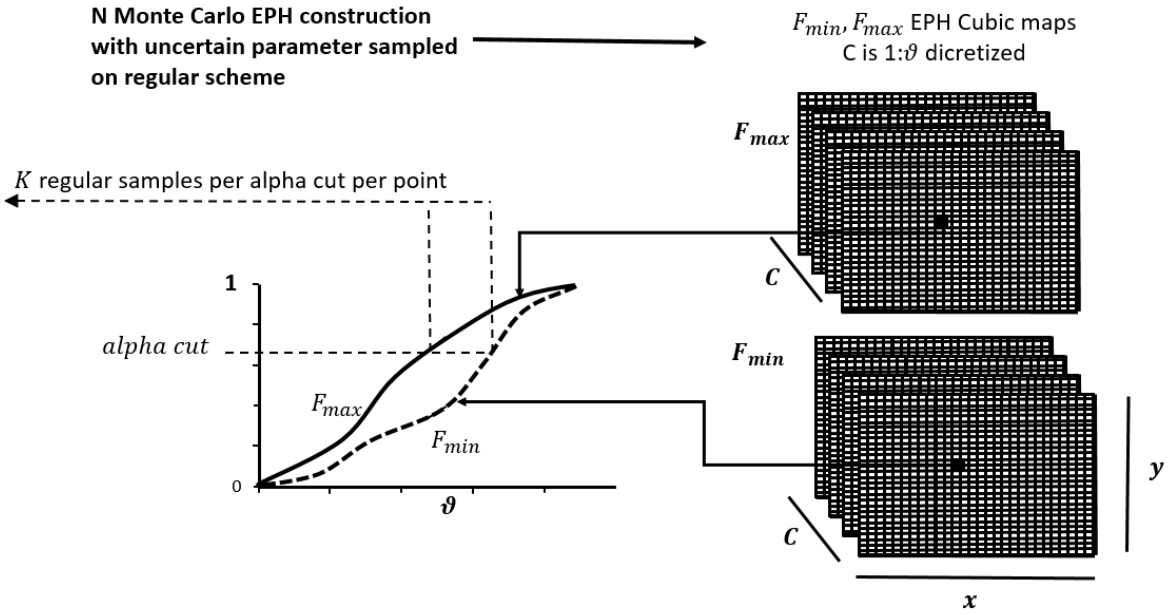


**Figure 64: Keeping only the min et max for each point on each N Monte Carlo run to feed the risk calculus**

EPH calculates an experimental distribution function of the data at each point. It is possible to calculate the distribution function $F$. In the cycle of a Monte Carlo calculation involving a parameter (N cycles), for each point K, we would have N distribution functions $F_k$, themselves discretized into $\vartheta$ increments. The memory load would be too high if we had to keep the ensemble for uncertainty processing. This is referred to as dimensional dispersion: the curse of dimensionality. For example, for GEMAS data, on a grid of 461 x 548 pixels (10 km x 10 km) with discretization $\vartheta$ = 700 and N=1000 scenarios, we would have $10^{11}$ values to store. One solution would be to keep only the calculation of the min ($F_k$) and max ($F_k$). In fact, at each point, we will have a min ($F_k$) and max ($F_k$), which can become plausibility (pl) and belief (bel) once standardized (Figure 65). For GEMAS data, the number of data points to be stored becomes close to 353 million, which is high but manageable using a calculation cluster.

From a practical point of view for ISLANDR, an uncertainty calculation of this kind can only be launched effectively on a selected part of the ITA where this type of in-depth analysis



Funded by
the European Union

is necessary. If we take the data from the Toulouse ITA, $\vartheta = 17$ on a 101 x 125 (250m x 250m) grid, the number of data points to be stored becomes a more acceptable 214,625. $2\vartheta = 34$ rasters are therefore needed to store $F_{min}$ and $F_{max}$, assuming poor soil recovery in the drill cores for these data points and, for example, varying the relative error of the dataset of measurements between 10 and 100%. It is possible to generate two tensor cubes of 17 rasters each containing a high-resolution $F_{min}$ and $F_{max}$ (**Error! Reference source not found.**).

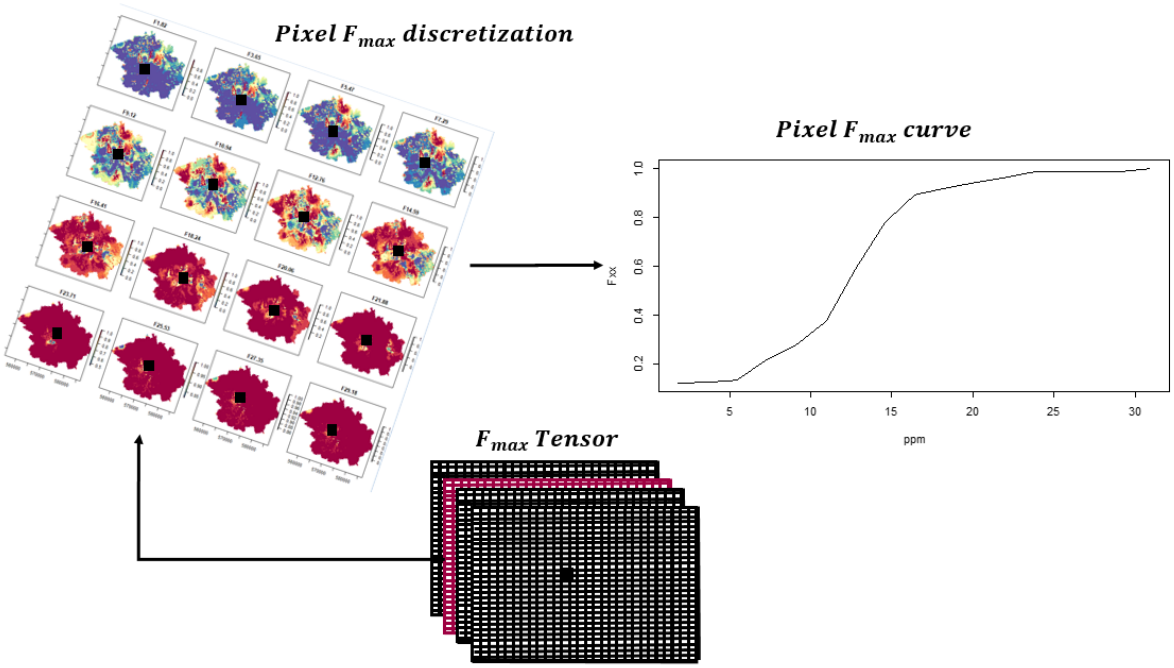


**Figure 65: Sample of Fmax tensor for arsenic values in the Toulouse ITA. The Fmax curve per pixel in the center of the city is extracted for further processing.**

Even if the $F_{min}$ and $F_{max}$ of such Monte Carlo draws cover the whole range of possibilities, they may be far apart in some SIC cases; in that event, it is better to opt for cross-section tensors $\boldsymbol{F}_{\alpha}^{+}$ and $\boldsymbol{F}_{\alpha}^{-}$ with $\alpha$ selected by an expert. Figure 66 shows the algorithm's output.





Funded by the European Union

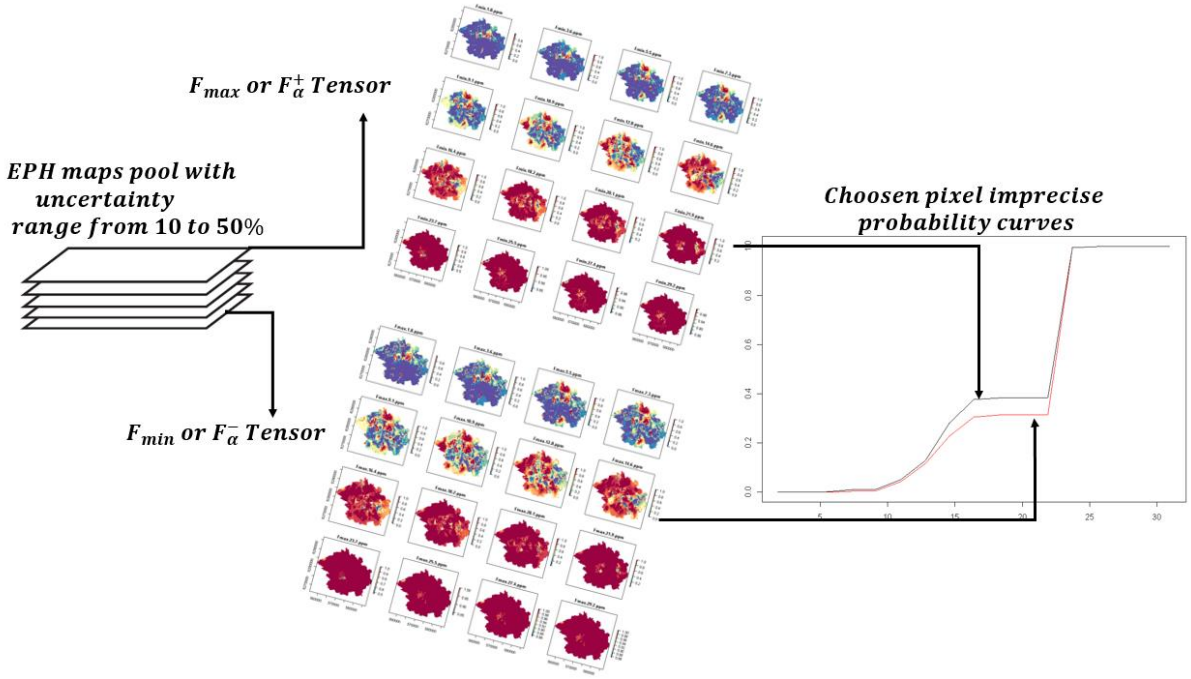


**Figure 66: Dual probability calculations implemented in ISLANDR calculus**

Exporting the two tensors ⟦ F⟧ _α^+ and ⟦ F⟧ _α^-, which constitute belief and plausibility, to a Hyrisk-type platform can then be performed using the export functions described in the previous chapter.

## 4.6.Anisotropy of phenomena

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 (Figure 67).

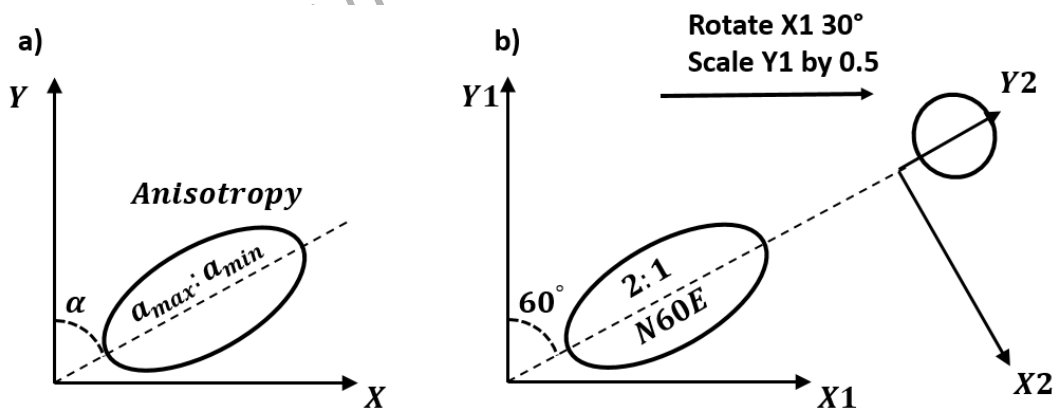


**Figure 67: Correction of geometric anisotropy**

*with a) notations and b) example of a 2:1 N60E anisotropy transformed into an isotropy.*

In 2D, the correction to be applied is as follows (Boisvert et al., 2009):



Funded by the European Union

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1/a_{max} & 0 \\ 0 & 1/a_{min} \end{bmatrix} \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

Using an anisotropy on our EPH of 13 sand data points allows us to recover the two possible geologist interpretations (Figure 67) on the Dahlberg (1975) dataset.

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 neighboring data and perform a local EPH. This is especially true in the case of groundwater, where flow lines change direction at soil permeability interfaces.

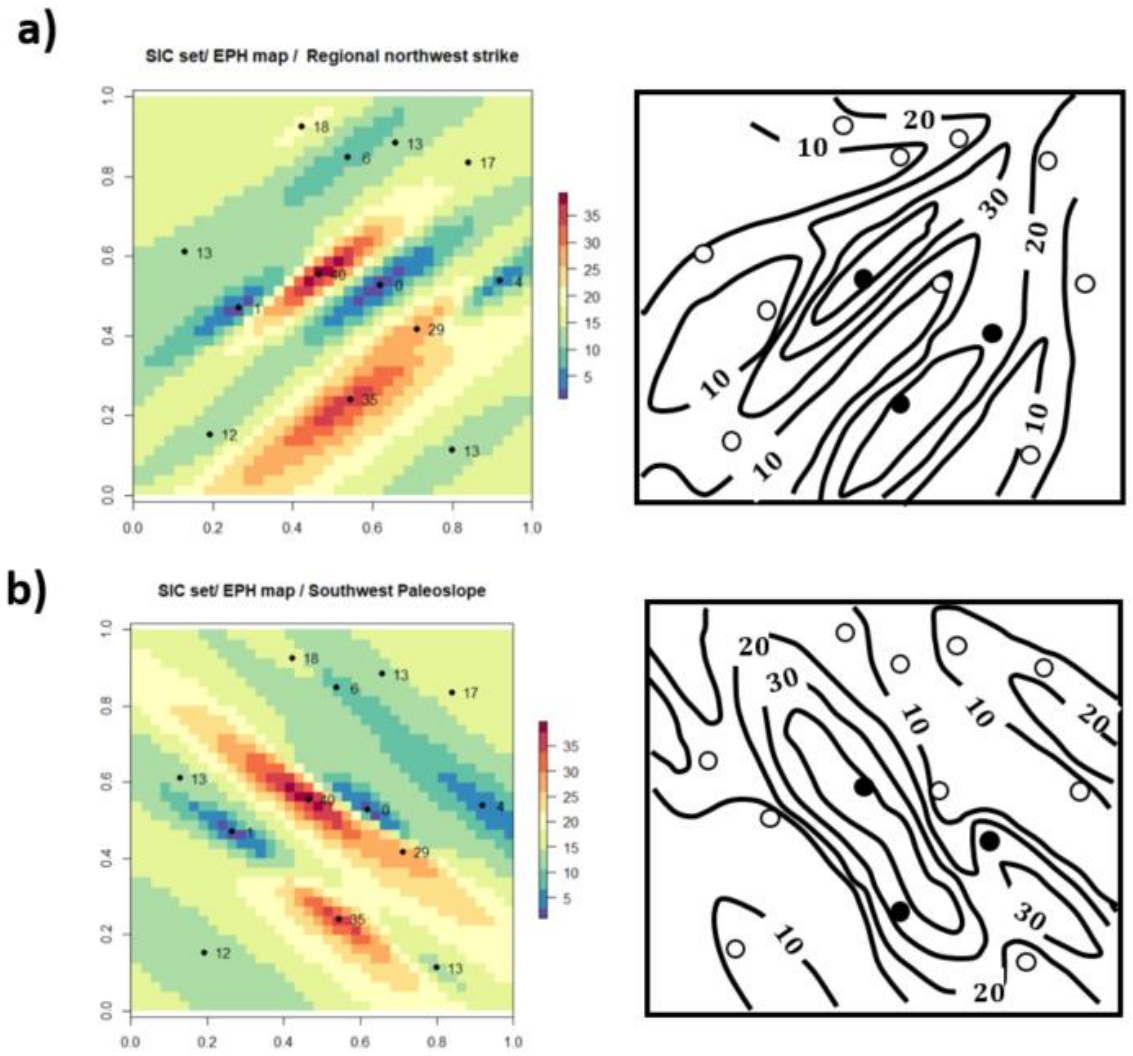


**Figure 68: 13 data points from Dahlberg (1975), corresponding EPH-generated map, hand-drawn map**

*with a) expected value map with 2:1 N60E anisotropy assuming channel sand model b) expected value map with 2:1 S60W anisotropy assuming a regional northwest strike with a southwest paleoslope.*



Funded by the European Union

## 4.7. Range of phenomena

The experimental variogram is the king of tools for studying the range of a physical phenomenon. It is a powerful tool that provides information on the behavior of our contaminants. Suppose there are two points $Z(x)$ and $Z(x+h)$ separated by distance $h$. The variogram is defined as the quantity:

$$2\gamma(x,h) = E(|Z(x) - Z(x+h)|^2) \approx \frac{1}{N_{(h)}} \sum_{i=1}^{N_{(h)}} |Z(x) - Z(x+h)|^2$$

With $N_{(h)}$ number of pairs of experimental points.
It is therefore a measurement of the spatial correlation between measurement points. The ranges measured on the variogram correspond to the notion of range of influence (Figure 69).

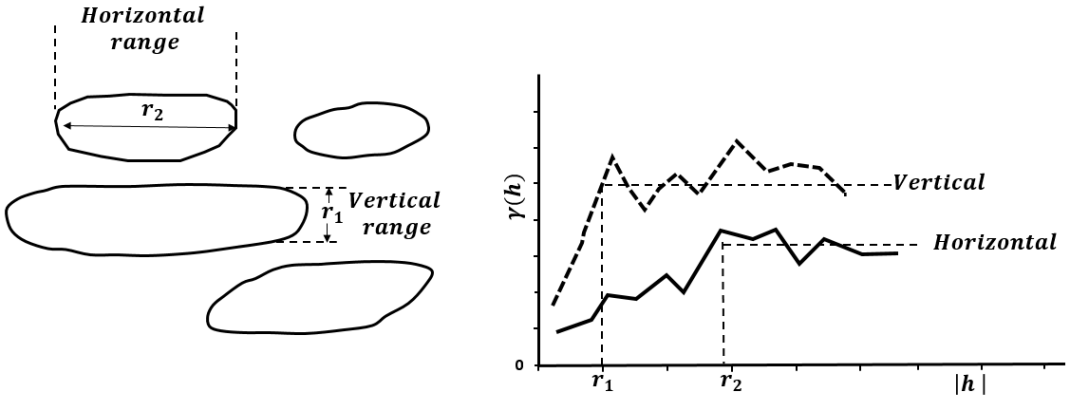


**Figure 69: Structural interpretation of variogram from Journel and Huijbregts (1978)**

A geostatistician will model this variogram and make multiple assumptions in order to produce a map, which we cannot do with SIC data. Nevertheless, the variogram provides us with information on the zones we can have, their average extension, and their anisotropy, and benefits from being calculated systematically on all the data or a group of data points thought to have a particular behavior (Figure 70)





Funded by the European Union

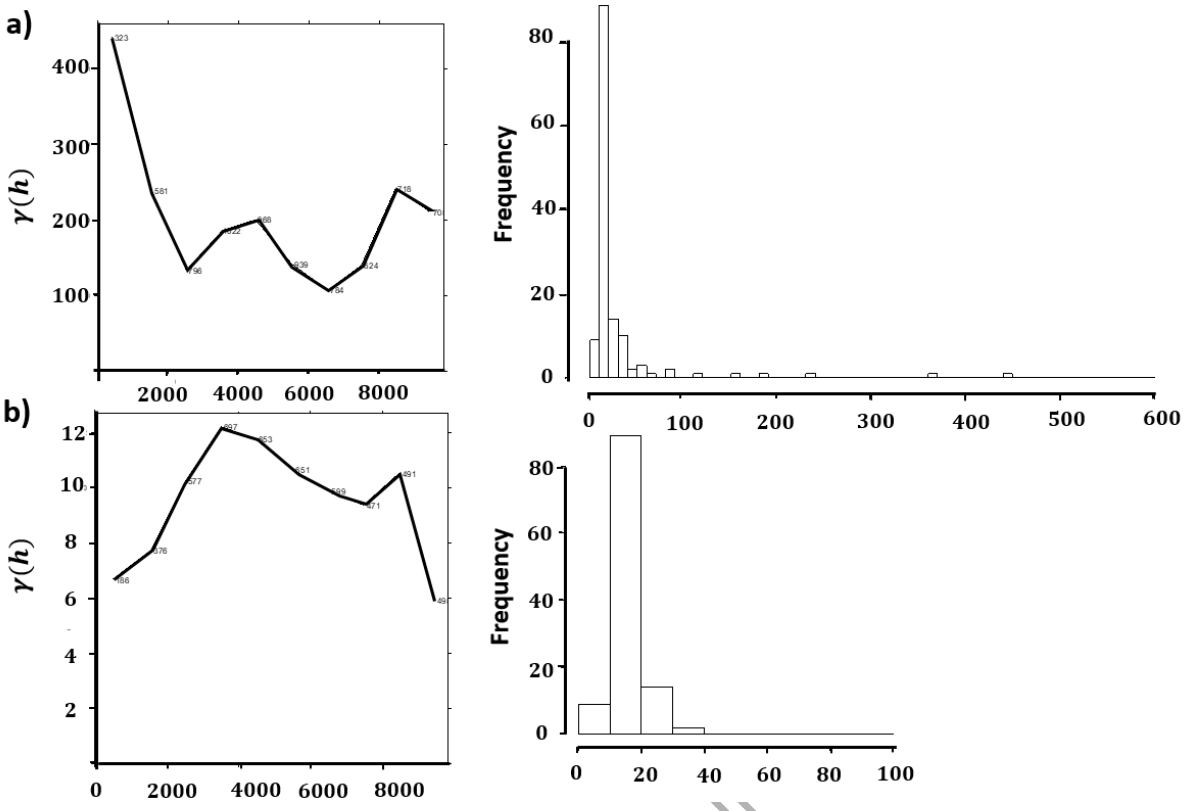


**Figure 70: Variogram showing a population structure on the Toulouse ITA**

*with a) variogram and histogram, Toulouse ITA TPH, 138 soil surface samples, b) variogram and histogram, Toulouse ITA TPH, 116 selected < 35 mg/kg soil surface samples.*

On the Toulouse ITA, the variographic study with the whole set showed no interesting structure. A variogram is very sensitive to anomalies and these must be removed, as explained in Chiles and Delfiner (2002). A specific population with levels below a 35 ppm kilometer range was then revealed (Belbeze et al., 2019). As a result, if it is established, for example, by expert opinion or by means of an experimental variogram that the neighborhood measurements no longer have any influence from an RoI distance, it is possible to modify this dmax quantity in the EPH, which adapts the slope of the entropy accordingly (Figure 71). Like with kriging, the map produced then takes into account a range of phenomena, but loses the neutrality that is the algorithm's strength. It should also be noted that the map produced is equivalent to Dahlberg's manual smoothing (Figure 72).





Funded by the European Union

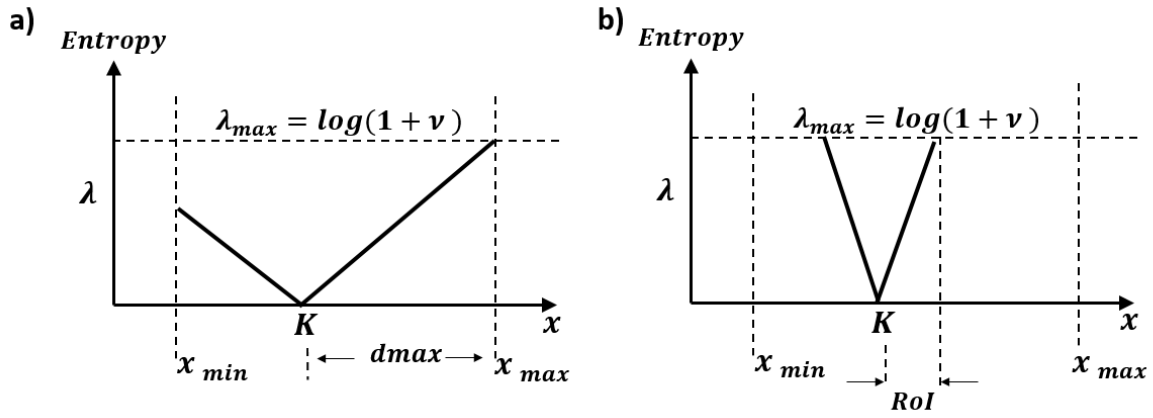


**Figure 71: Modified entropy growth with distance.**

Where RoI is the range of influence.

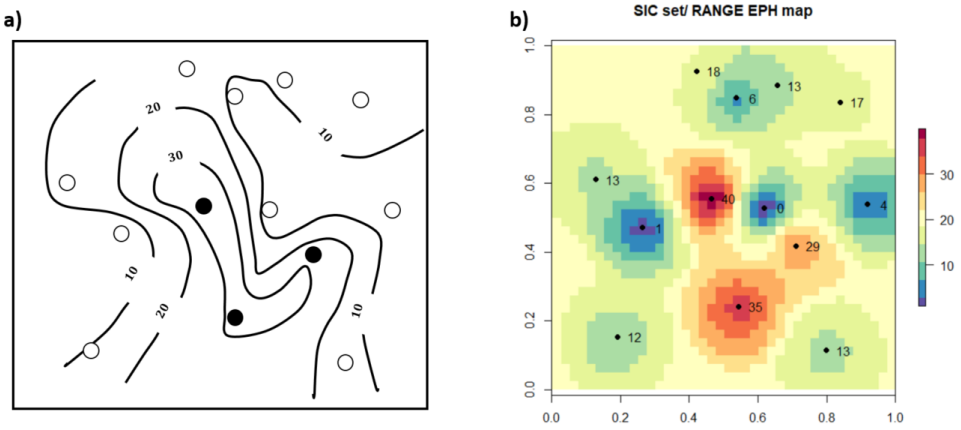


**Figure 72: : two maps of 13 data points from Dahlberg (1975),**

*with hand drawing (a) and corresponding EPH-generated map with b) expected value with altered range to 0.1 distance units.*

## 4.8.Data auto-variography

Another option for EPH 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 (Figure 73). This process is similar 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 on its own.





Funded by the European Union

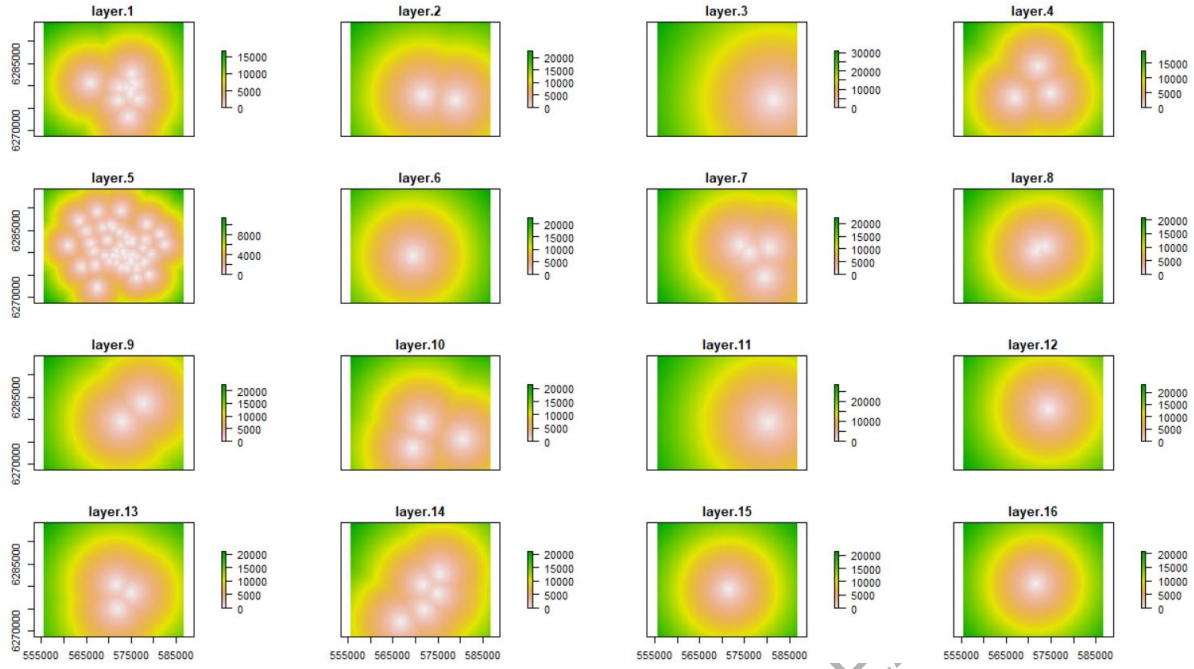


**Figure 73: Various Euclidean distances derived from classes of concentration. Toulouse ITA surface samples for TPH.**

For our EPH, distances are already accounted for once via the diffusion coefficient, but taking into account the differences between concentrations goes even further, revealing a statistical data structure comparable to geostatisticians' variograms.

This algorithm has been coded for our EEPH but, **like the variogram, it is highly dependent on the number of points (a minimum of 50–100 pts).** Therefore, when used on a good point density, this algorithm produced results comparable to kriging for an algorithm quantile random forest (QRF) (Hengl et al., 2018).

For SIC data such as that from Toulouse, the result appears less convincing, displaying characteristic circular bullseye profiles when compared with an EPH with optimized covariates (Figure 74). 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.





Funded by the European Union

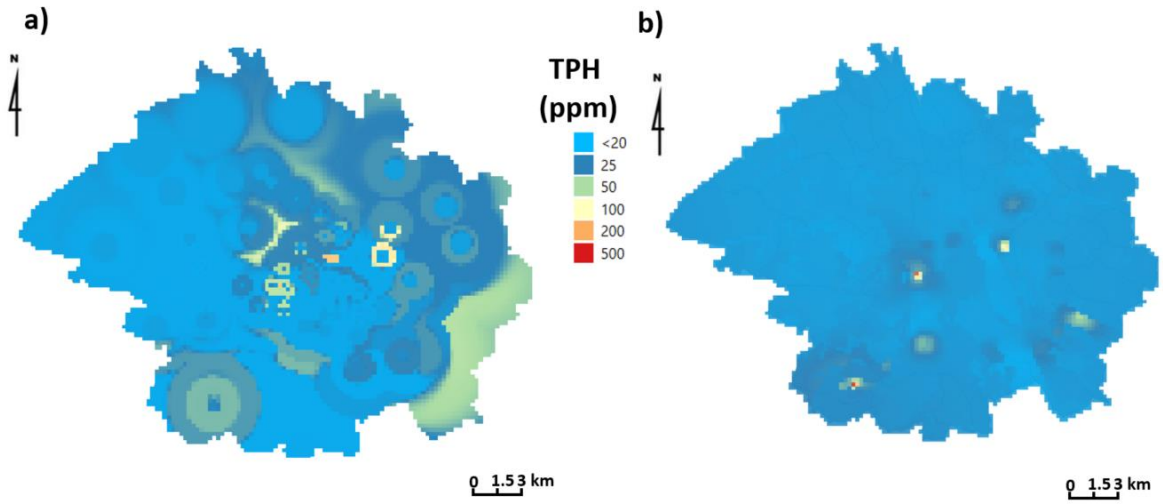


**Figure 74: Two interpolation methods for mapping total petroleum hydrocarbon (TPH) in the city of Toulouse**

*with a) EPH interpolation with Euclidean fields from classes of concentrations and b) EEPH interpolation with optimized LANU covariate.*

Topsoil samples from Belbeze et al. (2019), n=139, 0-10cm, TPH analysis by multiple laboratories, LOQ: 10 mg/kg (8 samples), 20 mg/kg (57 samples).

As we will see in EPH continuity processing, when the concentration variation is genetically linked to a known covariate such as land use, EPH captures the data structure on this covariate. There is no need for auto-variography. However, the EPH then loses what makes it so useful: the clear highlighting of data anomalies.

## 4.9.Establishing continuity for phenomena

Our SIC data does not itself carry continuity or even trend information; a strong epistemic assumption is attributed to them by the geostatistician or expert. To establish "continuity" in EPH, we need to introduce at least one parameter that conveys that continuity. If the main causal factor of our content is the geological nature of the subsoil, using this information as a parameter will restore a continuity of phenomena that could not be deduced from the data alone (Figure 75).

**Figure 75 : Images of the notion of continuity from plus to minus provided by a geological-type parameter (adapted from Sinclair and Blackwell, 2002).**

Thus, if the EPH of arsenic levels is calculated with geological information reduced to a 10 km grid as parameter, we obtain the map in Figure 76, which presents an improved continuity that is remarkably congruent with the anomaly marking carried out by Tarvainen et al. (2013).





Funded by
the European Union

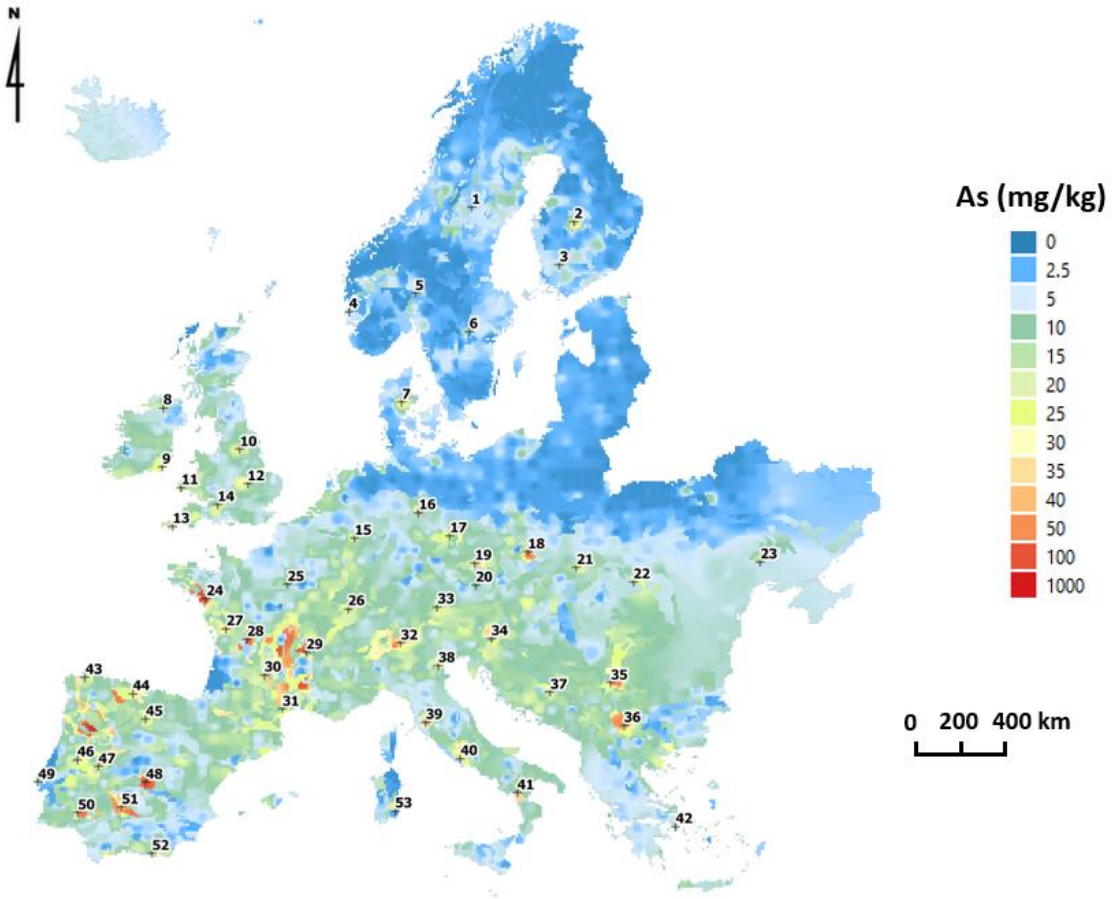


**Figure 76: Continuous expected value EEPH with geological covariable, Gemas Survey, Ap (0-20 cm), < 2mm, n=2217, 1 site/2500 km2, aqua regia, ICP-MS**

This map is definitely not as smooth as the usual maps, such as KED and GWR, but it is a map showing all the anomalies (no data has been discarded) and with minimized underlying assumptions. Geologists will be pleased that the shape of the anomalies matches the orientation of the layers and fractures.

## 4.10. Data clusters

While EPH is unaffected by outliers, it is algorithmically very sensitive to data clusters, which can bias its spatial probability calculation. As a countermeasure, a a CC bias correction weight $W_{clus}$ has been developed based on the Hclust 3.6.2 version of R Core Team (2022). To test this weight on difficult cluster data, a free dataset from GSLIB (Deutsch and Journel, 1997) was used. It consists of 140 survey data points where clusters have been produced around the highest magnitudes (Figure 77 a). The expected value calculated by EPH using this survey data loses its neutral character and wrongly displays a continuity linking the clusters (Figure 77 b). The modified EPH algorithm generates a weight $W_{clus}$ that enables the expected value by EPH to become “neutral” again, without the presupposition of continuity (Figure 77 c).





Funded by the European Union

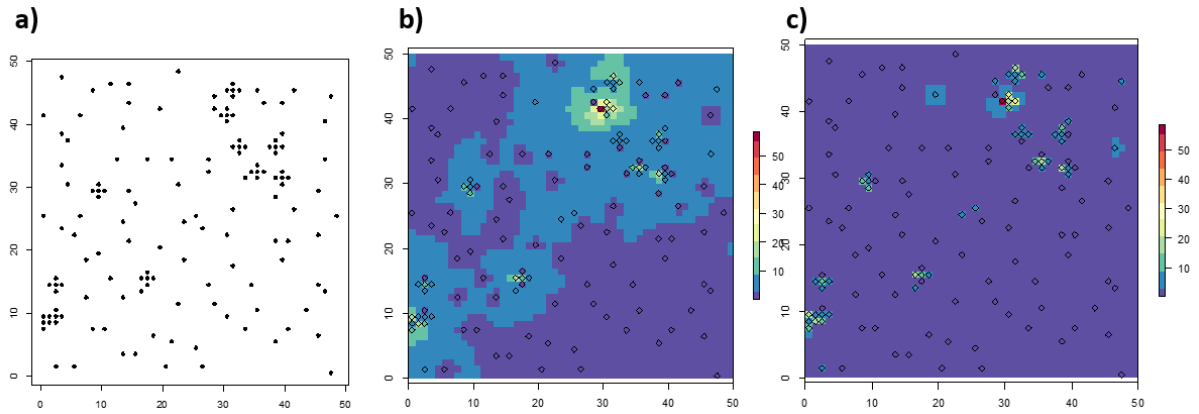


**Figure 77: Demonstration of the cluster correction from the Enhanced EPH algorithm for ISLANDR**

*with a) 140 data clustered sets from Deutsch and Journel (1997), b) resulting biased expected value calculated by EPH from the dataset, and c) declustered expected value calculated by EPH from the dataset.*

## 4.11. Sampling support

### 4.11.1. Physical bases

Taking the GEMAS example, a surface soil is a 2-2.5 kg sample taken 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. In the case of sampling of coherent zone or land unit, it will group together samples with the same proportions of factors that affect content. It should be noted that the notion of scale applies here, as it does to all mapping (Lindberg, 1994). This is because contaminant levels are subject to physical laws, the most important of which is 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.

This can be seen on small scales, such as soil samples where the contaminant variance is inversely proportional to the mass of soil sampled (Pitard, 2020). For a pool of samples taken from a coherent zone or land unit (LANU), the total variance $S^2$ of this pool breaks down as follows: the total variance $S^2$ = variance due to small-scale inhomogeneity + variance due to large-scale inhomogeneity + variance due to reduction and sub-sampling error (Visman et al.,1979).

$$S^2 = \frac{A}{W} + \frac{B}{N} + (SE)^2$$

With A , B, W, N, SE

on large scales, such as geostatistical mine grids, only the variance due to large-scale variations is modeled. By default, the variance due to reduction, under-sampling, and analysis is assumed to be low, and is referred to as the nugget effect. For these, if the content is z over a volume V:





Funded by the European Union

$$z(V) = \frac{1}{V}\int_V z(x)dx$$

For a region V decomposed into $v_i$disjoint regions (discrete case)

$$z(V) = \frac{\sum_i v_i z(v_i)}{\sum v_i}$$

For a region V decomposed into N equal disjoint regions (discrete case)

$$z(V) = \frac{\sum_i z(v_i)}{N}$$

Lastly, like with the small scale, this additivity translates into what is known as the variance dispersion formula. The variance of a small support is the sum of the variance of this small support (sample) in a mean $v_1$ (grid), plus the variance of the mean support in the entire V.

$$s^2(v|V) = \ s^2(v|v_1) + s^2(v_1|V)$$

A whole branch of geostatistics ("change of support") focuses on this subject and proposes solutions, such as transforming variables into a stationary space to perform calculations or adjusting the neighborhood and the model used (Chiles and Delfiner, 2013).

### 4.11.2. Applicable solutions

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. It mainly arises when merging two differently sampled campaigns. Figure 78 shows the effect of sampling support on content.

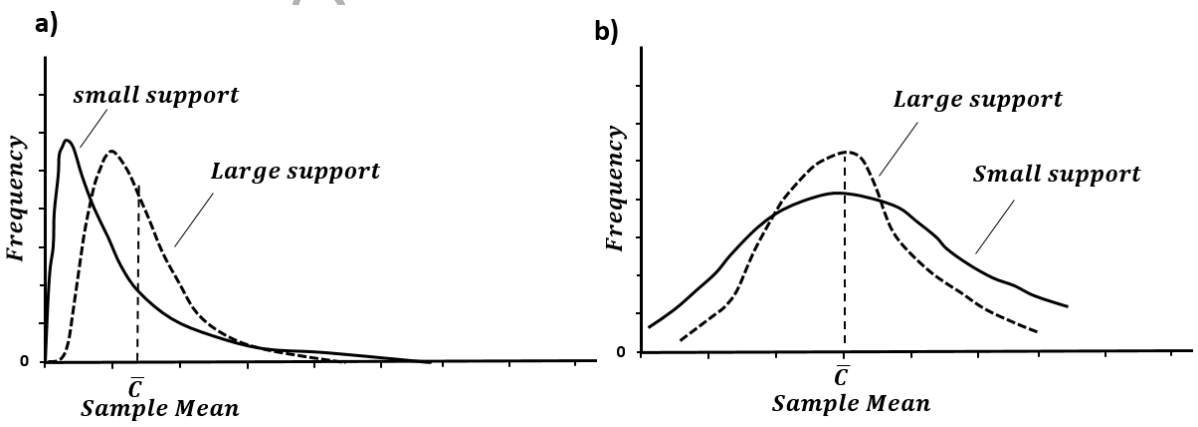


**Figure 78: Effect of support on content histograms with a) based on BRGM ore prospect data and b) based on Sinclair and Blackwell (2002).**

We see that although the mean remains fairly stable between supports, the variance decreases sharply.





Funded by the European Union

Geostatisticians are experts at changing support techniques, particularly when soil measurements are taken on a punctual support (such as soil cores or trenches), and the contents have to 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 (Chiles and Delfiner, 2013). These techniques require excellent variogram modeling and therefore a significant amount of data. Once the models have been calibrated, a theoretical punctual model is calculated; the calculations are performed before being re-transformed into output support. Figure 79 shows an exemplary example of such work by Kasmaeeyazdi et al. (2018).

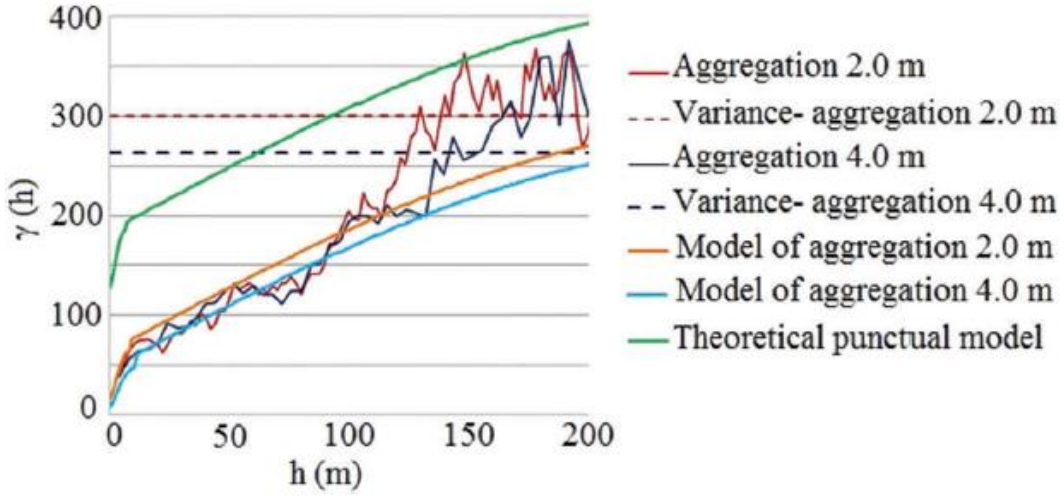


**Figure 79: Variogram models fitted on vertical sample variograms obtained from 2.0 and 4.0 m aggregation and the deduced theoretical punctual model. From Kasmaeeyazdi et al. (2018).**

If the ITA data allows it, this type of technique would be used. There is an R package called RTOP that is very useful for this kind of calculation (Skoien et al., 2014). But with SIC data, which is the very reason that EPH was developed, imposing a variance model on our data is not allowed, let alone deriving a theoretical point model from it.

However, there are some geostatistical solutions that we can adopt for EPH, such as compositing, regularization, and amalgamation procedures to obtain probabilities knowing the various supports to be calculated (Figure 80).





Funded by the European Union

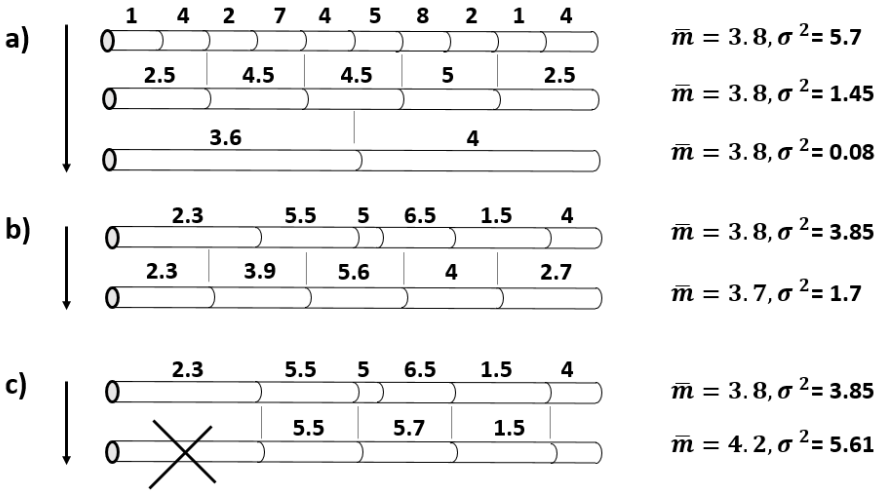


**Figure 80: Corrective actions for heterogeneous sample support**

with a) compositing, b) regularization c) amalgamation, data from Sinclair et Blackwell (2002). Amalgamation concept from Kasmaeeyazdi et al. (2018). Weighted mean and variance are shown for all three options.

Just a reminder: compositing involves gathering/mixing small samples to make a large one, to which a mean is assigned. Regularization involves assigning to a given size the weighted average of the samples it covers. This technique creates information, and should be used with caution. Amalgamation involves excluding samples that are too large, keeping samples that are the right size, and compositing smaller samples.

Several cases are possible, depending on the type of sampling encountered.

1. Sampling is of good quality, in line with mining standards between sampling campaigns. Supports are known and followed. This is based on strict equi-probabilistic sampling rules that guarantee additivity. In this case, content is corrected for its support to give accumulations, and maps are made for these before being switched back to map support.

Accum(x) = grade(x) * thickness(x) or Accum(x) = grade(x) * thickness(x) * density (x) if mineral density plays a role (like for gold).

Accum(x) and thickness(x) or Accum(x) and density(x) * thickness(x) are estimated using the same parameters and grids to maintain consistency. The block content is calculated by dividing the estimates.

The block content becomes $Block\ grade = \frac{\widetilde{\text{Accum}}(x)}{\widetilde{\text{thickness}}(x)}$ $ou$ $\frac{\widetilde{\text{Accum}}(x)}{\text{density}(x) * \widetilde{\text{thickness}}(x)}$

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 EPH equations.





Funded by the European Union

2. Sampling between sampling campaigns is 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 was 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 not the case, we assign to the smallest support a virtual value that it would have in support S2 associated with significant local error (Lajaunie, 1996). For our EPH, 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 (Figure 81). This interval will be taken uniformly and produce map Figure 82.

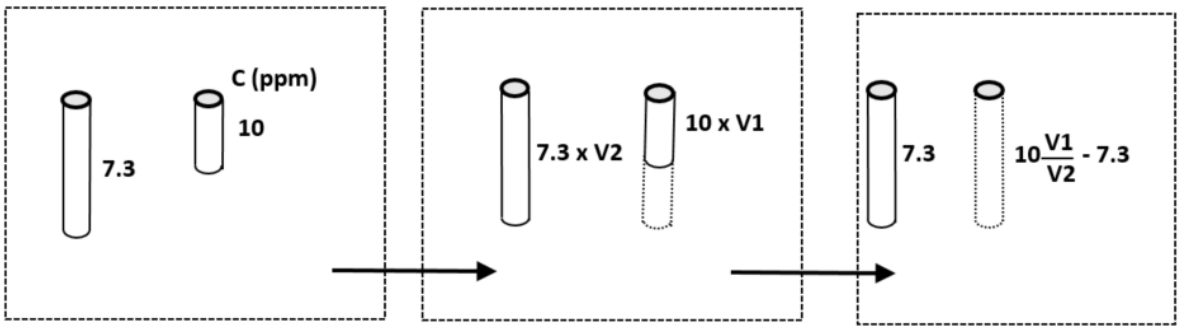


**Figure 81: Proposed discretization/regularization + weight approach to deal with heterogeneous support with medium-quality samples in ITA boreholes.**





Funded by the European Union

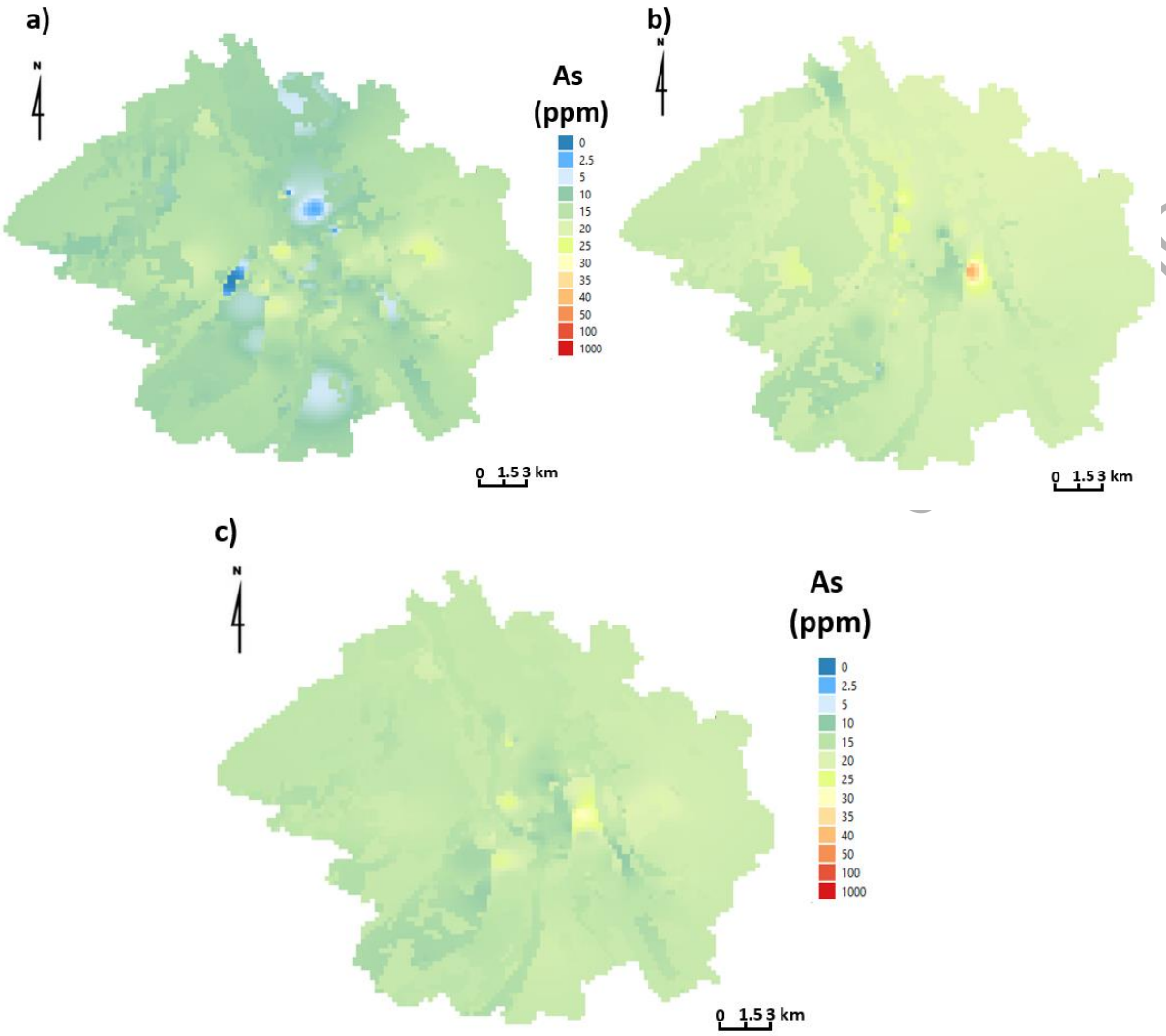


**Figure 82: As expected value map in soil 0-1m ±0.2, multi-support EPH calculus with geologic covariable (822 samples)**

*With a) surface soil sanmple only b) 1 meter sample only c) resulting multisupport calculus*

It may be noted that this mapping is consistent with the geochemical background of the area as calculated with conventional methods by Belbeze et al., 2019.

3. The sampling is of poor quality. Either the mass sampled is insufficient according to the Vismann equation, or the sampler has gleaned the soil. A composite will have to be made to restore correct averaging, which means degrading the resolution of the datasets to make them compatible. For example, by keeping only samples between 0.5 and 1 m. This may be too restrictive (Figure 83). An amalgamation (Kasmaeeyazdi et al. *2018)* of samples over 1 m with a tolerance of 0.2 m (Figure 84) is better.





Funded by
the European Union

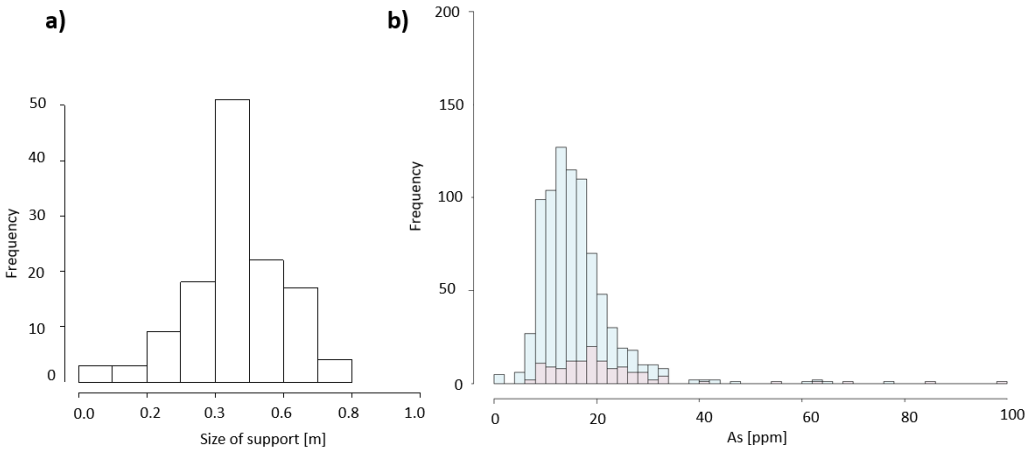


**Figure 83: Toulouse soil samples between 0 and 1 m depth, whole set (882 samples), selected [0 – 1] ±0.25 m set (127 samples). Toulouse ITA set**

*with a) size of support of the selected set 0.5–1 m ±0.25 cm and b) As histogram superposition for the whole set [0–1] ±0.25 cm and selected sets.*

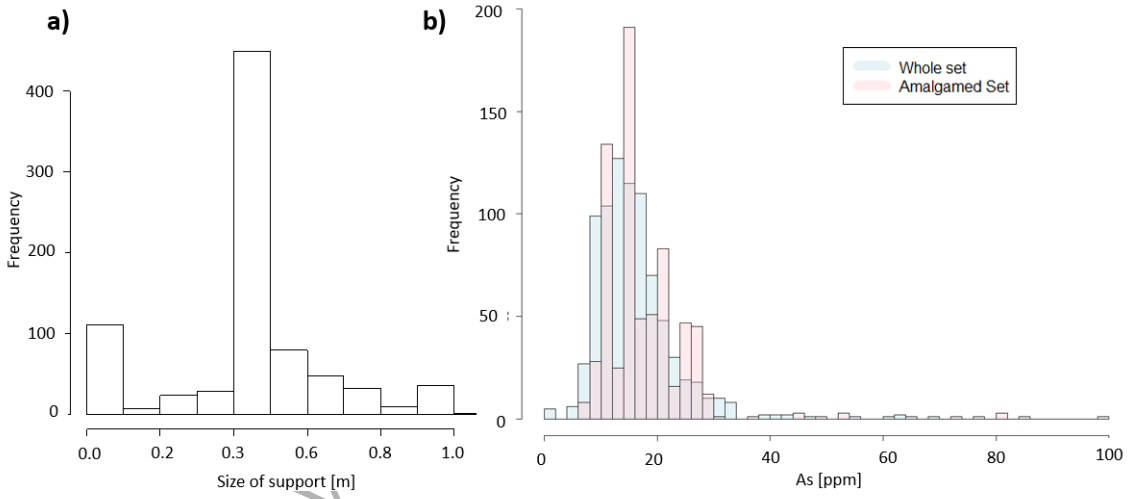


**Figure 84: Toulouse soil samples between 0 and 1 m depth, whole set (882 samples), amalgamed 1 m set (702 samples). Toulouse ITA set**

*with a) size of support of the whole set and b) As histogram superposition for the whole set [0–1] m ±0.25 cm and amalgamated 1 m set.*

Once the map has been obtained, restoring the additivity of supports between larger sample grids is simply a matter of convolving the map to the desired support. We apply the distribution:

$$s^2(v|V) = s^2(v|v_1) + s^2(v_1|V)$$

A good geologist will try to keep as close as possible to the conditions for equiprobable sampling. To achieve this, he will multiply the number of "elementary" samples (i.e. small samples, without having to go down to the fragment level) to make up the sample. The ppm content of this sample then approaches known mathematical distributions such as hyper-geometric, binomial, poisson...Poor-quality sampling can be detected by the double Poisson





Funded by the European Union

distribution, which can be calibrated to the data. If this phenomenon is not taken into account, it can have adverse consequences for the study, as shown by the example in Table 21 and Figure 85, taken from Ingamells and Pitard (1986). In this case, a dozen boreholes were drilled in a cobalt mine in the 1980s. The results appear to indicate that there is little cobalt in the panel boreholes, with the exception of S9 and S10. 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 as many environmentalists would tend to think, but a double Poisson distribution that always shows cyclicity (Figure 85).

**Table 21: Cobalt levels in 12 boreholes in a panel on a lateritic deposit from Ingamells and Pitard (1986)**

| Prof. (m) | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.03 | 0.1 | 1.07 | 0.64 | 0.34 | 0.14 | 0.09 | 0.16 | 0.21 | 0.2 | 0.28 | 0.22 |
| 2 | 0.07 | 0.2 | 0.16 | 0.24 | 0.2 | 0.24 | 0.25 | 0.36 | 0.73 | 2.42 | 0.81 | 0.53 |
| 3 | 0.02 | 0.02 | 0.03 | 0.41 | 0.31 | 0.46 | 0.29 | 0.33 | 0.28 | 0.41 | 0.35 | 0.11 |
| 4 | 0.09 | 0.04 | 0.04 | 0.03 | 0.09 | 0.08 | 0.09 | 0.12 | 0.5 | 0.28 | 0.09 | 0.47 |
| 5 | 0.02 | 0.03 | 0.05 | 0.28 | 0.23 | 0.33 | 1.01 | 0.17 | 0.1 | 0.07 | 0.03 | 0.08 |
| 6 | 0.11 | 0.22 | 0.21 | 0.24 | 0.21 | 0.2 | 0.2 | 0.2 | 0.21 | 0.18 | 0.14 | 0.13 |
| 7 | 0.05 | 0.04 | 0.04 | 0.03 | 0.03 | 0.04 | 0.04 | 0.03 | 0.05 | 0.1 | 0.16 | 0.12 |
| 8 | 0.02 | 0.02 | 0.01 | 0.03 | 0.01 | 0.02 | 0.06 | 0.05 | 0.08 | 0.17 | 0.35 | 0.28 |
| 9 | 0.02 | 0.02 | 0.03 | 0.03 | 0.05 | 0.03 | 0.02 | 0.03 | 0.03 | 0.08 | 0.09 | 0.05 |
| 10 | 0.02 | 0.02 | 0.03 | 0.02 | 0.08 | 0.14 | 0.12 | 0.3 | 1.34 | 1.04 | 0.5 | 0.27 |
| 11 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.07 | 0.12 | 0.16 | 0.3 | 0.43 |
| 12 | 0.2 | 0.26 | 0.17 | 0.12 | 0.12 | 0.1 | 0.22 | 0.23 | 0.27 | 0.29 | 0.22 | 0.18 |

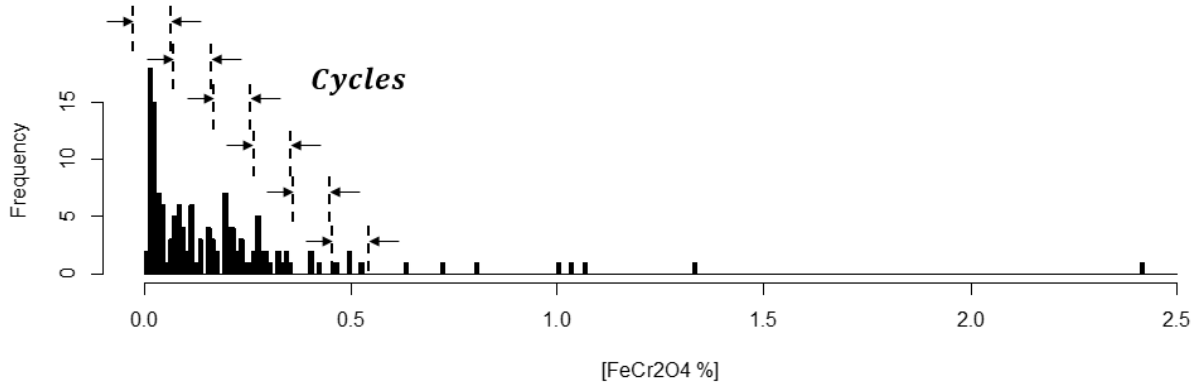


**Figure 85: Histogram of cobalt levels in 12 boreholes from laterite mineralization from Ingamells and Pitard (1986).**

*Note the apparent periodicity of the histogram curve, with a cycle every c= 0.1%.*





Funded by the European Union

The physical model can be deduced from our knowledge of probabilities as follows: Let's suppose that the contaminant or metal under investigation is present in the soil in the form of drops, slag, or grains. Let P(x = r) be the probability that a number r of these grains are in the sample. This probability, as demonstrated by Pitard (2020), follows a Poisson distribution:

$$P(x = r) = \frac{\theta^n}{r!} e^{-\theta}$$

With r = 1,2,3, etc.
If m is the number of samples in the composite, the variance in the sample is

$$\theta = mpq \approx mp$$

In other words, approximately equal to the mean.

If the sampling grid itself contains only a few clusters of these grains (μ on average), they will be sub-sampled by the five boreholes usually distributed in a cross pattern over the grid, each sampling $\theta$ on average. This type of sampling will cascade in a double Poisson distribution (the probability that r grains will be taken from a borehole or grid containing only n grains):

$$P(x = r) = \sum P(y = n) \cdot P(x = r) = \sum \frac{\mu^n e^{-\mu}}{n!} \cdot \frac{(nf)^r e^{-nf}}{r!}$$

$$P(x = r) = \frac{f^r e^{-\mu}}{r!} \sum_{n=0}^{\infty} \frac{\mu^n e^{-nf} n^r}{n!}$$

With r = 1,2,3, etc. And $f = \frac{\theta}{\mu}$

With the double Poisson, the variance may be higher or lower than the mean, as is the case with many anthropogenic pollution datasets. This is known as distribution dispersion (mean/variance). In practice (with the exception of diamondiferous surveys, for example), rather than measuring the number of grains in the sample, we measure a concentration. We then use C, the conversion factor between grain and concentration. This factor represents the grain's contribution.

Pitard (2020) proposes an iterative method for calculating these factors for a Poisson or double Poisson distribution. It uses the periodicity observed on the histogram and a concentration a_H that represents the geogenic background. For the example of cobalt ore, the author obtains the contribution of a cobalt grain to the panel composite: 0.1%. The solution to the issue of biased datasets has already been described above: make a composite or amalgamate the samples to find a representative central tendency.

Consequently, for ISLANDR, if ITA monitoring is high quality and extensive, so that the lengths sampled correspond to the lithological formations of the soil and the cores are sampled equiprobably, it will be possible to carry out accumulation calculations and thus





Funded by the European Union

integrate all samples of various supports into the EPH calculation. If ITA monitoring is of average or SIC quality, the interpolation algorithm can be run in multi-support mode to integrate all available information from disparate samples into the mapping. Finally, if sampling is of poor quality, we need to apply compositions or amalgamation to find a representative central tendency. If this cannot be obtained, the dataset must be abandoned.

## 4.12. Closure for multivariate calculations

Any content z is a compositional data point (ppm, mg/kg) that carries only relative information between the various constituents. Their sum is constant (1 kg, 1 million, etc.). Given 〖( x〗_1,x_2,...,x_p), the p elements measured in the sample (geochemists sometimes measure more than 50 in their studies) are compositional data points (usually in %, ppm, or mg/kg), we have

$$\sum_{i=1}^{p} x_i = 1$$

This condition is called closure (Aitchison, 1986); it introduces biases, particularly in the multivariate interpretation of correlations and PCA. Some authors use a new approach in geochemistry called CODA (COmpositional Data Analysis), developed since the 1990s by Aitchison (1986), Pawlowsky-Glahn and Buccianti (2011), and Pawlowsky-Glahn et al. (2015). Cooper and Caritat (2010) deal with Australian geochemistry, particularly mining prospects, Albanese et al. (2007). Petrik et al. (2018) used these techniques as part of the European EuroGeoSurvey GEMAS project, following protocols established by Reimann (2012).

To provide results congruent with the concepts of closure and advanced multivariate techniques, the R package algorithm features modules for normal score transformation, ILR, inverse ILR, PCA, and maximum autocorrelation function (MAF) calculations. However, the decision to transform data or not may bring additional uncertainties and must be decided by the panel of experts studying the ITA.

Spatial additivity and closure impose constraints that must be observed on our interpolation, particularly with regard to the scale of our maps. This can be achieved with upstream data transformations, such as ILR for closures in multi-element calculations, as well as with defuzzification constraints. For example, if we calculate an envelope or a grade density function for each v the defuzzification must maintain additivity for a volume V. To facilitate additivity in the case of square grid interpolation, we interpolate on a grid at least 4 times finer than the desired grid, which we then amalgamate to the correct resolution, which corresponds to Gaussian smoothing that guarantees additivity.

## 4.13. Comparing imprecise probabilities to a guide value





Funded by the European Union

A guide value T (Threshold) is needed to help decision making and communication with the public. When this is applied to a map with imprecise probabilities, we obtain a partition of the map into 3 zones presenting knowledge of the zone (>T, <T, U ). Starting from the two concentration planes sum $\overline{R}$ and $\underline{R}$ resulting from the imprecise probability calculation, if we apply the guide value to each pixel of these set to one if R ≥ T and 0 otherwise. Zones are merged by simple sum :

($\overline{R}$≥ T: 1, 0 otherwise ) + ( $\underline{R}$ ≥ T: 1, 0 otherwise).

If the limits T = 25 m and T = 30 m are applied to the Dahlberg (1975) data map, we obtain Figure 86.

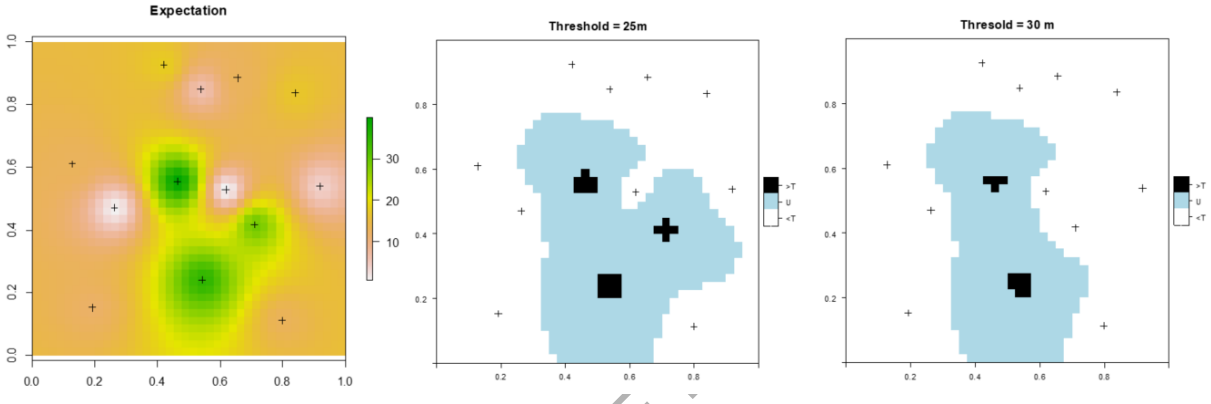


**Figure 86: Expectation of sand value, then public maps with threshold 25 m and 30 m, SIC dataset adapted from Dahlberg (1975).**

If the threshold itself is blurred, for example T= 25–30m, it is no longer possible to apply the preceding method. It is possible to use possibilities (Dubois and Prade, 1987), Dempster-Shafer formalism (Dempster, 2007), or neutrosophy (Smadaranche, 2006). To compare, according to neutrosophy, for an R content map with a limit value T, we calculate the pixel's exceedance of T neutrosophically through formulas that are widely used in brain imaging (Koundal et al., 2018 a and b; Ali et al., 2018; Haji and Yousif, 2019; Ahmed and Aboul, 2019; Anter and Hassenian, 2019; Özyurt et al., 2019; Salama et al., 2019; Sert and Avci, 2019; Wady et al., 2020).

To do this, we associate an index of truth (t), uncertainty (i) and falsity (f) with the value of each pixel $P$ considered with the guide value T.

$$P_{NS>T}(i,j) = \{t(i,j), i(i,j), f(i,j)\}$$

$$t(i,j) = \frac{\overline{g}_{(i,j)}}{\overline{g}_{\max} - \overline{g}_{\min}}$$

$$i(i,j) = \frac{\delta_{(i,j)} - \delta_{min}}{\delta_{max} - \delta_{min}}$$

$$f(i,j) = 1 - T_{(i,j)}$$

$$g_{(i,j)} = R_{(i,j)}/T$$





Funded by the European Union

$$\overline{g}_{(i,j)} = \frac{1}{A^2} \sum_{m=i-a/2}^{i+a/2} \sum_{n=j-a/2}^{j+a/2} g_{(m,n)}$$

$$\delta_{(i,j)} = \left| g_{(i,j)} - \overline{g}_{(i,j)} \right|$$

On such numbers, addition or union is performed as follows:

$$P_1 \oplus P_2 = \{ t_1 + t_2 - t_1 t_2 , i_1 i_2, f_1 f_2 \}$$

Multiplication or intersection:

$$P_1 \otimes P_2 = \{ t_1 t_2 , i_1 + i_2 - i_1 i_2, f_1 + f_2 - f_1 f_2 \}$$

Applied to our imprecise probabilities, by crossing plausibility (lo) and belief (hi) with 2 guide values, we obtain: 4 datasets of three rasters (T, I, F) as shown in Figure 87.

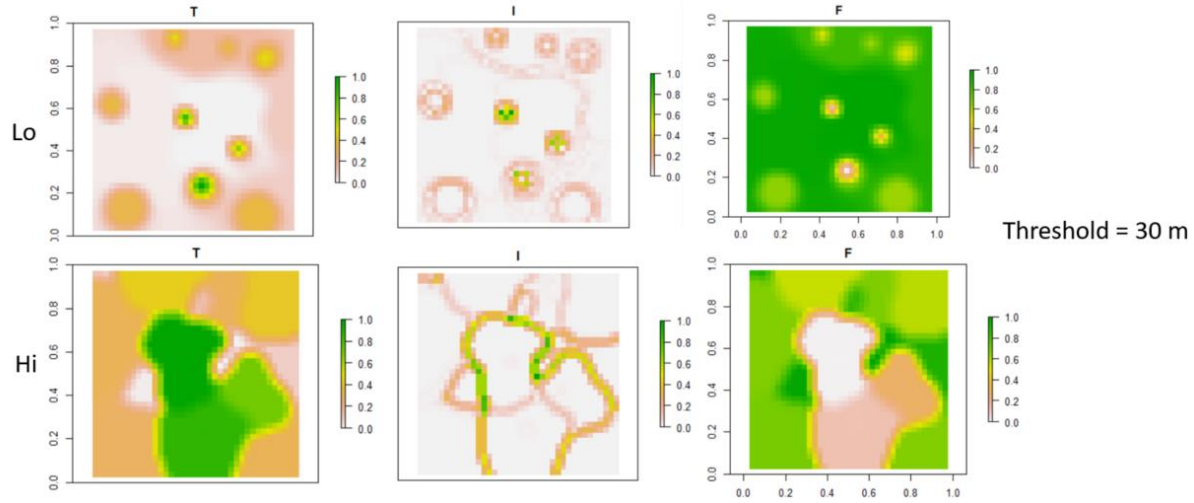


**Figure 87: Neutrosophic raster of the treatment of our imprecise levels using a guide value T = 30, SIC dataset adapted from Dahlberg (1975).**

Zones are merged for a given guide value by the simple sum $R < T_1 = (\overline{R} \geq \mathrm{T}_1 : < \mathrm{T}, \mathrm{I}, \mathrm{F} > ) + (\underline{R} \geq \mathrm{T}_2 : < \mathrm{T}, \mathrm{I}, \mathrm{F} > )$. To obtain the intersection of $R > T_1$ and $R > T_2$ we multiply them (Figure 88). We then note that the maximum envelope of T is indeed the envelope of application for small Threshold 25 m as in Figure 88 and that the fuzzy threshold value is well diffused on the truth plane T with probabilistic degrees (Bera and Mahapatra, 2017).





Funded by the European Union

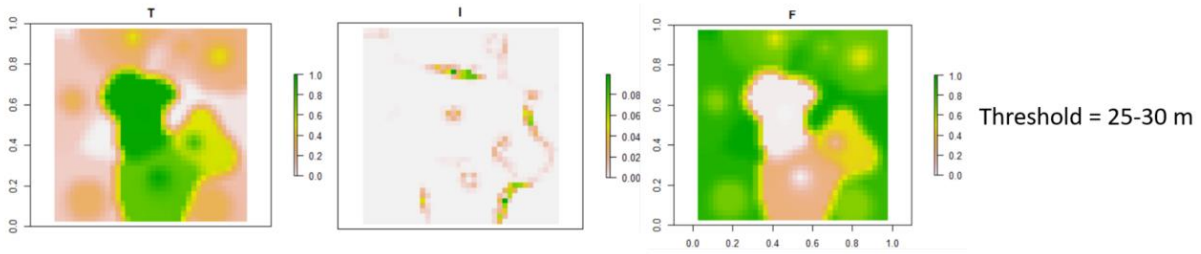


**Figure 88: Levels from Dahlberg's dataset with a threshold of 25–30m, SIC dataset adapted from Dahlberg (1975).**

Results in spectral colors (Figre 89) may appeal to the public. Of course, we must address the uncertainty of the green zones.

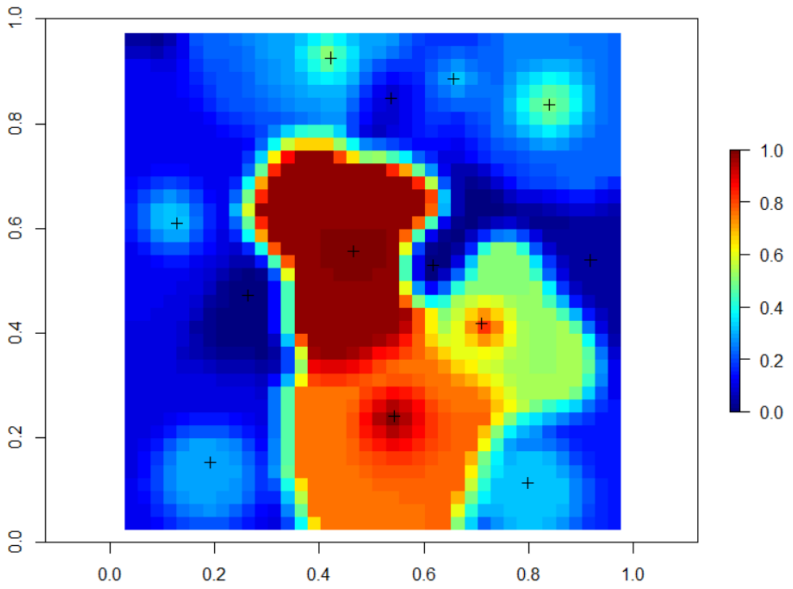

**Figure 89: Soil likely to exceed a threshold of 25–30 m, SIC dataset adapted from Dahlberg (1975).**

## 4.14. Contradictory data

Two Environmental Consultancies (EC1 and EC2) find contamination, but one found more than double the other did (u% = 100). In most cases, this is due to sampling bias (layers are not sampled in the same way, losses have occurred during transport, etc.). Since we are arriving after the fact, it is impossible to consult the distribution or calibration error, as we would for a sensor. Since EPH is a probabilistic tool, it is easy to switch to imprecise probabilities. To do so, we need to:

- Calculate a probability tensor by EC, then convert these tensors into CFD tensors.
- Construct the two min/max CFD tensors by pixel (which are simply plausibility and belief).
- Calculate imprecise probabilities in the same way that EPH calculates uncertainties on an explanatory variable or parameter.





Funded by the European Union

We thus obtain the uncertainty on the cartographic analysis of the situation, taking into account the fact that the consultancies did not measure the same thing in the same place (Figure 90

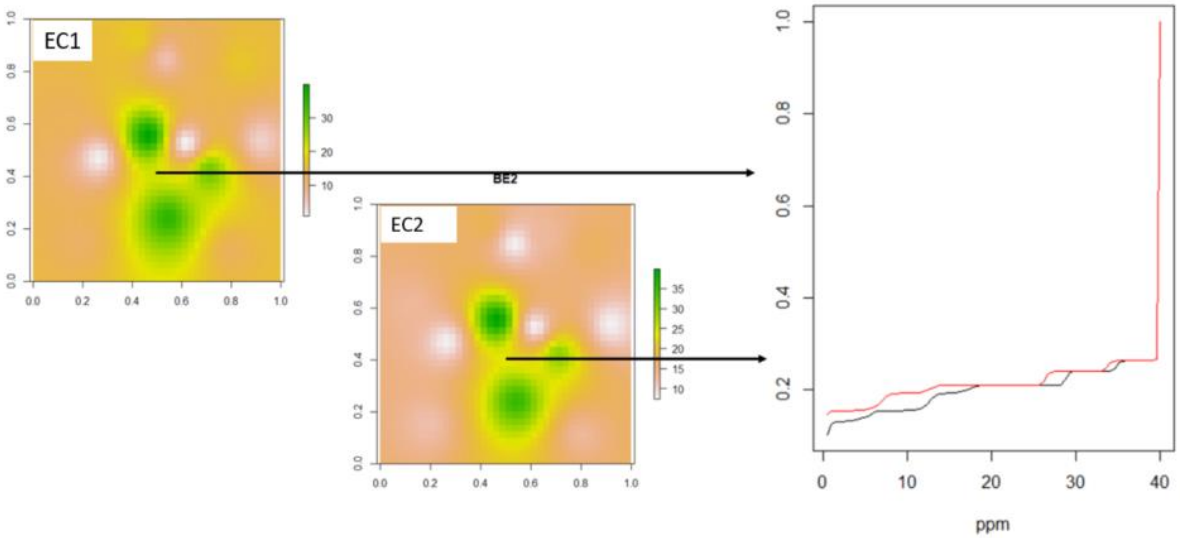


**Figure 90: Imprecise dual probability resulting for two different sets of measurements**

As appealing as it is, this technique would be flawed if there were total disagreement between the two consultancies. We would then have to introduce the notion that everything and its opposite is possible. This means moving on to possibilities (Dubois and Prade, 1987), Dempster-Shafer formalism (Dempster, 2007), or neutrosophy (Smadaranche, 2006). A neutrosophic technique, for example, would be to create as many Pxx maps as there are consultancies, and to generate their possibilities. We could then present the merging operation schematically as an intersection of these possibilities per pixel with one classified as a reference (the one with the most precise quality or measurements).





Funded by the European Union

## 4.15. Situating EEPH in relation to other interpolation systems

### 4.15.1. The Spatial 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 difficult intercomparison datasets such as SIC2004 (Dubois et al., 2005). This test (Figure 91) 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, joker dataset). Various more or less automatic techniques were then tried 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 situating our method because:

- It requires fast, neutral interpolation without covariates,
- The joker dataset has a radiological background (task 1.2), it is a natural dataset, and it is not derived from a model, so it is likely to favor methods that would be related to this process.
- The joker dataset presents an anomaly based on this background, which, from an ISLANDR perspective, must be identified (objective task 1.3). The test as carried out at the time focused only on the restitution of the mean in the presence of the outlier.
- Among these authors, we note two methods related to our EEPH: one that uses deviations between measurements instead of EEPH distances (Fang, 2005) and another with a neural network equipped with a Nadaraya-Watson kernel (Timonin and Savalieva, 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 favor EEPH (it only delivers its full power if n<50), but they do highlight its role as an 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 (Figure 91) and 1 s for Dataset 2 with anomaly detection Figure 92) on a portable P5. The formal results (Table 22) for Dataset 1 give us a score of 26, very close to the 24–25 obtained by experts in geostatistical modeling (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 mean and amplify just enough to define anomalies (Figure 93).





Funded by the European Union

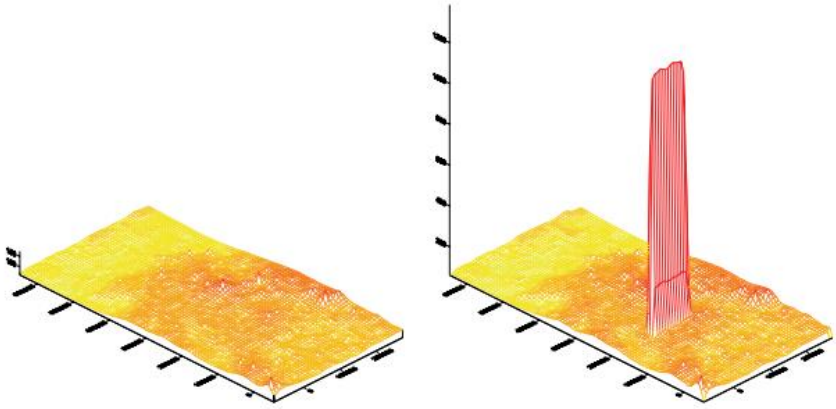

**Figure 91: Datasets 1 and 2 of the exercise, SIC2004, based on Dubois 2005 and Dubois and Galmarini (2005)**

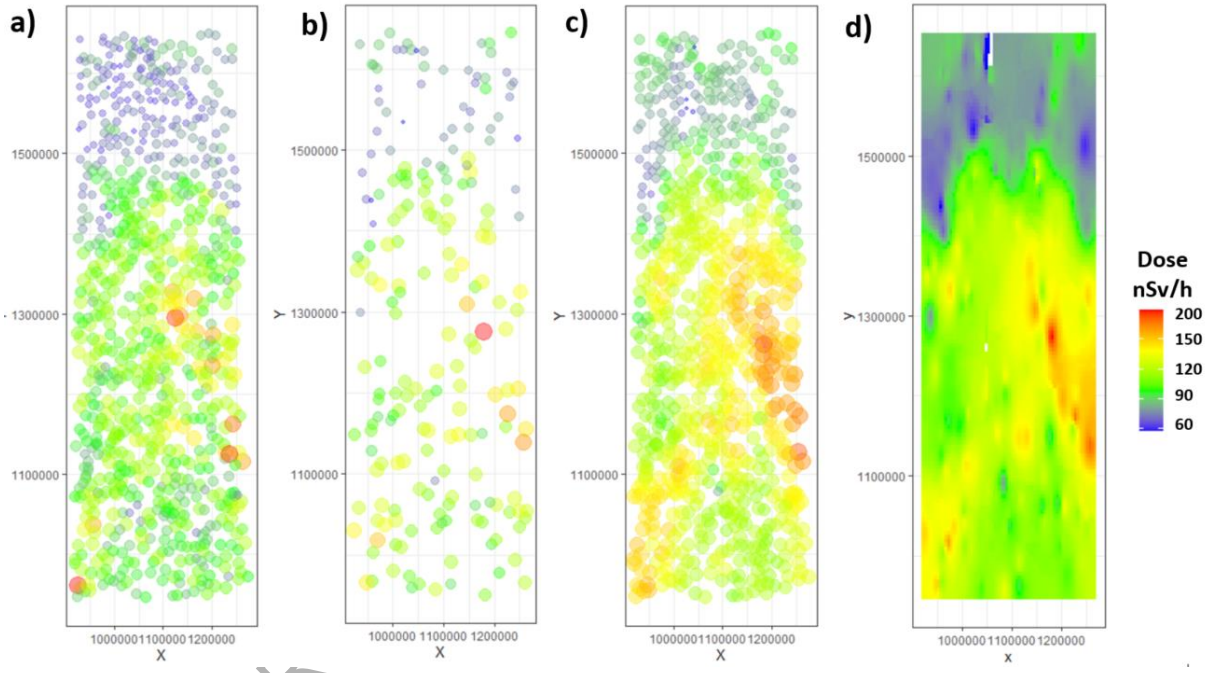


**Figure 92: EEPH tests on Dataset 1 from the SIC2004 exercises obtained in 0.15s (dataset from Dubois and Galmarini, 2005).**

*With a) true set 1, b) training set 1, c) EEPH neutral results on out points, and d) EEPH raster*





Funded by
the European Union

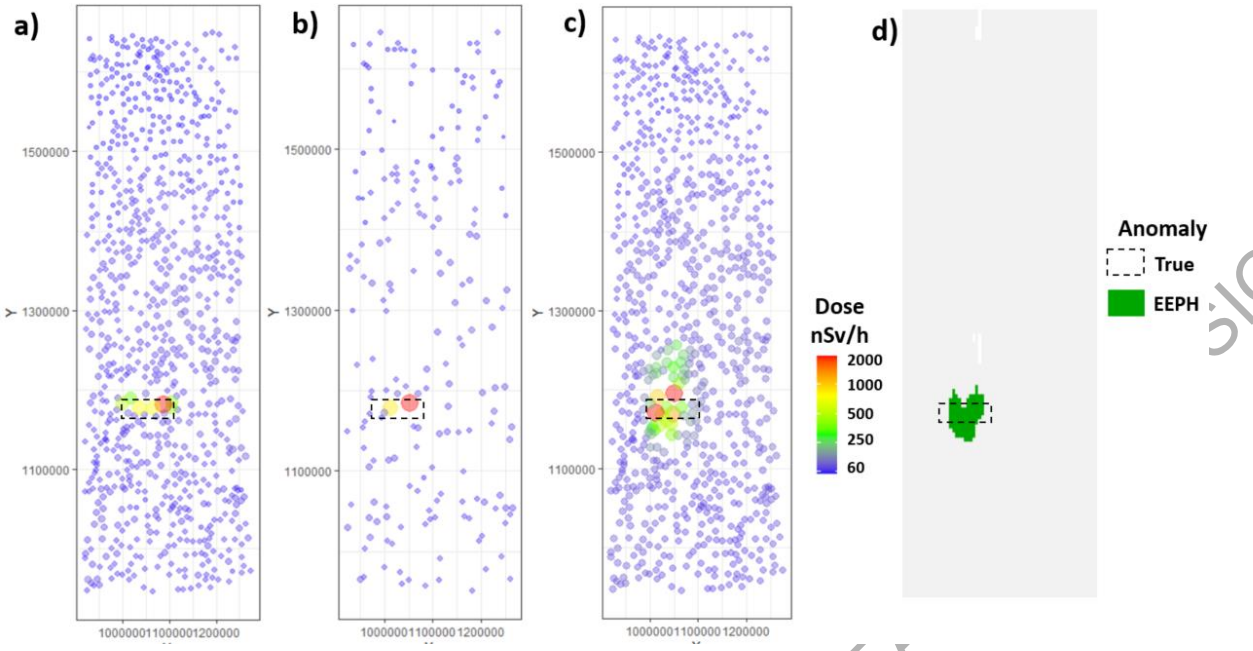


**Figure 93: EEPH tests on Dataset 2 "joker set" from the SIC2004 exercises obtained in 1 (dataset from Dubois and Galmarini, 2005).**

*With a) true set 1, b) training set 1, c) EEPH neutral results on out points, and d) anomaly detection*





Funded by the European Union

**Table 22: Results of SIC2004 exercises with the addition of EEPH results, adapted from Dubois G, Galmarini, S., 2005**

MAE = mean absolute error, ME = mean error or bias, RMSE = root mean squared error, r = Pearson's correlation coefficient r between true and estimated ones. GEOSTATS denotes geostatistical techniques, NNW refers to neural networks, and SVM is support vector machine. Skill score = MAE + abs(ME) + RMSE + 10*(1 - Pearson's r). Results are sorted from the MAE and the skill 1 results. Bolded results are obtained by the author in less than a second. Dark gray is the overall winner of the test with a GRNN and light gray is the winner of test 1 with a SVM.

| First authors | Method | Classification (Dubois & Galmarini 2005) | Expert action | MAE (1) | MAE (2) | ME (1) | ME (2) | RMSE (1) | RMSE (2) | r (1) | r (2) | skill1 | skill2 | overall skill |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Palaseanu-Lovejoy | GEOSTATS | Bayesian Kriging | Specify the prior distributions | 9.05 | 19.76 | 1.4 | 2.33 | 12.46 | 74.54 | 0.79 | 0.5 | 25 | 102 | 127 |
| Fournier | GEOSTATS | Robust Kriging | Structural modeling | 9.06 | 16.22 | -1.32 | -8.58 | 12.43 | 81.44 | 0.79 | 0.27 | 25 | 114 | 138 |
| Ingram | GEOSTATS | Gaussian process | Set hyperparameters | 9.08 AM | 21.77 | -1.44 | 0.72 | 12.47 | 79.57 | 0.79 | 0.35 | 25 | 109 | 134 |
| Ingram | GEOSTATS | Gaussian process | Set hyperparameters | 9.1 | 18.55 | -1.27 | -4.64 | 12.46 | 54.22 | 0.79 | 0.86 | 25 | 79 | 104 |
| Hofierka | GEOSTATS | R. splines with tension | Set hyperparameters | 9.1 | 18.62 | -1.3 | 0.41 | 12.51 | 73.68 | 0.79 | 0.5 | 25 | 98 | 123 |
| Savelieva | GEOSTATS | Ordinary kriging (OK) | Structural modeling | 9.11 | 19.68 | -1.39 | -2.18 | 12.49 | 69.08 | 0.78 | 0.56 | 25 | 95 | 121 |
| Pebesma | GEOSTATS | Ordinary kriging | Structural modeling | 9.11 | 20.83 | -1.22 | 0.92 | 12.44 | 73.73 | 0.79 | 0.5 | 25 | 100 | 125 |
| Pebesma | GEOSTATS | Ordinary kriging (OK) | Structural modeling | 9.11 | 23.26 | -1.22 | 4 | 12.44 | 76.19 | 0.79 | 0.42 | 25 | 109 | 134 |
| Pebesma | GEOSTATS | Auto ordinary kriging | Structural modeling | 9.11 | 146.4 | -1.22 | 19.71 | 12.44 | 212.1 | 0.79 | -0.27 | 25 | 391 | 416 |
| Fournier | GEOSTATS | Robust Kriging | Structural modeling | 9.22 | 19.43 | -0.89 | -0.22 | 12.51 | 73.5 | 0.78 | 0.48 | 25 | 98 | 123 |
| Pozdnoukhov | SVM | Support vector machine | Set hyperparameters | 9.22 | 16.25 | -0.04 | -6.7 | 12.47 | 81 | 0.79 | 0.28 | 24 | 111 | 135 |
| Fournier | GEOSTATS | Robust Kriging | Set hyperparameters | 9.29 | 19.44 | -1.12 | -0.12 | 12.56 | 71.87 | 0.78 | 0.53 | 25 | 96 | 121 |





Funded by the European Union

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Saveliev | GEOSTATS | Multilevel B-spline (MBA) | Set hyperparameters | 9.3 | 22.2 | 1.6 | 0.6 | 12.6 | 76.4 | 0.78 | 0.41 | 26 | 105 | 131 |
| **Belbeze** | **OTHERS** | **Experimental probabilistic hypersurface** | **None** | **9.35** | **23.58** | **-1.41** | **3.61** | **12.79** | **86.37** | **0.77** | **0.33** | **26** | **120** | **146** |
| Hofierka | GEOSTATS | R. splines with tension with CV parameters | None | 9.38 | 26.52 | -1.27 | 4.29 | 12.68 | 77.98 | 0.78 | 0.38 | 26 | 115 | 141 |
| Timonin | NNW | GRNN | Set hyperparameters | 9.4 | 14.85 | -1.25 | -0.51 | 12.59 | 45.46 | 0.78 | 0.84 | 25 | 62 | 88 |
| Ingram | NNW | MLP | Structure of network | 9.47 | 22.53 | -1.15 | 3.09 | 12.75 | 79.16 | 0.78 | 0.33 | 26 | 111 | 137 |
| Ingram | NNW | Gaussian process 2 | None | 9.48 | 48.41 | -1.22 | -3.01 | 12.73 | 90.89 | 0.78 | 0.38 | 26 | 149 | 174 |
| Saveliev | GEOSTATS | Kriging with drift | Structural modeling | 9.6 | 17 | 3 | 10.4 | 13 | 82.2 | 0.77 | 0.23 | 28 | 117 | 145 |
| Dutta | NN | RBF (GA division + unsupervised training ) | Structure of network | 9.62 | 28.2 | 0.9 | -0.22 | 12.7 | 80.1 | 0.78 | 0.31 | 25 | 115 | 141 |
| Fang | *OTHERS* | *Dirac Monte Carlo method* | None | 9.67 | 19.91 | -1.29 | 3.26 | 13.21 | 66.8 | 0.75 | 0.61 | 27 | 94 | 121 |
| Lophaven | GEOSTATS | Universal kriging | Structural modeling | 9.7 | 22.2 | 1.2 | -4.1 | 13.1 | 71.2 | 0.76 | 0.54 | 26 | 102 | 129 |
| Ingram | NNW | Radial basis function | Structure of network | 9.72 | 38.29 | -1.54 | 8.38 | 13 | 84.24 | 0.76 | 0.3 | 27 | 138 | 165 |
| Dutta | NNW | RBF model (random training) | Structure of network | 9.92 | 17.5 | 0.2 | 5.1 | 13.1 | 80.6 | 0.76 | 0.29 | 26 | 110 | 136 |
| Dutta | NNW | Multilayer feed forward network (MFFN) | Structure of network | 9.93 | 38.5 | 2.18 | 17.98 | 13.3 | 87.3 | 0.76 | 0.27 | 28 | 151 | 179 |
| Pebesma | *OTHERS* | Inv dist weight (pow 2) | None | 9.94 | 21.03 | -1.35 | 4.5 | 13.32 | 72.12 | 0.78 | 0.51 | 27 | 103 | 129 |
| Rigol-Sanches | NNW | ANN b-p (add training) | Structure of network | 12.1 | 20.3 | -1.2 | -9.4 | 15.8 | 84.1 | 0.67 | 0.12 | 32 | 123 | 155 |





Funded by the European Union

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dutta | NNW | RBF (GA division + supervised training) | Set hyperparameters | 12.2 | 28.9 | 1.5 | -1.29 | 15.9 | 79.9 | 0.64 | 0.33 | 33 | 117 | 150 |
| Rigol-Sanches | NNW | ANN back propagation | Structure of network | 16 | 25.3 | -1.7 | -11.1 | 20.8 | 87.5 | 0.55 | 0.02 | 43 | 134 | 177 |
| Rigol-Sanches | NNW | ANN back propagation | Structure of network | 21.4 | 30.5 | 5.3 | 3.8 | 45.8 | 96.6 | 0.24 | 0.2 | 80 | 139 | 219 |





Funded by the European Union

### 4.15.2. The HOUSES test

ANR Houses (Harmonized Operation of Uncertainties in Spatialized Environmental Systems) is a project of the French national research agency (ANR) aimed at developing innovative algorithms for digital mapping. The avenues explored include mainly possibilistic and Bayesian mathematics, rather than information diffusion as developed specifically for ISLANDR. With the agreement of their project leader, we decided to process their dataset to assess the performance of our SIC data processing algorithm.

The dataset consists of an interpolation of the OCS variable on the 500m x 500m grid of a truth plane, in the form of 5 experiments to be conducted (Figure 94) with and without 14 covariates Figure95, Table 23) for a total of 12 500x500 maps to be produced.

**Table 23: ANR HOUSES sparse, imprecise, and clustered dataset**

| Dataset | Number of points |
|---|---|
| Truth data | 27430 pts on a 500x500 raster map |
| Covariables | 14 covariables on a 500x500 raster map |
| Reference clustered case: Clustered | 200 pts |
| Sparse clustered case 1: Sparse 1 | 100 pts |
| Sparse clustered case 2: Sparse 2 | 50 pts |
| 10 Outliers outside the cluster: Outlier | 203 pts |
| 52 left censored data outside the cluster: LOQ=LOQ | 203 pts |

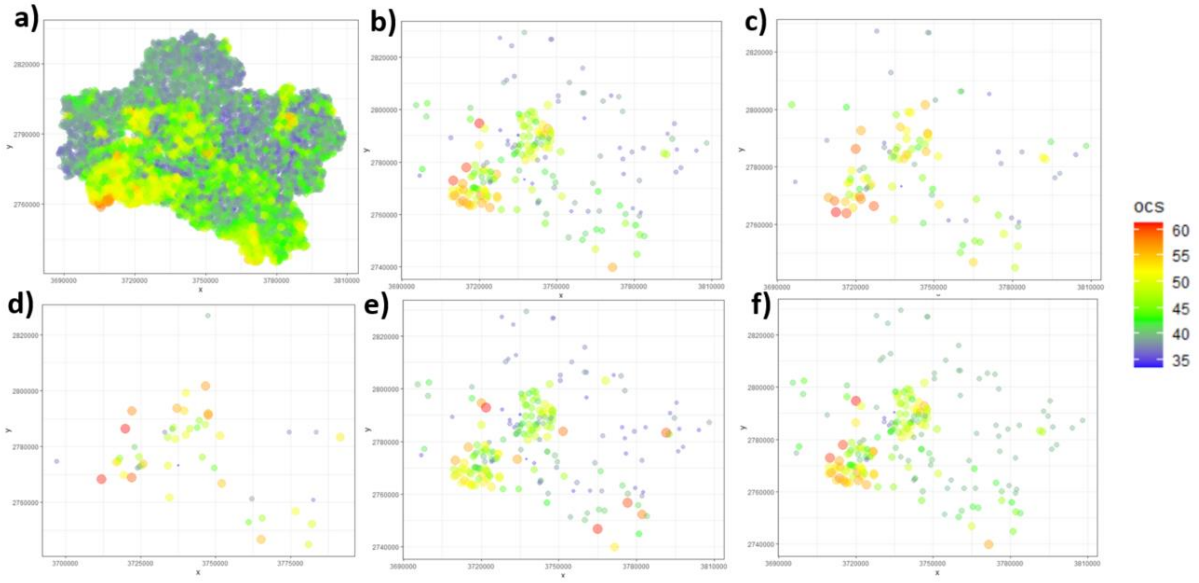


**Figure 94: Harmonized Operation of Uncertainties in Spatialized Environmental Systems – HOUSES datasets.**

*With a) true (10000 pts), b) clustered set (200 pts), c) sparse 1 set (200 pts), d) sparse 2 set (50 pts), e) outlier set (203 pts), and f) left censored (203 pts).*





Funded by the European Union

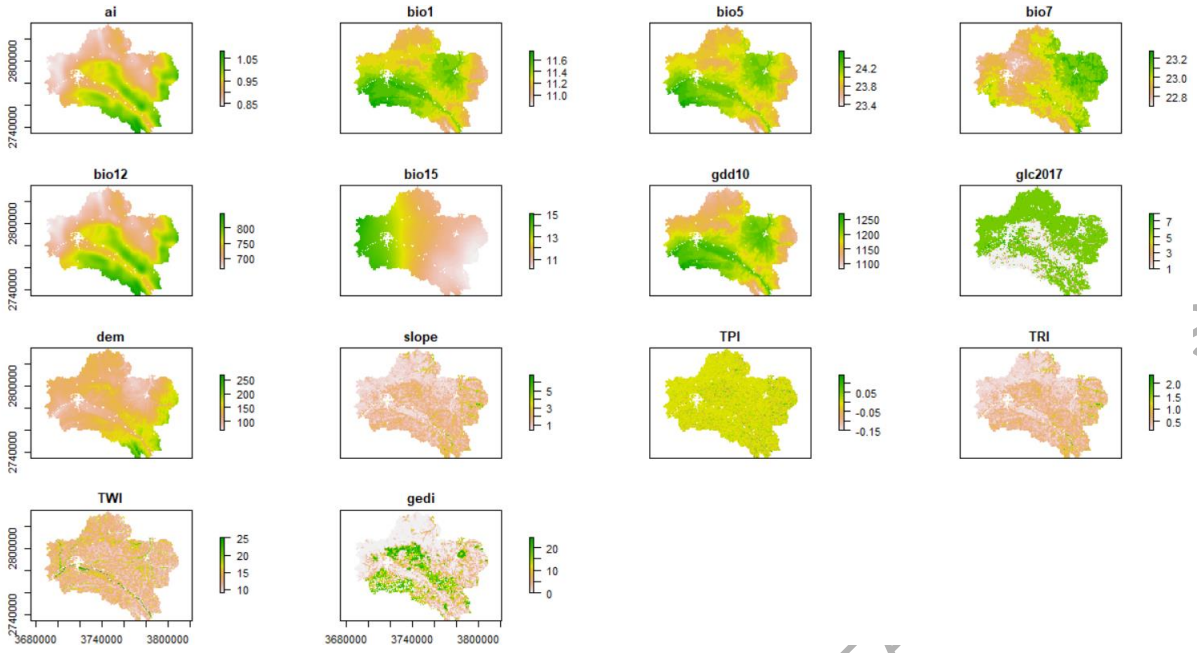


**Figure 95: 14 covariable HOUSES maps available for testing additional information in interpolation**

## Indicators

The MAE, ME, RMSE, and MAXE are used to assess the accuracy of deterministic forecasts. They address deviation from the central tendency, i.e., the mean. The w.PI, cov.PI, Mcov.PI, and MCRPS are used to evaluate stochastic forecasts.

- For the width of the 90% confidence interval (w.PI) of the prediction, the smallest interval could be more precise if his coverage is high (cov.PI). For the coverage of the 90% confidence interval (cov.PI), evaluate the performance of the 90% confidence estimator. It is a probability so the higher, the better.
- The mean absolute deviation of the accuracy plot (Mcov.PI).
- The mean continuous ranked probability score (MCRPS) is used to evaluate the evolution of the accuracy of probabilistic forecast ensembles. Its principle role is to compare the cumulative distribution function of the forecast ensemble with the Heaviside function (step function) of the observed value (Gneiting and Raftery, 2007). So the smaller the MCRPS, the better.interpolation.
- $MCRPS = \frac{1}{n}\sum_{t=1}^{n}\int_{0}^{\infty}(F(x_t) - H(x_t \geq y_t))^2 dx$

where $F(x)$ is the value of cumulative distribution function of the forecast ensemble at the discretization step t, $H(x \geq y)$ is the Heaviside function (step function) of the ranked observed value.

Results

A variety of EEPH algorithm options were used, giving the results in Table 24 with a neutral version and Table 25 with various covariates. The results are compared with a QRF of the same data, which serves as a reference technique. A skill score indicator summarizes the





Funded by the European Union

test results obtained for each dataset; the lower the score, the greater the success. This method has already been used for the previous SIC2004 intercomparison.

The performance of the neutral EEPH is better than the QRF for the sparse2 dataset of 50 data points with and without covariates, which is in line with expectations for our developments (Figure 96). With more than 50 data points, it is better to opt for QRF or kriging. It may be noted that the dimensional reduction of covariates greatly improves the behavior of the EEPH on sparse dataset 1 (200 data points,) to the point of equaling the score of the reference QRF.

**Table 24: Calculus without covariates on various HOUSE datasets with QRF and EEPH algorithm**

| Neutral (without covariates) | MAE | ME | RMSE | MAXE | W.PI | Cov.PI | Mcov.PI | MCRPS | PCOR | Skill score | Case | Neigh. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **QRF** | **2.7** | | **3.5** | **21.0** | **5.00** | **80%** | **0.05** | **1.86** | **0.59** | **10** | **cluster** | 200 |
| **EPH_interval** | 2.6 | 0.10 | 3.5 | 20.4 | 4.03 | 68% | 0.15 | 2.03 | 0.61 | 10 | cluster | 10 |
| **QRF** | **2.8** | | **3.6** | **21.0** | **4.70** | **80%** | **0.04** | **1.96** | **0.56** | **11** | **sparse1** | 200 |
| **EPH_interval** | 2.8 | -0.07 | 3.7 | 20.1 | 4.24 | 71% | 0.13 | 2.11 | 0.55 | 11 | sparse1 | 10 |
| **QRF** | **4.1** | | **4.8** | **19.0** | **5.10** | **73%** | **0.09** | **2.93** | **0.14** | **20** | **sparse2** | 50 |
| **EPH_interval** | 3.8 | 1.35 | 4.6 | 18.5 | 4.63 | 67% | 0.14 | 2.85 | 0.24 | 17 | sparse2 | 10 |
| **QRF** | **3.0** | | **3.9** | **20.6** | **6.80** | **85%** | **0.03** | **2.02** | **0.55** | **12** | **outlier** | 200 |
| **EPH_interval** | 2.9 | 0.95 | 3.9 | 19.7 | 5.64 | 73% | 0.13 | 2.21 | 0.59 | 12 | outlier | 10 |
| **QRF** | **3.0** | | **3.7** | **19.2** | **3.90** | **44%** | **0.26** | **2.23** | **0.57** | **12** | **lq = lq** | 200 |
| **EPH_interval** | 2.9 | 0.90 | 3.6 | 19.2 | 3.13 | 37% | 0.30 | 2.33 | 0.60 | 11 | lq = lqr | 10 |
| **EPH_interval** | 9.1 | 8.17 | 11.5 | 33.6 | 16.90 | 83% | 0.11 | 5.89 | 0.57 | 11 | < Lq | 10 |

*skill score = MAE + abs(ME) + RMSE + 10*(1- Pearson's r), MAE = mean absolute error, ME = mean error or bias, RMSE= root mean squared error, r = Pearson's correlation coefficient r between true and estimated ones, w.PI = width of the 90% confidence interval, cov.PI = coverage of the prediction interval, Mcov.PI = mean absolute deviation of the accuracy plot, MCRPS = mean continuous ranked probability score , Neigh is the number of neighbors chosen to estimate a point.*





Funded by the European Union

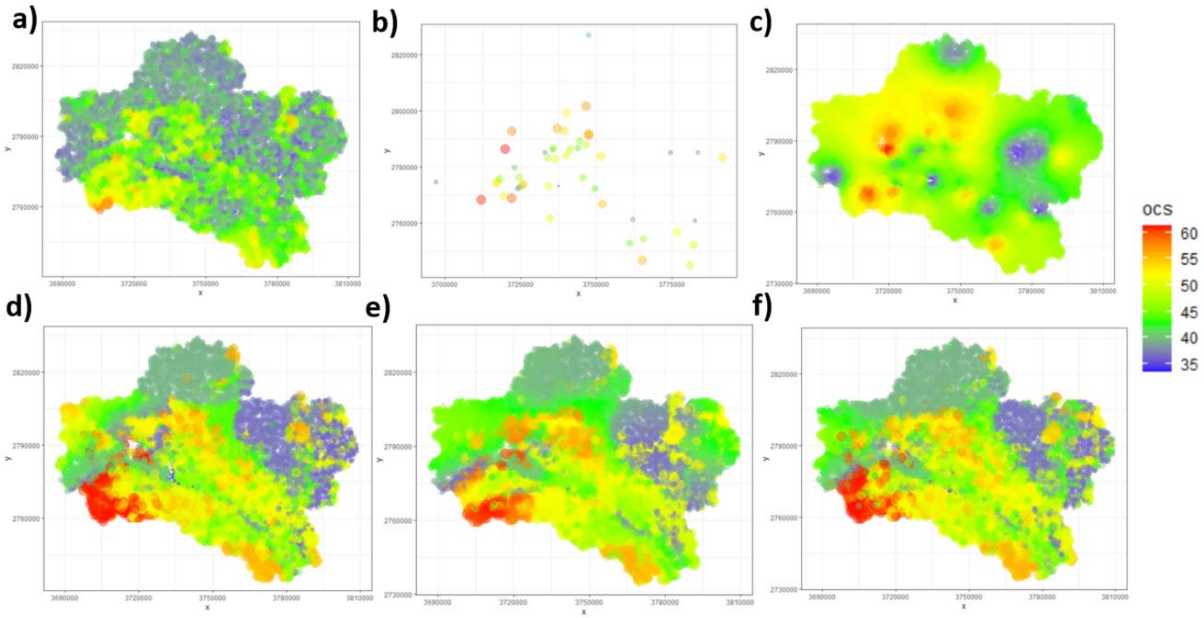


**Figure 96: Maps generated by EEPH on the Scares 2 set**

*With a) true, b) sparse 2 set, c) resulting EEPH without covariates, d) resulting EEPH with 14 covariates, e) resulting EEPH with 3 covariates (UMAP V1, V2 + glc2017), and f) resulting EEPH with 3 covariates (PCA V1, V2 + glc2017).*





Funded by the European Union

**Table 25: Calculus with 14 covariates on various setcases with QRF and EEPH algorithm**

| With covariates | MAE | ME | RMSE | MAXE | W.PI | Cov. PI | Mcov. PI | MCRPS | PCOR | Skill score | Case | Neigh. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **QRF 14 covars** | **2.2** | | **2.90** | **19.5** | **5.00** | **89%** | **0.01** | **1.55** | 0.76 | 8 | **cluster** | 200 |
| EPH_interval 14 covars | 2.4 | -0.18 | 3.35 | 20.0 | 0.95 | 25% | 0.35 | 2.36 | 0.67 | 9 | cluster | 10 |
| **QRF 14 covars** | **2.4** | | **3.20** | **19.0** | **4.40** | **83%** | **0.03** | **1.70** | 0.69 | 9 | **sparse 1** | 200 |
| EPH_interval 14 covars | 2.5 | -0.39 | 3.38 | 19.9 | 0.96 | 22% | 0.37 | 2.34 | 0.64 | 10 | sparse1 | 10 |
| EPH_interval 14 covars + WASS dist | 2.6 | -0.36 | 3.54 | 21.88 | 0.729 | 18% | 0.38 | 2.45 | 0.61 | 10 | sparse1 | 10 |
| EPH_interval 3 covars (UMAP V1, V2 + glc2017) | 2.4 | -0.50 | 3.26 | 20.4 | 1.50 | 34% | 0.32 | 2.17 | 0.67 | 9 | sparse1 | 10 |
| EPH_interval 3 covars (PCA V1, V2 + glc2017) | 2.4 | -0.44 | 3.23 | 19.8 | 1.58 | 35% | 0.31 | 2.15 | 0.67 | 9 | sparse1 | 10 |
| **QRF 14 covars** | **2.9** | | **3.60** | **16.7** | **4.90** | **82%** | **0.04** | **1.99** | 0.60 | 11 | **sparse 2** | 50 |
| EPH_interval 14 covars | 2.5 | -0.11 | 3.36 | 16.1 | 1.49 | 27% | 0.34 | 2.35 | 0.64 | 10 | sparse2 | 10 |
| **QRF 14 covars** | **2.3** | | **2.90** | **17.2** | **6.90** | **91%** | **0.04** | **1.63** | 0.77 | 9 | **outlier** | 200 |
| EPH_interval 14 covars | 2.7 | 0.38 | 3.93 | 20.7 | 1.26 | 27% | 0.34 | 2.69 | 0.62 | 11 | outlier | 10 |
| **QRF 14 covars** | **2.6** | | **3.10** | **19.0** | **4.10** | **51%** | **0.20** | **1.93** | 0.76 | 9 | **lq = lq** | 200 |
| EPH_interval 14 covars | 2.6 | 0.61 | 3.31 | 17.5 | 0.62 | 11% | 0.41 | 2.49 | 0.67 | 10 | lq = lq | 10 |
| EPH_interval 14 covars | 10.4 | -8.94 | 13.07 | 37.6 | 12.01 | 52% | 0.31 | 7.28 | 0.58 | 37 | < lq | 10 |

*Skill score = MAE + abs(ME) + RMSE + 10*(1- Pearson's r), MAE = mean absolute error, ME = mean error or bias, RMSE= root mean squared error, r = Pearson's correlation coefficient r between true and estimated ones, w.PI = width of the 90% confidence interval, cov.PI = coverage of the prediction interval, Mcov.PI = mean absolute deviation of the accuracy plot, MCRPS = mean continuous ranked probability score , Neigh is the number of neighbors chosen to estimate a point.*





Funded by the European Union

# 5. Adapting covariates to SIC domains

## 5.1. Principle

In probabilistic terms, the contribution of a covariate C to our estimate E is expressed by the joint distribution (E|C). To construct this distribution, a minimum number of data points is required. This number depends on the "universe" on which we wish to calculate the distribution. For example, if C can take nc values and E ne. Knowing the E|C distribution requires us to determine nc * ne probabilities, if we project this onto a table or distinct geographical entities, we will have nc*ne cells or entities, with at least 1 data point per cell. For example, if ne=5 and nc = 5, our table would have 25 cells. A minimum of 25 data points, with one in each cell. If we wanted to create a complete probability distribution within each box, we would need more. This can be understood as a set of boxes representing our universe of possible measurements, into which we would arrange our data. The number of measurements in each box makes it possible to calculate probability (Figure 97). The fuzzy interval version of this system would correspond to a measurement system in which the proximity of each measurement to the edges of the boxes would be measured (Huang and Moraga, 2004).

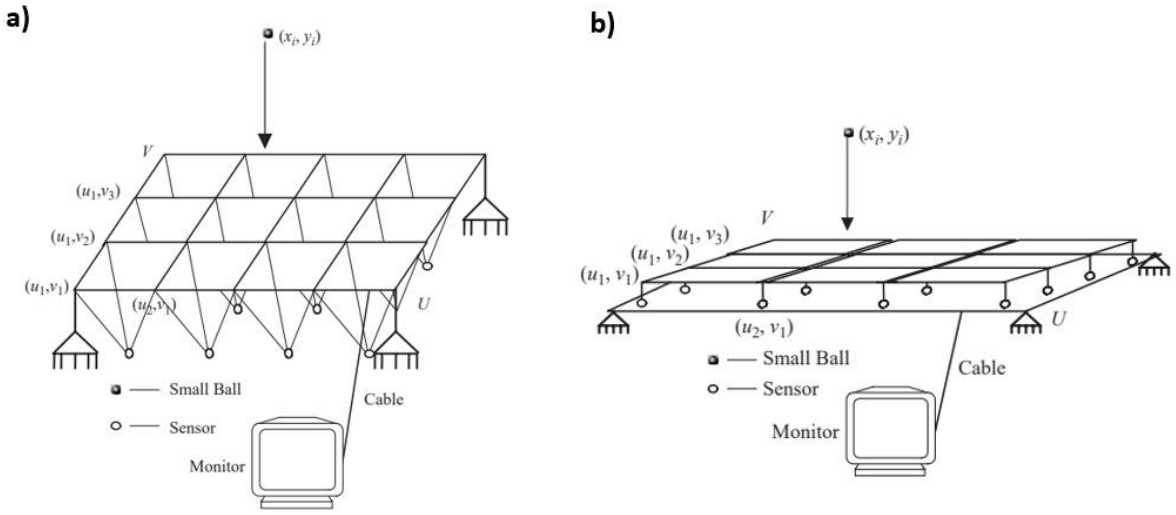


**Figure 97: Analogic model for fuzzy logic**

*with a) a crisp-interval information matrix, and b) Fuzzy-interval information matrix from Huang and Moraga (2004). Note: Let (x_i,y_i ),i=1..n be observations of a given sample X with domain U of input and range V of output.*

Other authors speak of 30 data points per square (Fisher, 1992). This oft-repeated number 30 has become widely accepted as the number at which the central limit theorem begins to apply. One hypothesis is that this number was not actually calculated, but was the limit of n in the Student's t and Normal function tables held on a single page and presented in books. A different explanatory approach can be proposed:





Funded by the European Union

Let X_(1 )… X_(n ) be measurements of a Gaussian random variable of mean μ and variance σ^2. Experimentally, we calculate estimators

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ without Bessel's correction (n-1 instead of n)

$E[S^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}((X_i - \mu) - (\bar{X} - \mu))^2\right]$

$$E[S^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + (\bar{X} - \mu)^2\right]$$

But $(\bar{X} - \mu) = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$, by replacing in the previous equation and combining the terms

$$[S^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\bar{X} - \mu)\, n\,(\bar{X} - \mu) + (\bar{X} - \mu)^2\right]$$

$$[S^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X} - \mu)^2\right]$$

$$[S^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] - E\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{X} - \mu)^2\right]$$

$$[S^2] = \sigma^2 - E\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{X} - \mu)^2\right]$$

$$[S^2] = \sigma^2\left(1 - \frac{1}{n}\right)$$

If n is statistically sufficient, we get $[S^2] \cong \sigma^2$. An additional sample will also satisfy this relationship. Trivially, we have:

$$\sigma^2\left(1 - \frac{1}{n+1}\right) \cong \sigma^2\left(1 - \frac{1}{n}\right) \text{ hence } \frac{1}{n+1} \cong \frac{1}{n}$$

If we plot the two curves as a function of n (Fogure 98) we see that the approximation occurs near 30.





Funded by the European Union

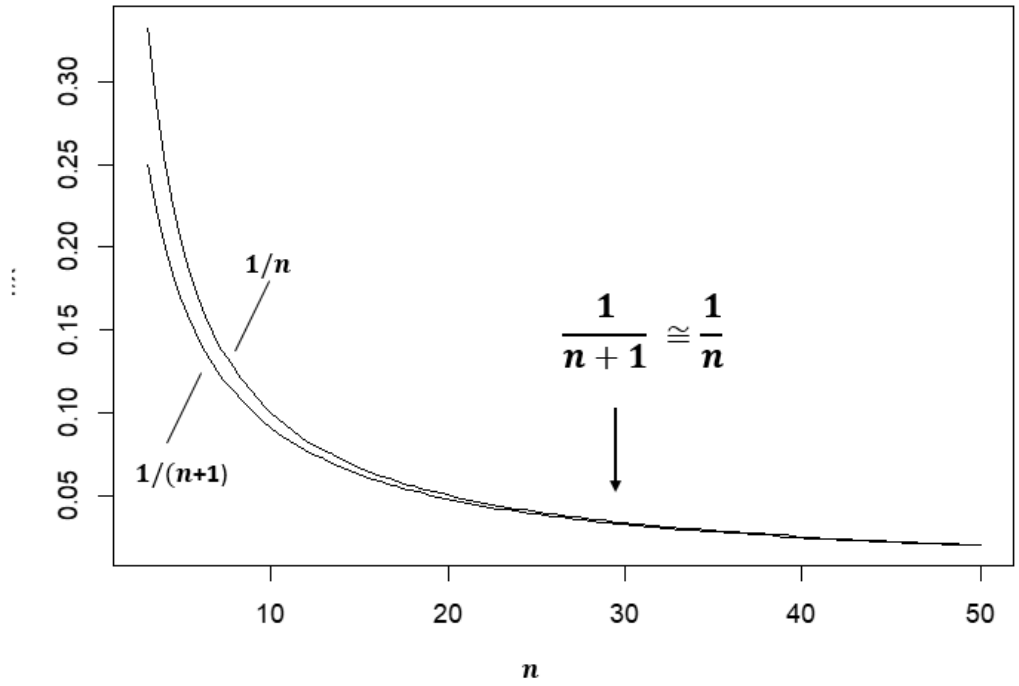


**Figure 98: Inverse n and inverse (n+1) curves near touch around n=30.**

In practice, the rules are empirical; Hair et al., 2010 (*Multivariate Data Analysis*, 7th Edition, 2010), suggest a minimum of 5 observations per variable (5:1), and state that an acceptable size would be 10 per variable (10:1). Pyrcz and Deutsch, (2014) recommend 10 data points per cell, giving the example of a histogram with 10 classes of 10 data points each: "A general rule is that a minimum of 10 data points are needed for a particular 'statistic'; for example, 10 data points would be needed to compute a mean with any reasonable level of precision. Two hundred data points would be needed to compute a 90% probability interval (10 data points below the 5% limit and 10 above the 95% upper limit). This is of course overly simplistic, but useful nonetheless." For a single variable, this adds up to a minimum of 100 data points, and is a complete number when expressed in percentagesé".

If the shape of the E|C distribution is known to us (sometimes studies have been conducted in the past that guide the expert in the choice of possibilities/probabilities), it is possible to estimate the desired number. For example, if the distribution is multi-normal (as in the case with the distribution of large numbers, i.e., big data), it is possible to estimate the number of data points required as a function of the dimension (Table 26). We can see that there are around 20 samples minimum in 2D and almost 70 in 3D.





Funded by the European Union

**Table 26: Some sample sizes when estimating a standard multivariate normal density from Silvermann (1986)**

| Dimensionality | Required sample size |
|---|---|
| 1 | 4 |
| 2 | 19 |
| 3 | 67 |
| 4 | 223 |
| 5 | 768 |
| 6 | 2790 |
| 10 | 842000 |

Note: Some sample sizes required to ensure that the relative mean square at zero is less than 0.1 when estimating a standard multivariate normal density using a normal kernel and the window width that minimizes the mean square error at zero, from Silverman (1986).

Consequently, if we have SIC data and want to use covariables, which means locking E|C, we make suggestions for the SIC data, since we cannot increase ne nor adapt our claims greatly for nc by intelligently changing C. If C cannot be changed, it is better to be careful.

- For numerical data, we can use Pearl's logic, as in the example to be developed, to effectively truncate each covariate at its PNS>0. It is also possible to merge the often-redundant numerical parameters using a dimensional reduction technique. There are several such techniques, mainly used in genetic research. The best known are linear PCA, MDS and its variants (Torgerson, 1954; Kruskal, 1964; Silva and Tenenbaum, 2003), and new techniques such as diffusion (Coifman et al., 2005; Moon et al., 2019) , T-SNE (Van der Maaten and Hilton, 2008), and UMAP (McInnes et al., 2018). An article by Cook et al. (2018) and a digital book by Cook and Laa (2023) present these algorithms interactively and suggest an interactive visualization under R of the results. For ISLANDR, we have no algorithmic preference; several tests can be conducted quickly under R to choose the most suitable.

- For categorical data, such as a land-use map, the number of classes nc can be adjusted by merging so that there are at least 3 data points in each cell, or with a target objective, so that the merged cells are close in distance to the measurement. This technique is correctly explained by Heißerer et al. (2016), and in line with the probabilistic concepts presented in this report. A system of multi-distance covariate pruning has therefore been coded under R.

For each parameter or parameter combination, a numerical layer and an optimized categorical layer can be produced. The next two paragraphs show an example of how this was achieved prior to contaminant mapping for ITA3.





Funded by the European Union

## 5.2. Statistical measurement of weight between covariates

Various covariates can be made available on our ITA to help EEPH interpolate more accurately. The idea we propose is to introduce an optimized mixture of covariates as an EEPH parameter. Within a chosen neighborhood, a statistical measure can be made meaningful, and covariates can be ranked by relative importance for the content, using a statistical distance between their distributions. The weight assigned to the various parameters is then injected as a weight into the EEPH (w3). Several distance calculation methods are then possible to compare F_1 and F_2 globally and thus establish a possible dependency. Choosing between the two methods is not easy, as it depends on both the number of data points and the size of the distance calculated (Niles-Weed and Rigolet, 2019). Both types of distance have been coded for ISLANDR. The distance measurement must be positive and zero if, and only if, the two distributions are independent.

In a context of sparse data, only an expert can confirm that a distance statistic makes sense for weighting our covariates and interpolating our data. Without any prior knowledge, it's best to work like a geostatistician, selecting the most linear correlated covariates and ranking them.

### 5.2.1. The Wasserstein distance

Derived from Wasserstein's work (1969), this distance is calculated as follows.

$$W(X,Y) = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|$$

This measurement is always positive and is zero if, and only if, X is equal to Y. It is also referred to as linear error in probability space (LEPS) by Webster and Oliver (2007) and Heißerer et al. (2016). For a Euclidean space, it performs well, but only converges slowly (speed of $n^{-1/d}$). It performs best for sparse data (Niles-Weed and Rigollet, 2019). For large dimensions, an energy measurement is preferable.

### 5.2.2. Energetic distance

According to Szekely and Rizzo (2004), energetic distance is calculated as follows:

$$\varepsilon_{n,m}(X,Y) = 2\,\frac{1}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m}|x_i - y_k| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}|x_i - x_k| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{k=1}^{m}|y_i - y_k|$$

This measurement is always positive and is zero if, and only if, X is equal to Y. This measurement outperforms the Wasserstein distance on large dimensions.





Funded by the European Union

For a robust statistic with good properties under the assumption of equality of the two distributions, Szekely and Rizzo (2004) propose the following test:

$$T = \frac{nm}{n+m}\varepsilon_{n,m}(X,Y)$$

### 5.2.3. Probabilistic approach using tau, nu and lambda models

Several probabilistic treatment solutions for covariate weight and redundancy have been proposed, such as the tau model (Journel, 2002; Krishnan et al., 2005; Krishnan, 2008), the nu model (Polyakova and Journel, 2007) and the lambda model (Hong, 2010). As we have just seen, these models apply weights to conditional probabilities to take into account their importance and possible redundancy, but only for a non-SIC setting.

The tau and nu models apply to what are known as indicator (binary) geostatistical variables. We prefer the lambda model, which we'll call w3 in our notation, and which can replace distance measures if the number of data points allows. This is a data-intensive conditional probability method, which we will only apply if:

1) the calibration data is sufficient (as there is dimensional leakage, we'll quickly leave the SIC setting) and 2) cross-validation of the various parameter combinations adapted by the previous methods is unsuccessful.

If $D_i$ are i th covariates, and z is content

$$p_w(z|D_1,\dots D_K) = \frac{f(D_1,\dots D_K|z)p(z)}{f(D_1\dots D_K)} = \frac{f(D_1|z)^{w_1} \times \dots \times f(D_K|z)^{w_K}p(z)}{f(D_1\dots D_K)}$$

$$= p(z)\left(\frac{p(z|D_1)f(D_1)}{p(z)}\right)^{w_1} \times \dots \times \left(\frac{p(z|D_K)f(D_K)}{p(z)}\right)^{w_K} \times C$$

$$= p(z)\prod_{k=1}^{K}\left(\frac{p(z|D_k}{p(z)}\right)^{w_k} \times C$$

Coefficient C is calculated so that $p(z|D_1,\dots D_K) \in [0,1]$ and we replace

$$p_w(z|D_1,\dots D_K) = \frac{p(z)\prod_{k=1}^{K}\left(\frac{p(z|D_k}{p(z)}\right)^{w_k}}{\sum_{z=1,..,u}\left(p(z)\prod_{k=1}^{K}\left(\frac{p(z|D_k}{p(z)}\right)^{w_k}\right)} \text{ and } \sum_{z=1,..,u}\left(p_w(z|D_1,\dots D_K)\right) = 1$$

To find the weights $w_k$, we can use data in a cross-validation process. For a known value of z, only one $p_w(z|D_1,\dots D_K)$ is worth 1 and the others are zero.

$$\frac{p_w(z|D_1,\dots D_K)}{p_w(\bar{z}|D_1,\dots D_K)} = \frac{p(z)\prod_{k=1}^{K}\left(\frac{p(z|D_k)}{p(z)}\right)^{w_k}}{p(\bar{z})\prod_{k=1}^{K}\left(\frac{p(\bar{z}|D_k)}{p(z)}\right)^{w_k}} = \frac{p(z)}{p(\bar{z})}\prod_{k=1}^{K}\left(\frac{p(\bar{z})p(z|D_k)}{p(z)p(\bar{z}|D_k)}\right)^{w_k}$$





Funded by the European Union

$p_w(z|D_1, \dots D_K)$ is known to us at the known n points $A_i$.

## 5.3. Examples of SIC adjustments

### 5.3.1. Example of SIC adjustment of categorical parameters

First, the measurements are subjected to statistics and measurements by category (Table 27) before inter-category distances are calculated (Table 28). We see that some categories have similar mean + var, Wasserstein, or energetic measurements and would benefit from being grouped together. Authors such as Heißerer et al. (2016) consider that if the LEPS differ by only 5% (noted value δ), the categories can be merged.

**Table 27: Base statistics and TPH measurements for the simplified geologic maps**

| Geologic ID | Nb | Mean | Standard devia. | Energy | Wasserstein LEPS |
|---|---|---|---|---|---|
| 1 | 2 | 16 | 5.66 | 0.00006 | 0.004 |
| 2 | 24 | 48.46 | 93.24 | 0.00436 | 0.041 |
| 3 | 57 | 36.74 | 56.51 | 0.00692 | 0.050 |
| 4 | 13 | 25.54 | 10.37 | 0.00181 | 0.026 |
| 5 | 20 | 18.20 | 6.22 | **0.00429** | **0.040** |
| 6 | 20 | 26.10 | 32.34 | **0.00391** | **0.039** |

**Table 28: Categorical inter-distances TPH measurements for the simplified geologic maps.**

| Mean + variance distance | | | | | | | | Wasserstein - LEPS distance | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean+var | 1 | 2 | 3 | 4 | 5 | 6 | | | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 0 | | | | | | | 1 | 0 | | | | | |
| 2 | 2972 | 0 | | | | | | 2 | 0.0363 | 0 | | | | |
| 3 | 14783 | 4508 | 0 | | | | | 3 | 0.0458 | 0.0094 | 0 | | | |
| 4 | 8418 | 1395 | 965 | 0 | | | | 4 | 0.0216 | 0.0147 | 0.0242 | 0 | | |
| 5 | 17448 | 6033 | 226 | 1627 | 0 | | | 5 | 0.0357 | 0.0006 | 0.01 | 0.0142 | 0 | |
| 6 | 16975 | 5746 | 142 | 1489 | 10 | 0 | | 6 | 0.0342 | 0.0021 | 0.0115 | 0.0126 | 0.0015 | 0 |
| | | | | | | | | | | | | | | |
| **Energetic distance** | | | | | | | | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | | | | | | | | |
| 1 | 0 | | | | | | | | | | | | | |
| 2 | 0.0043 | 0 | | | | | | | | | | | | |
| 3 | 0.0069 | 0.0026 | 0 | | | | | | | | | | | |
| 4 | 0.0017 | 0.0026 | 0.0051 | 0 | | | | | | | | | | |
| 5 | 0.0042 | 0.0001 | 0.0026 | 0.0025 | 0 | | | | | | | | | |
| 6 | 0.0039 | 0.0004 | 0.003 | 0.0021 | 0.0004 | 0 | | | | | | | | |





Funded by the European Union

A covariate pruning algorithm was then produced. Note that for each contaminant, the distribution of measurements in the various categories varies (Figure 99); a specific categorical adjustment would be required.

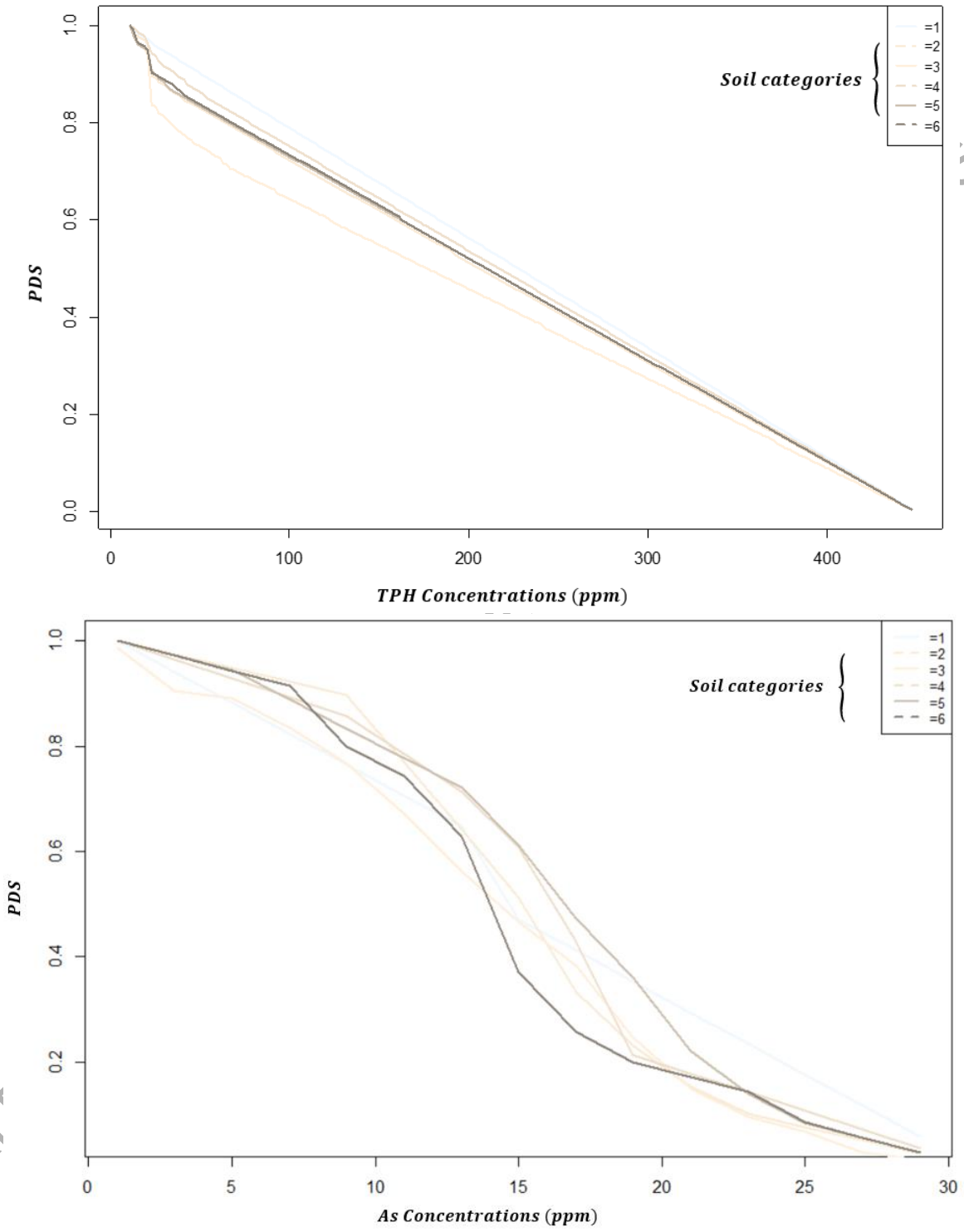


**Figure 99: TPH and As concentration by PDS geologic categories**

If we know the hierarchy of the categorical variables influencing our content, it is possible to adjust the least relevant parameter to a few categories with numbers close to 30 (in



Funded by the European Union

practice 27) before crossing it with another categorical map. The result is then adjusted so that a minimum number of samples (typically 3) per category are present (Figure 100).

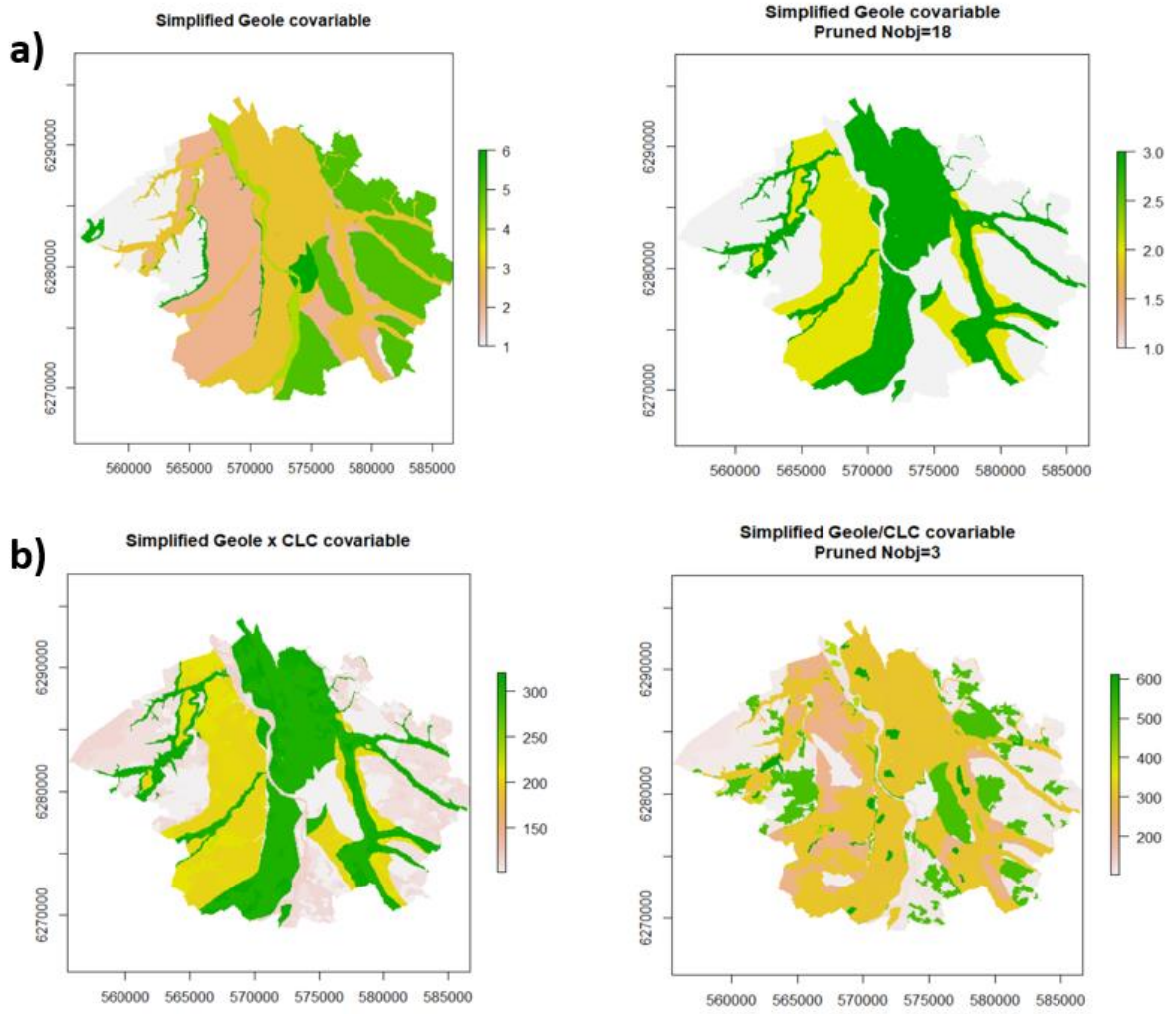


**Figure 100: Categorical covariate pruning. Example for the Toulouse ITA3 TPH concentration. Simplified Geologic map and CORINE land cover map pruning for better spatial interpolation**

*with a) simplified Geole covariable (6 classes) and the optimized pruned Geologic covariable (3 classes) for TPH estimation and b) Geole x CLC covariable (27 classes) and the optimized pruned Geolex CLC (14 classes) covariable (3 classes) for TPH estimation.*

### 5.3.2. Adjustment of numerical parameters for probabilistic SIC

If we take the hydrocarbon data (TPH) from ITA Toulouse and view the TPH content as a function of distance from the road, we obtain Figure 101 and a linear correlation coefficient of 0.015, considered very low or unlikely in IPCC terminology (Mastrandrea et al, 2010).





Funded by the European Union

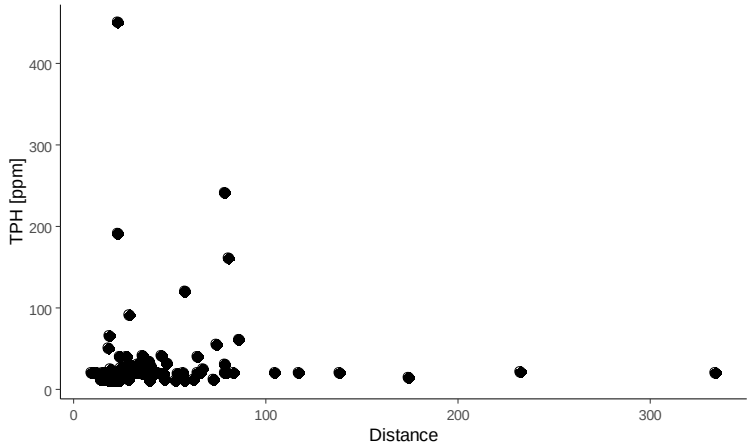


**Figure 101: TPH concentration vs. road distance on the TPH Toulouse ITA**

Nevertheless, the probabilistic approach outlined in Appendix 1 and studying $P$(TPH ≥c$|distance > m) > P(TPH \geq c|\, distance < m)$ with a variable m of one reveals a pivotal PDS position around 30 m from roads (Figure 102).

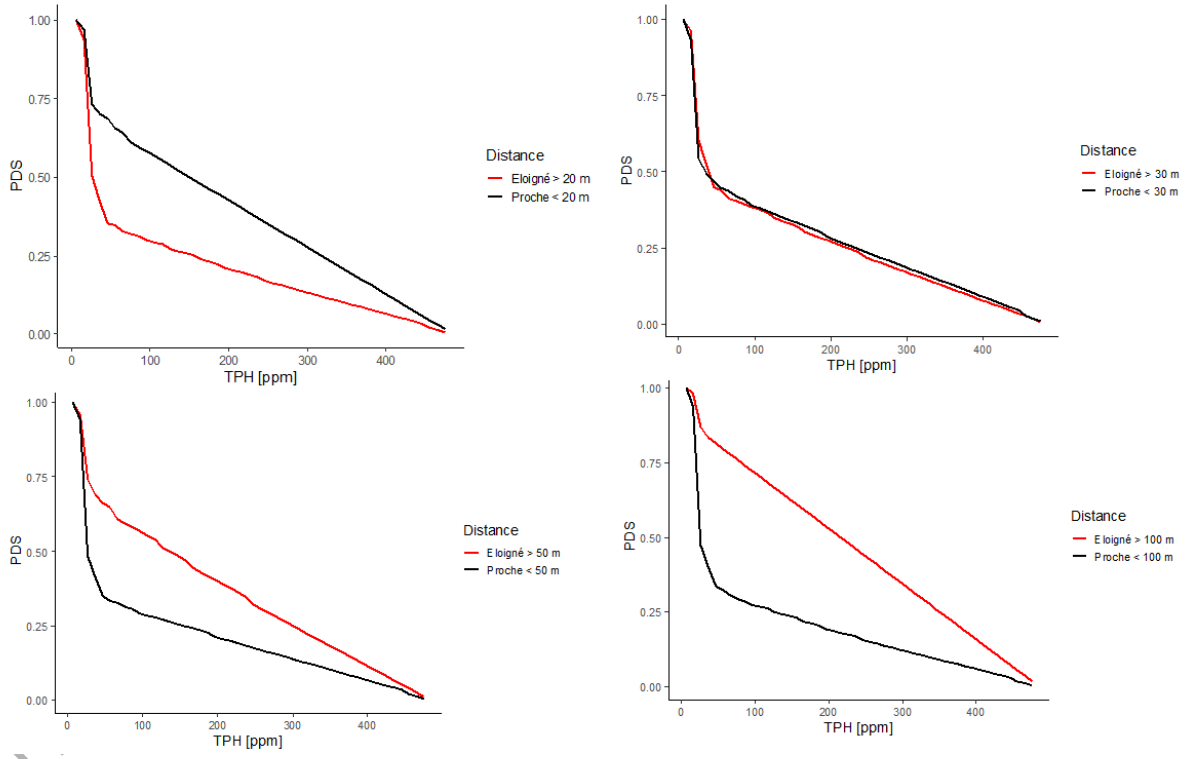


**Figure 102: PDS application for discovering the non-linear TPH/proximity to road relationship**

Locally and up to around 30 m from roads, there is a greater chance of finding high TPH values in probabilistic relation to them. Beyond this distance threshold, the relationship decreases, and any sources of TPH contamination come from other potential sources. This is a non-linear effect over a range of 22–47 ppm, as shown by the joint probability density calculated for the data (Figure 103).



Funded by the European Union

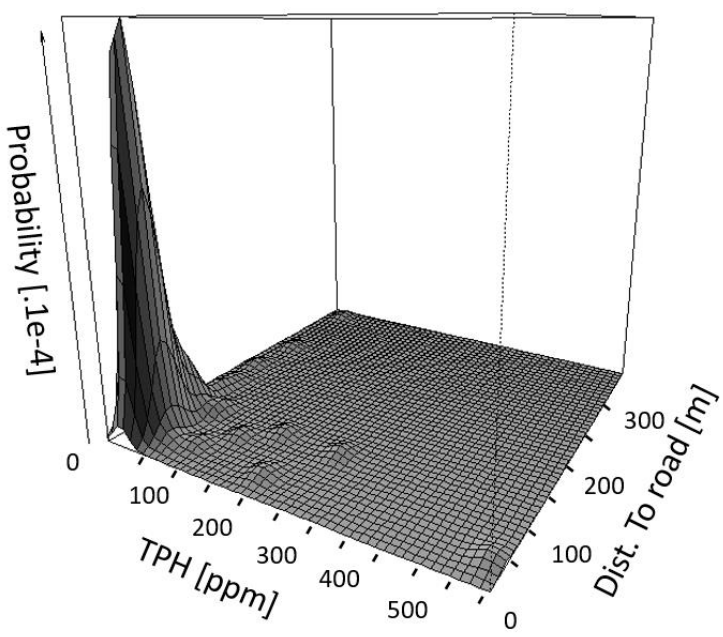


**Figure 103: Probability density of TPH presence at fixed distance from road.**

The example of hydrocarbon dependence with distance from roads in the Toulouse ITA is a good illustration of applying Pearl's logic (2000). Most of them are found at low levels near roads, but not always. This may depend on wind conditions, the density of road traffic at the site, or a past accident involving a spill, not to mention that, using the scale of measurement, a BASIAS industrial site merges 23 times out of 138 with the road beside it.

The road is not always necessary for the presence of hydrocarbons, nor is it always sufficient, as shown by the PS calculations—probability of sufficient causation, probability of necessary causation, and PNS (probability of necessary and sufficient causation)—in Table 29. In this calculation, we again find the pivot around the 30 m distance discussed in the previous chapter.





Funded by the European Union

**Table 29: Probabilities of dependency between TPH and distance to the nearest road, according to Pearl (2000).**

$p = P(TPH|DR)$

| ppm/m | 0–30 | 30–50 | 50–100 | 100–150 | 150–200 | 200–250 |
|---|---|---|---|---|---|---|
| 0–20 | 0.44 | 0.13 | 0.11 | 0.01 | 0.01 | 0.01 |
| 20–30 | 0.05 | 0.02 | 0.02 | 0 | 0 | 0.01 |
| 30–40 | 0.04 | 0.01 | 0.02 | 0 | 0 | 0 |
| 40–50 | 0 | 0.01 | 0.01 | 0 | 0 | 0 |
| 50–100 | 0.01 | 0.01 | 0.01 | 0 | 0 | 0 |
| 100–200 | 0.02 | 0 | 0 | 0 | 0 | 0 |
| 200–500 | 0.01 | 0.00 | 0.01 | 0 | 0 | 0 |

$\overline{p} = P(TPH|\overline{DR})$

| ppm/m | 0–30 | 30–50 | 50–100 | 100–150 | 150–200 | 200–250 |
|---|---|---|---|---|---|---|
| 0–20 | 0.28 | 0.60 | 0.61 | 0.71 | 0.71 | 0.71 |
| 20–30 | 0.05 | 0.08 | 0.08 | 0.10 | 0.10 | 0.10 |
| 30–40 | 0.03 | 0.07 | 0.05 | 0.07 | 0.07 | 0.07 |
| 40–50 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| 50–100 | 0.03 | 0.04 | 0.04 | 0.03 | 0.04 | 0.04 |
| 100–200 | 0 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| 200–500 | 0.01 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 |

$PN(DR \rightarrow TPH)$

| ppm/m | 0–30 | 30–50 | 50–100 | 100–150 | 150–200 | 200–250 |
|---|---|---|---|---|---|---|
| 0–20 | 0.37 | 0 | 0 | 0 | 0 | 0 |
| 20–30 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30–40 | 0.33 | 0 | 0 | 0 | 0 | 0 |
| 40–50 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50–100 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100–200 | 1 | 0 | 0 | 0 | 0 | 0 |
| 200–500 | 0.5 | 0 | 0 | 0 | 0 | 0 |

$PS(DR \rightarrow TPH)$

| ppm/m | 0–30 | 30–50 | 50–100 | 100–150 | 150–200 | 200–250 |
|---|---|---|---|---|---|---|
| 0–20 | 0.22 | 0 | 0 | 0 | 0 | 0 |
| 20–30 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30–40 | 0.02 | 0 | 0 | 0 | 0 | 0 |
| 40–50 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50–100 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100–200 | 0.02 | 0 | 0 | 0 | 0 | 0 |
| 200–500 | 0.01 | 0 | 0 | 0 | 0 | 0 |

$PNS(DR \rightarrow TPH)$

| ppm/m | 0–30 | 30–50 | 50–100 | 100–150 | 150–200 | 200–250 |
|---|---|---|---|---|---|---|
| 0–20 | 0.16 | 0 | 0 | 0 | 0 | 0 |
| 20–30 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30–40 | 0.01 | 0 | 0 | 0 | 0 | 0 |
| 40–50 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50–100 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100–200 | 0.02 | 0 | 0 | 0 | 0 | 0 |
| 200–500 | 0.01 | 0 | 0 | 0 | 0 | 0 |

The impact of these probabilistic causality calculations on the potential mapping of hydrocarbons in Toulouse sheds light on the role of roads: the SIC data do not make it possible to clearly establish their role. As a consequence, hydrocarbons have been interpolated for the city of Toulouse by a double inequality kriging using no covariates at all (Belbeze et al., 2019) because we're going for parsimony and above all not trying to panic the public, if the map is to be shown to them (Figure 104).





Funded by the European Union

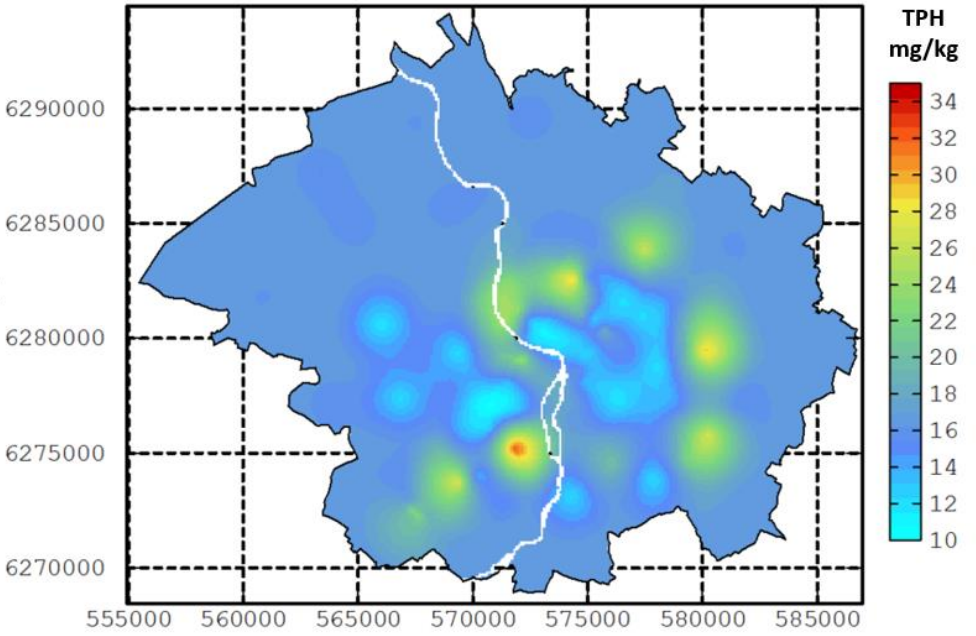


**Figure 104: TPH FPGU by double inequality kriging – no covariates.**

*Topsoil samples from Belbeze et al. (2019), n=139, 0-10cm, TPH analysis by multiple laboratories, LOQ: 10 mg/kg (8 samples), 20 mg/kg (57 samples).*

We also understand that, to correctly represent the phenomenon of hydrocarbons near roads, it would be necessary to carry out systematic roadside analyses to precisely determine the zone(s) where they are present. In the SIC context, we don't have them.

If the road covariate is forced to be taken into account for hydrocarbon prediction, false positives appear:

- We can take this information into account using a belief function and enter into an uncertainty formalism like Dempster-Shafer (Dempster, 2007). In this case, 4 maps will have to be produced. Figure 105 shows the maximum belief function calculated by Belbeze et al. in 2023. This type of formalism then highlights the edges of the roads in red to take into account the low identified dependency.





Funded by the European Union

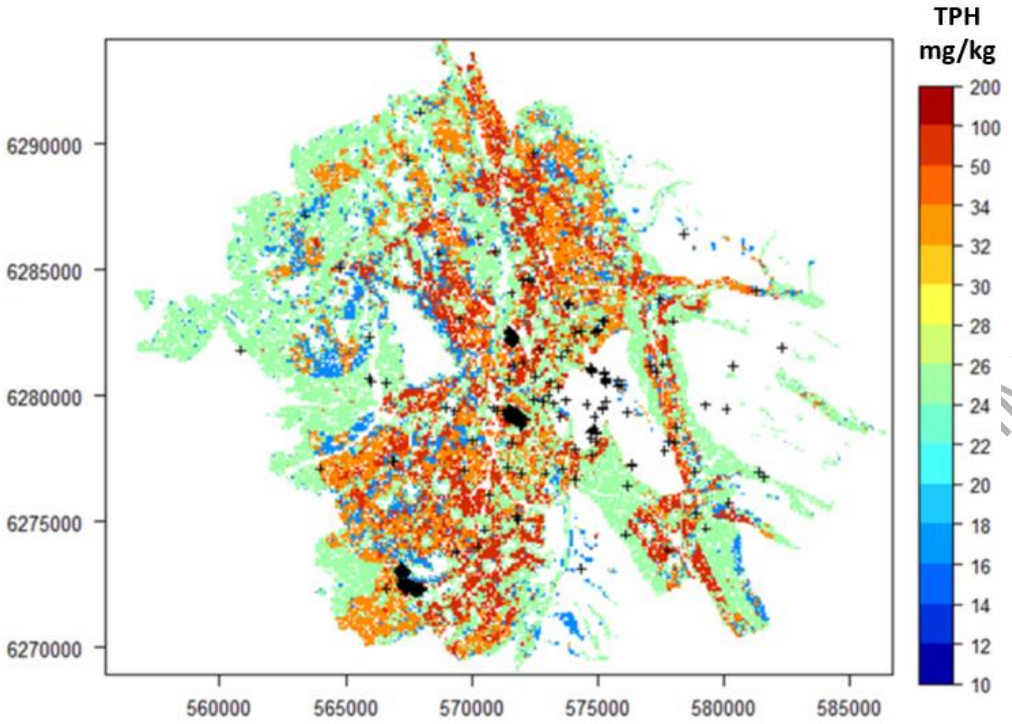


**Figure 105: TPH FPGU relief map by DST with covariates**

Topsoil samples from Belbeze et al. (2019), n=139, 0-10cm, TPH analysis by multiple laboratories, LOQ: 10 mg/kg (8 samples), 20 mg/kg (57 samples). Small X's on map are sampling points.

- For EEPH, the distance calculation should be modified to take into account the loss of road proximity causality beyond a certain distance. Verstraete proposed a technique using a fuzzy rule-based approach to retranscribe a dependency effect in the interpolated result in 2021. This is an elegant method based on re-gridding the various covariates using triangular number inference. To remain within our probabilistic approach, an equivalent result can be proposed by modifying the covariate map based on the information provided by our data. In doing so, only covariate pixels with dependency information are retained (Figure 107).





Funded by the European Union

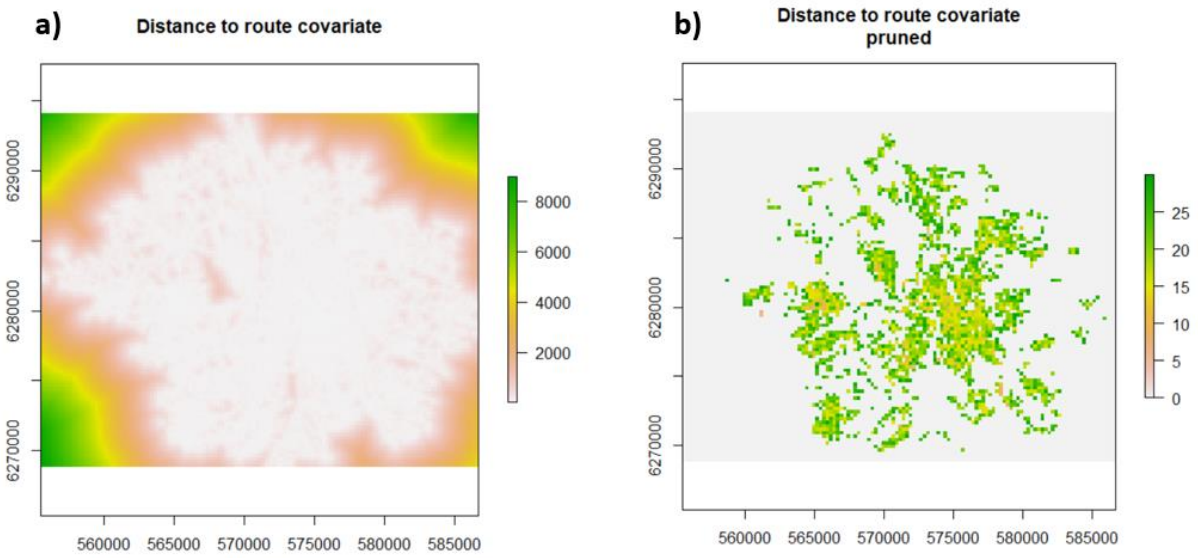


**Figure 106: ITA3 distance to road calculation from Belbeze et al. (2019).**

*With a) the whole range of distance and b) distance to road pruned to 30 m so that PN, PS, PNS > 0.*

Figure 107 shows the effect of introducing the DR covariate on hydrocarbon content estimates (Figure 107 a). If the distance to roads covariate is introduced as is into the calculation (Figure 107 b), we note the presence of false positives in the circles (in fact, a forest and an air strip are marked as contaminated, although this is not necessarily established by boreholes), which justifies our idea of modifying the distance to road covariate. Figure Figure 107 c shows the result of interpolation while retaining only distances to roads <30m: the previous artifacts have disappeared. The parameter then influences the shape of the anomalies, which tend to follow certain roads (see the southwestern part of the map). However, this is just one of the possible configurations that we need to explore through a specific calculation of imprecise probabilities. It's also worth noting that this calculation does not involve the uncertainty of the distance to roads parameter, but the uncertainty of the dependency relationship designating it as a parameter of the model. Moreover, in the case of a multi-parameter model, we will need to address potential information redundancy.





Funded by the European Union

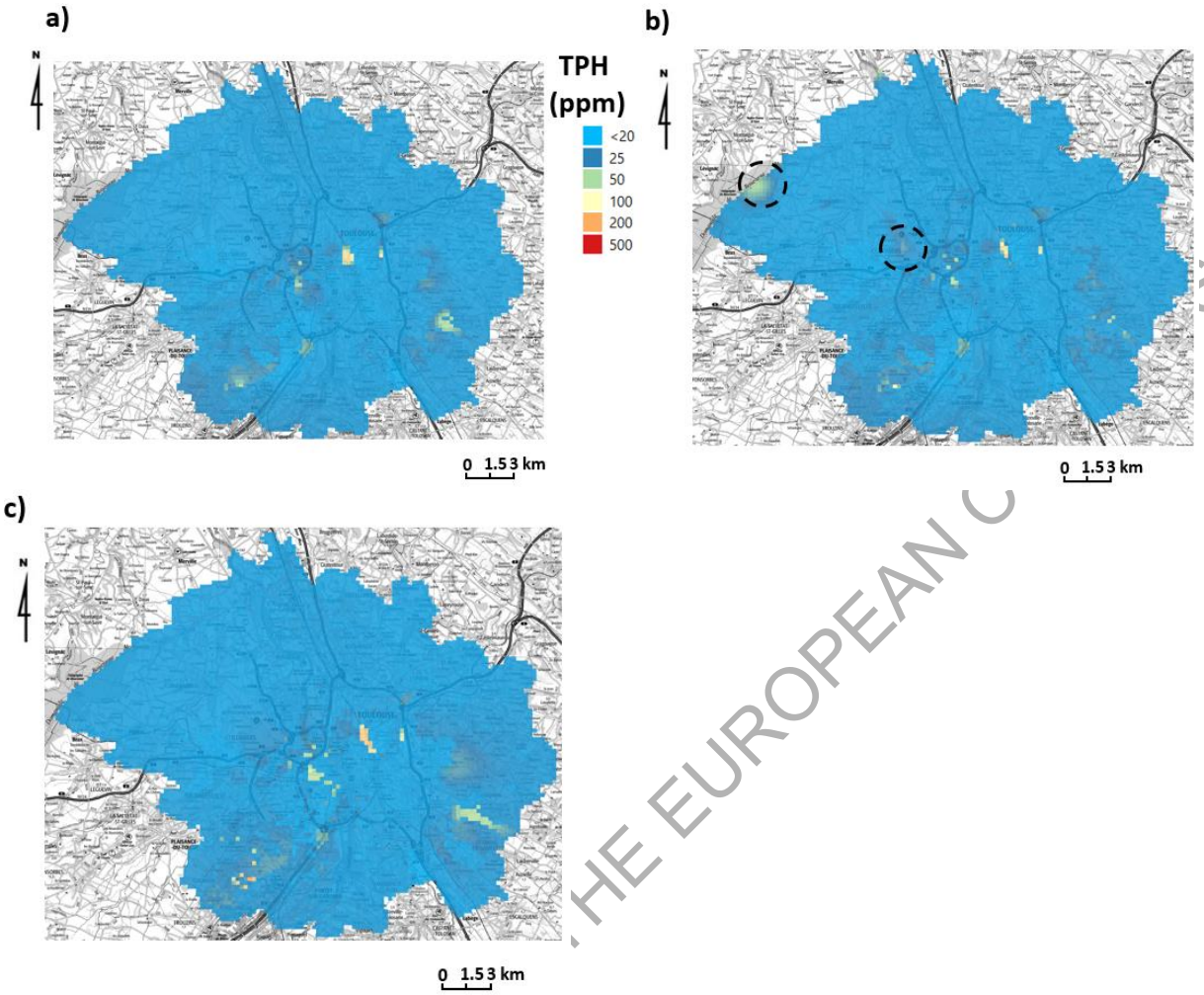


**Figure 107: Three interpolation methods for mapping total petroleum hydrocarbon (TPH) in the city of Toulouse.**

*With a) EEPH expectation with LANU covariable, b) EEPH expectation with LANU and road distance covariables., and c) EEPH expectation with LANU and pruned road distance covariables.*

Topsoil samples from Belbeze et al. (2019), n=139, 0-10cm, TPH analysis by multiple laboratories, LOQ: 10 mg/kg (8 samples), 20 mg/kg (57 samples). False positive are dotted circles.





Funded by
the European Union

Calculating the EEPH in imprecise probability mode gives 2 cdf min and max for each pixel, framing reality. Figure 108 shows the cdf for a pixel taken from the center of the map.

$$p = TPH|LANU.DR \; et \; \bar{p} = TPH|LANU.\overline{DR}$$

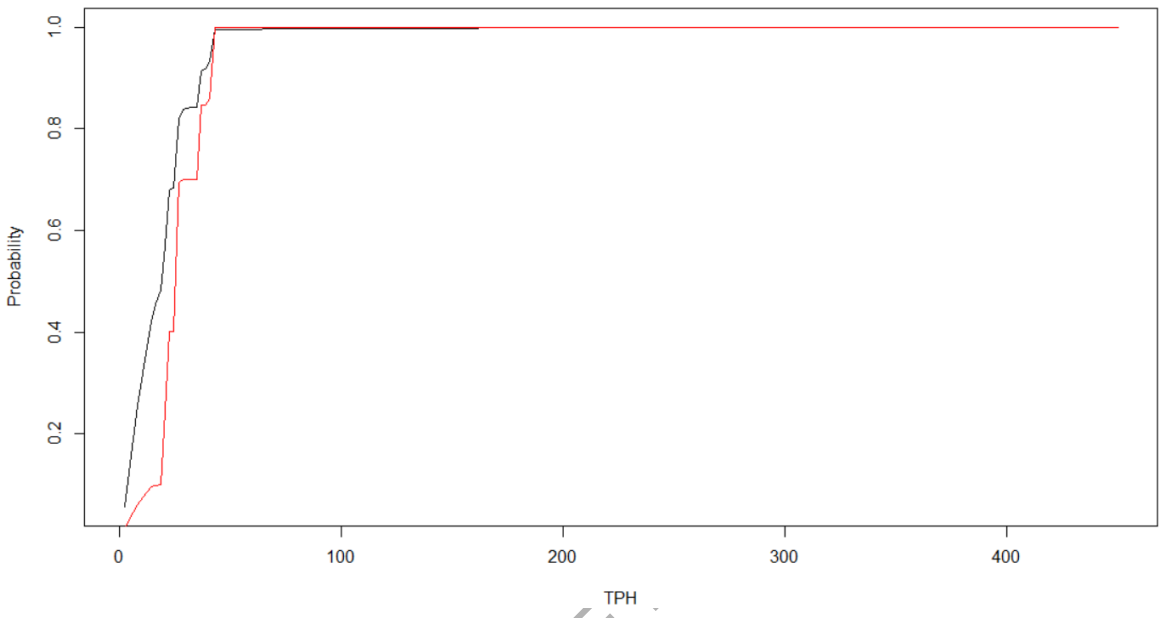


**Figure 108: EEPH imprecise probability cdf LANU with and without pruned road distance covariables on point in the middle of the map.**

This approach is well illustrated on two covariates (LANU and DR), and underscores that various biases arise at each interpolation with expert opinion prescribing a covariate: whether or not to choose this covariate and the possible redundancy of information between the various covariates the expert has chosen.





Funded by the European Union

### 5.3.3. Handling redundancy using dimensional reduction of numerical covariates

First, the simplest reduction technique to learn the PCA is applied. The scree plot (Figure 109) shows that 2 to 3 dimensions are sufficient to effectively summarize the five covariates (DE: distance to the nearest river, DBAS: distance to the potential polluted sites, DBOL: distance to polluted site, DC: distance to the town center, DR: distance to road). As a result, the first two PCA axes and a UMAP were mapped (Figures 110).

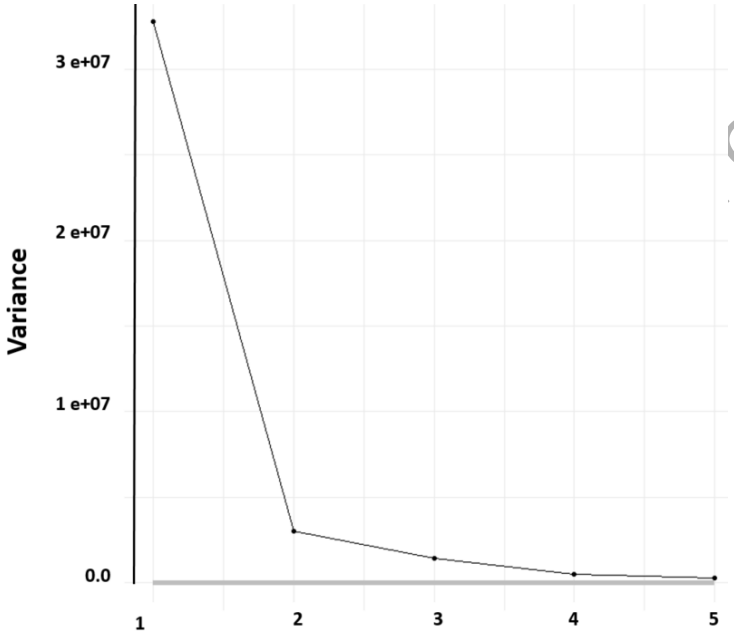


**Figure 109: Scree plot of the PCA of five numerical covariates for the Toulouse ITA**

*With a) PC 1,2,3 as three axes of PCA analysis and V1 and V2 the two axes resulting from an UMAP.*





Funded by the European Union

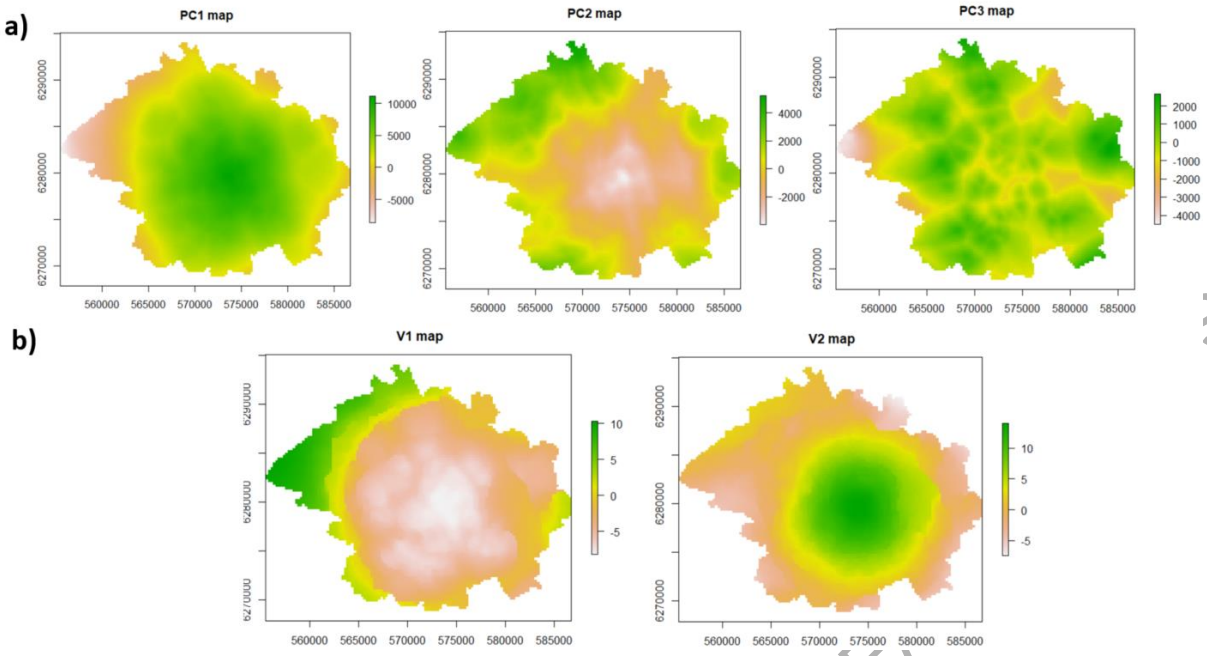


**Figure 110: Linear dimension reduction (PCA) and non-linear dimension reduction (UMAP) for the Toulouse ITA dataset.**

Figure 111 shows that the 2D UMAP transformation would be the most efficient without altering the slight linear correlation observed on the factors. Seen from this angle, and given the high degree of redundancy observed on this dataset, UMAPs should be used, or even the DR (road distance) map alone could be kept if improved.





Funded by the European Union

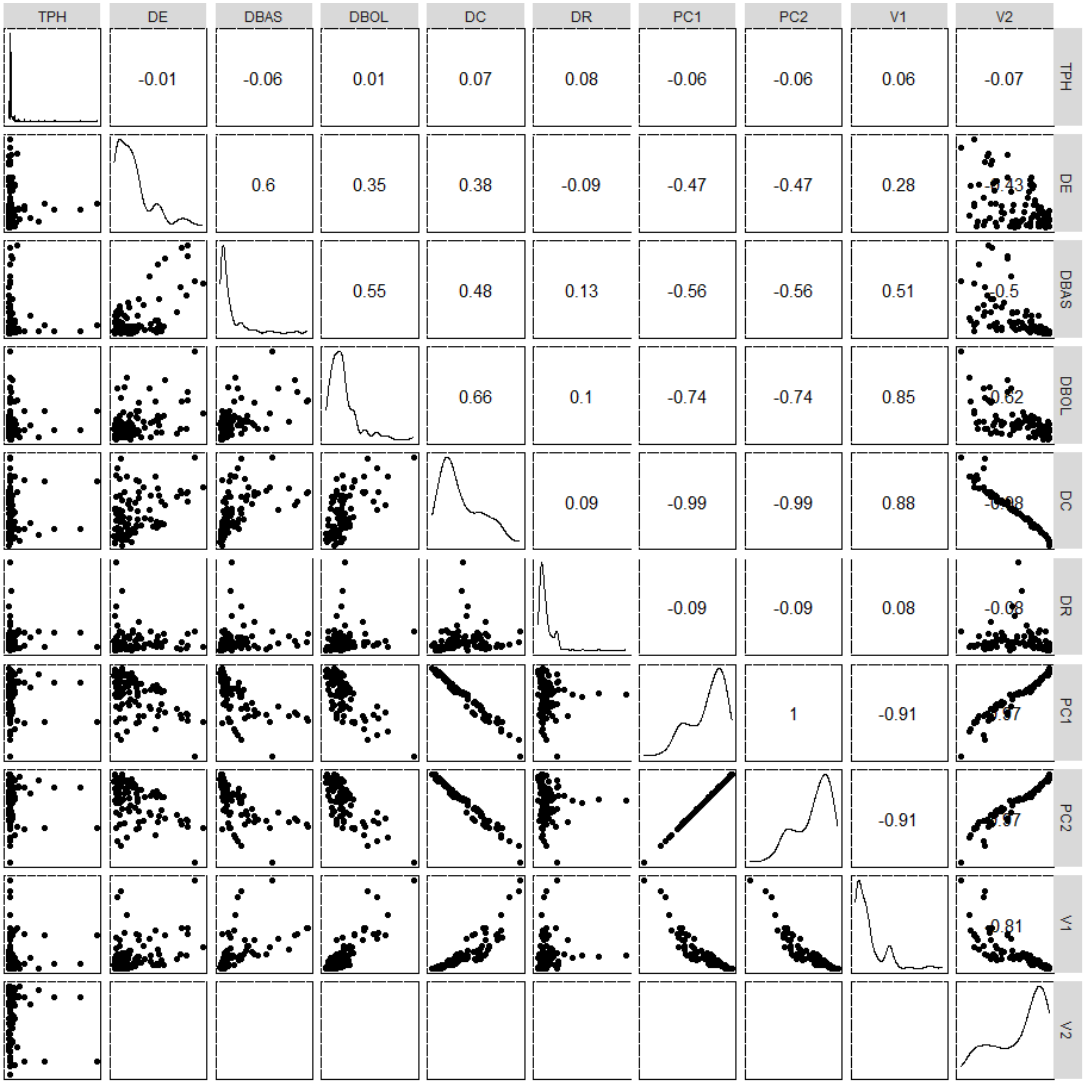


**Figure 111: Enhanced correlation matrix between covariate and possible dimension reduction axes**

*With TPH topsoil TPH concentrations, DE: distance to the nearest river, DBAS: distance to the potential polluted sites, DBOL: distance to polluted site, DC: distance to the town center, DR: distance to road, PC1, PC2: first PCA axis, V1, V2: UMAP axis*





Funded by the European Union

### 5.3.4. Change in scale for covariates

The work of Behrens et al., (2019) shows the influence of covariate scaling on correlation with estimated content. As we have non-linear measures of dependence, we can draw on this work to find the optimal scale to use in our maps. Two examples from the Toulouse ETA3.

Figure 112 shows a Gaussian pyramid created with the lithological covariate, allowing geological information to be presented at multiple scales. We then take the map whose correlation or probabilistic distance is optimal for our interpolation.

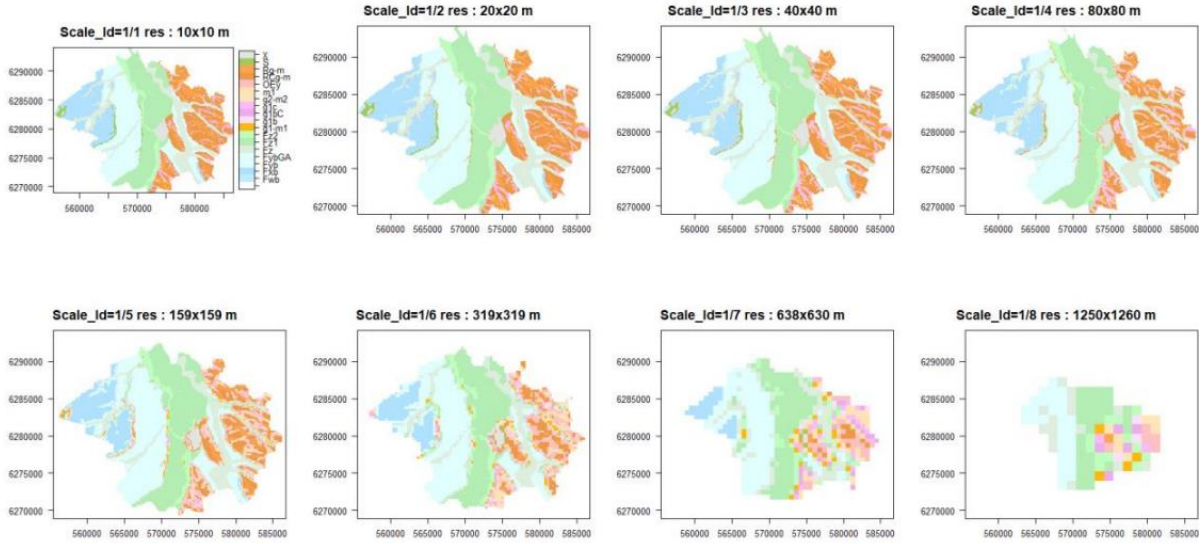


**Figure 112: Truncated Gaussian pyramid of lithological data for the Toulouse metropolitan area.**

The following figure shows the energetic distance for the parameter distance to roads for TPHs; it is an aid to choosing the resolution of the covariate providing the maximum information.
In addition, if an exponential can be calibrated to the data (Figure 113), a latent diffusive process on calculations at scale becomes a credible hypothesis, as theorized by Koendering (1984) and Lindeberg (1994). It is then possible to do one of the following:

- Interpolate on a fine grid, which we deteriorate by applying a diffusive kernel (we then find a method comparable to that of Aberle et al., 2023, who used diffusion on Voronoi polygons).
- As we do it, we degrade the covariate using a Gaussian pyramid and apply diffusion with EEPH.





Funded by the European Union

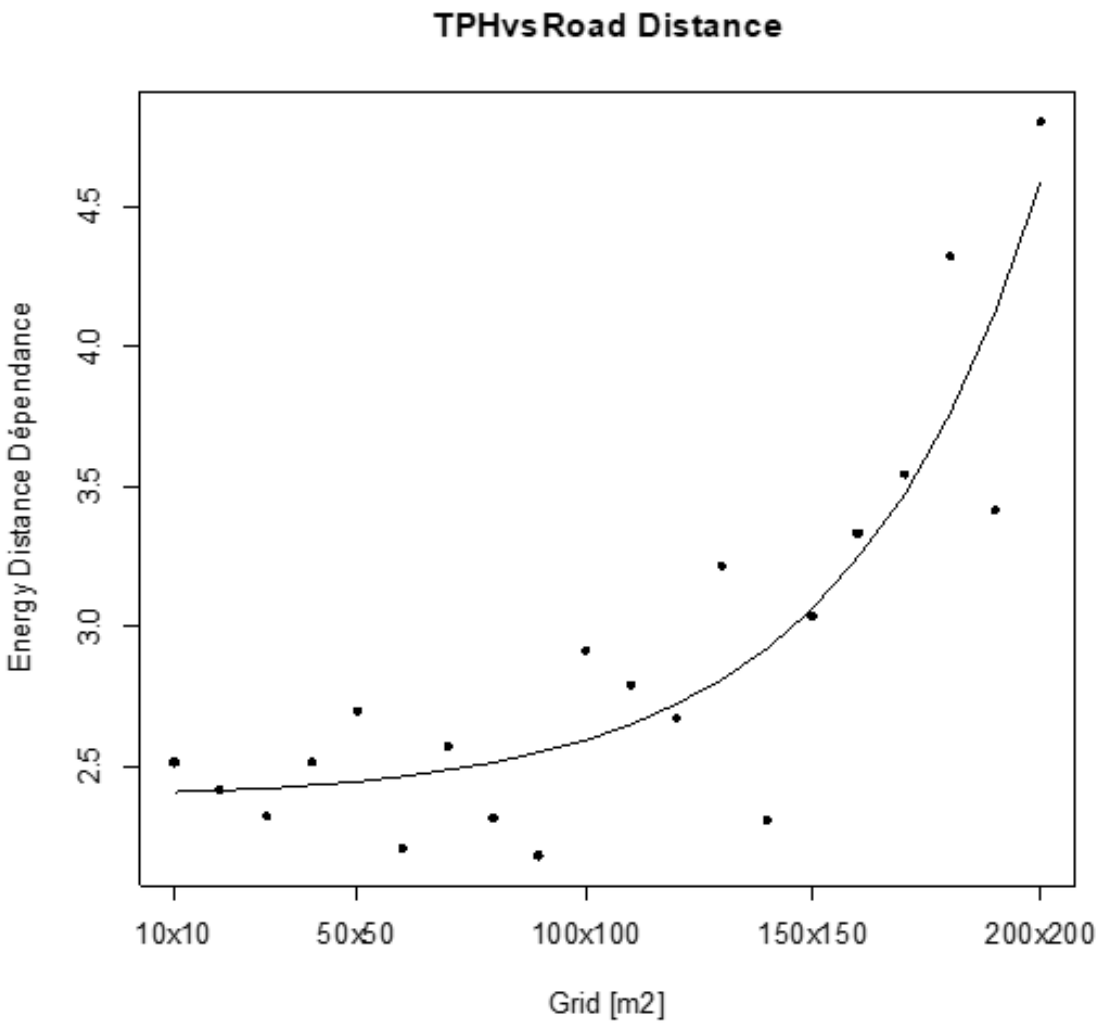


**Figure 113: TPH vs. road distance between scale and energy distance dependence**





Funded by
the European Union

## 5.4. Toulouse ITA3 covariate applications

### 5.4.1. Overall statistics

We see inTable 30 that the parameters with the greatest statistical explanatory power for TPH content are LANU uses.

**Table 30: Calculation of energetic and Wasserstein distances for various covariates for TPH prediction**

| | Euclidean field | Geological | | Land uses | | | LANU based on ADEME (2018) dichotomy | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Distance | ALT | LITHO | GEOLE | CLC | Copernic | | GEOBAP | COPER H2 | GEOBAP H2 | GEO_COPER H2 |
| Energy | 0.00 | 0.92 | 0.60 | 0.99 | 1.00 | | 0.73 | 0.01 | 0.48 | 0.70 |
| Wasserstein | 0.05 | 0.95 | 0.70 | 0.98 | 1.00 | | 0.78 | 0.09 | 0.57 | 0.79 |

| Distance | dist_basias | dist_basol | dist_center | dist_water | dist_route |
|---|---|---|---|---|---|
| Energy | 0.002 | 0.003 | 0.001 | 0.001 | 0.01 |
| Wasserstein | 0.03 | 0.04 | 0.02 | 0.02 | 0.06 |

### 5.4.2. Confirmation of statistical weight measurements by cross-validation of covariate combinations

Using a cross-validation method (Hawkins et al., 2003), we can use our measured $z$ sparse contents to assess the quality of interpolation $\hat{z}$ with the mean absolute error (MAE), the root mean squared error (RMSE) and the linear error in probability space (LEPS)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|z_i - \hat{z}_i\ |$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i\ )}$$

$$LEPS = \frac{1}{n}\sum_{i=1}^{n}|F_{zi} - \hat{F}_{\hat{z}i}\ |$$

if we apply this cross-validation to Toulouse's covariates and combine them (127 possibilities). We obtain multiple runs (Figure 114) that allow us to know the best combination reproducing our 136 data points. It is also possible to test a possible overall weighting of the data.





Funded by the European Union

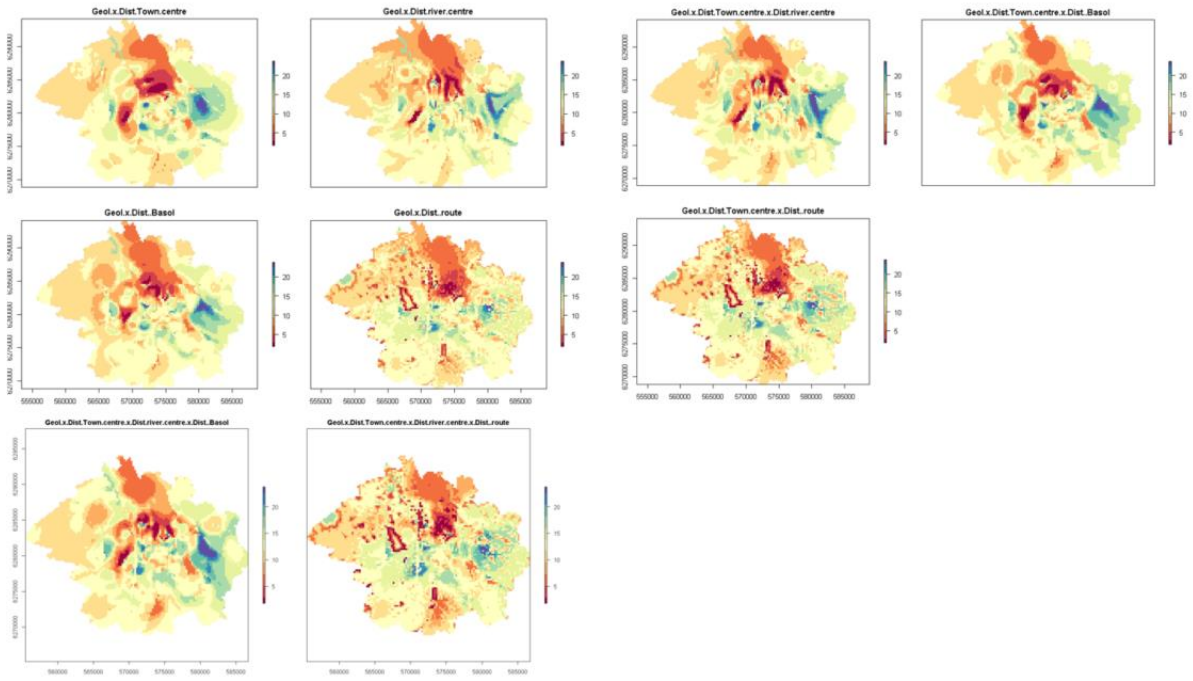


**Figure 114: Possible spatial EEPHs obtained by combining arsenic covariates with ITA 3 data**

The cross-validation results of the 127 possible covariate combinations are compared, and those giving minimum indicators are retained. For our Toulouse dataset, the dataset is SIC and the three indicators MAE, RMSE and LEPS developed for datasets with more than 5000 data points (Gnann et al., 2018, Abraj et al., 2022) give the results found in Table 31.

**Table 31: Cross-validations of TPH results with 127 combinations of covariables for ITA3: Toulouse**

| | Min | Med | Max | Arg min parameter set |
|---|---|---|---|---|
| MAE | 19.39 | 23.4 | 24.71 | LANU |
| RMSE | 57.49 | 70.05 | 72.63 | LANU |
| LEPS | 0.0048 | 0.0054 | 0.0078 | LANUxDBASIASxDBASOLxDR |

This test on 127 combinations confirms the choice of the single covariate LANU for the best interpolation in terms of the lowest MAE and RMSE. For an imprecise probability calculation, all covariates can be used in a belief as suggested by LEPS.

In a SIC context such as that of ITA3 Toulouse, we use the cross-validation method because we don't have enough data to do otherwise. A much more efficient mathematical method would be to divide the data into two parts (by random selection), reconstitute a pollution map using the first set and compare it with reality, i.e. the second set. This high-performance method was successfully used on the HOUSES dataset.





Funded by the European Union

# 6. A plan for the public in the face of uncertainty

## 6.1. Literature review

Generally speaking, members of the public have a poor understanding of probabilities, error bars, and abstract concepts, and react better if they have experience with such graphics and are prepared to take an interest in them. Presenting uncertainties to the public in the field of contaminated sites and soils has many dimensions, and if we want to communicate uncertainty effectively. Graphic representations of uncertainty can benefit from the advances made in communicating global warming, cyclone paths, disasters and so on.

- Legends can be shaped or colored, but there should also be an explanatory sentence with them so they will be easier to understand (Joslyn and Savelli, 2021). With regard to colors, spectral legends are the most popular.
- Impact maps (Cheong, 2016; Bostrom et al., 2018), spaghetti diagrams (Tak and Toet, 2014; Padilla et al., 2017; Kale et al., 2019; Mulder et al, 2019; Kinkeldey et al, 2015), sampling prediction ensembles (Liu et al. 2016) and animations (Witt et al., 2020, 2021; Hullman et al., 2015) are much better understood than probabilities, imprecise probabilities or anything else.
- For treatment probabilities, everyone understands their "natural" representation using black and white dots, as in the medical field.

The clear winners for presenting uncertainty are the impact map, text and interval, spaghetti diagrams, kinematics, and sampling prediction ensembles. Figures 115 to 121 show what these advanced visual aids look like.



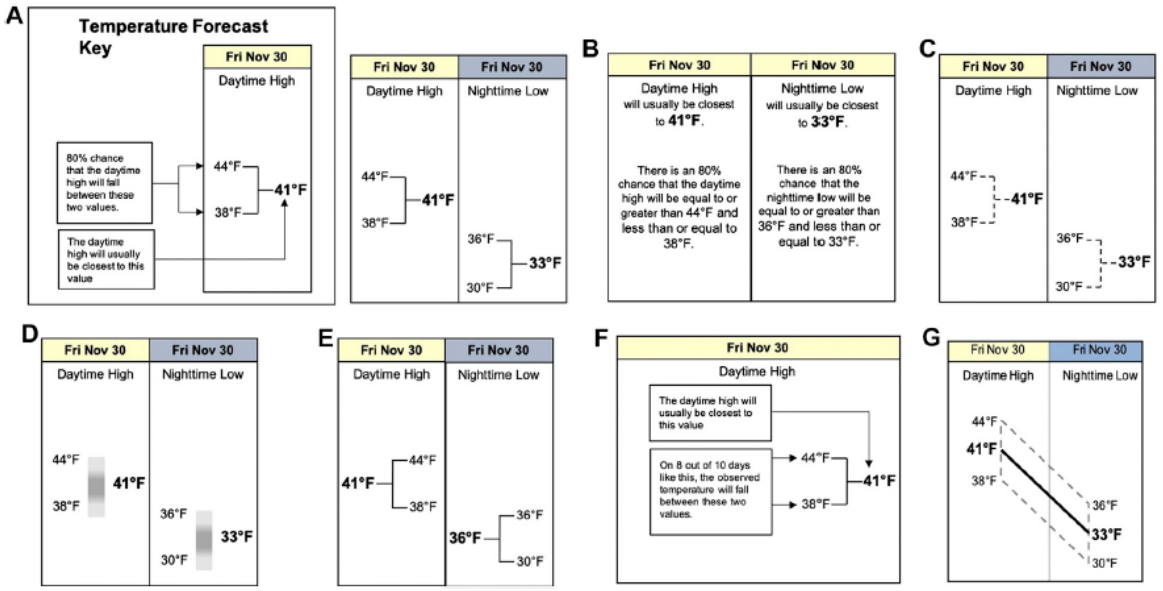


FIGURE 1 | Predictive intervals, each accompanied by a key (shown in a) describing "41°F" as the "best forecast".

**Figure 115: 8 weather prediction figures (Joslyn and Savelli, 2021).**



Funded by the European Union

*Note: figure B is the most effective. Figure A is an example of a text legend.*

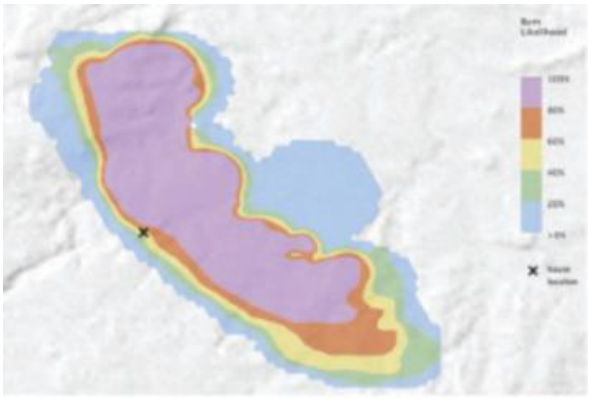

**Figure 116: Fire destruction map (80% to 100% destruction probability), Cheong (2016)**

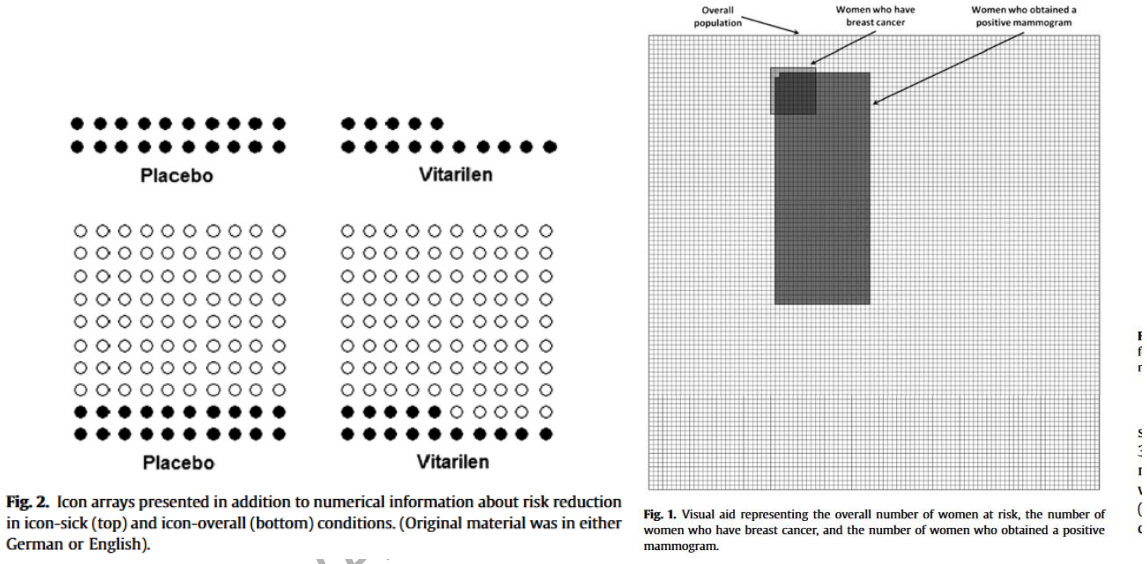


**Fig. 2.** Icon arrays presented in addition to numerical information about risk reduction in icon-sick (top) and icon-overall (bottom) conditions. (Original material was in either German or English).

**Fig. 1.** Visual aid representing the overall number of women at risk, the number of women who have breast cancer, and the number of women who obtained a positive mammogram.

**Figure 117: Natural frequency displays**

*with a) visualization of probabilistic results of a Garcia-Retamero and Galesic (2010) drug test and b) breast cancer rates in the general population and false positives from Garcia-Retamero and Hoffrage (2013)*

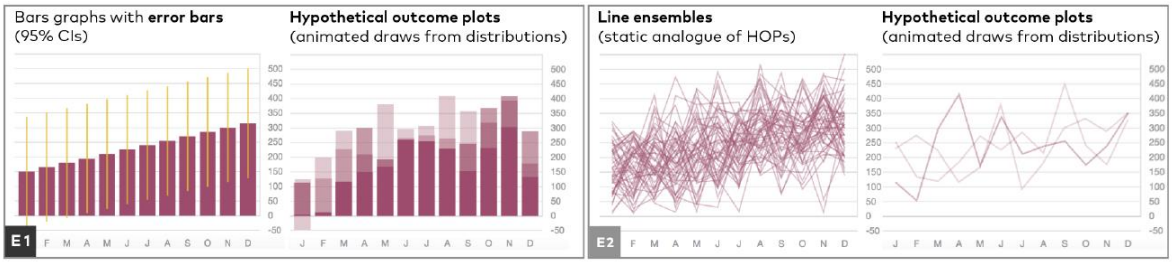


Fig. 1. We present two experiments (E1 and E2) evaluating four different uncertainty visualizations (from left to right): bar graphs wi error bars, bar hypothetical outcome plots (HOPs), static line ensembles, and line HOPs.

**Figure 118: 4 maps of the same thing, including the spaghetti map in 3), which is the most informative about uncertainty, according to Kale et al., 2019**



Funded by the European Union

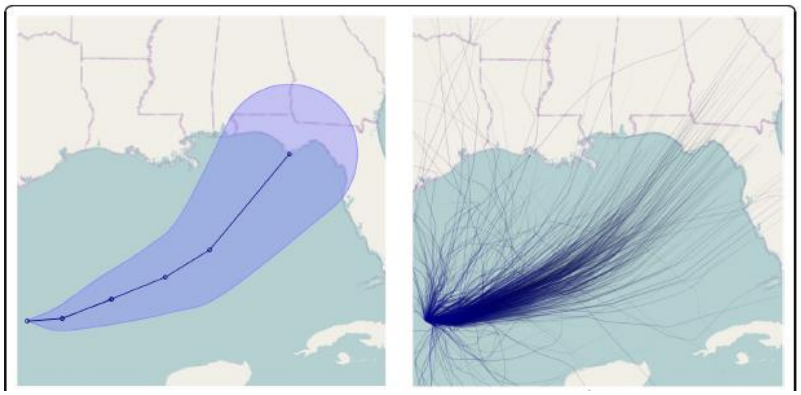

**Figure 119: 2 maps of the same thing, which is the most informative, according to Kale et al., 2017**

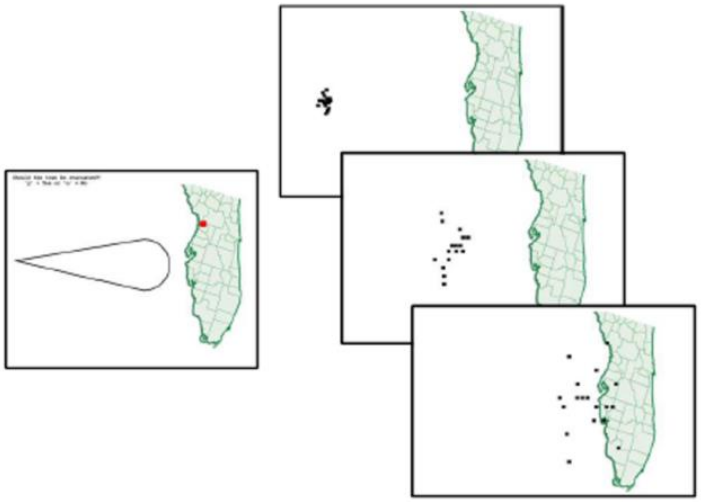

**Figure 120: Dynamic ensembles according to Witt et al. (2021)**

*With a) cyclonic evacuation cone and b) dynamic ensemble of the cone showing better results for understanding and evacuating the area.*

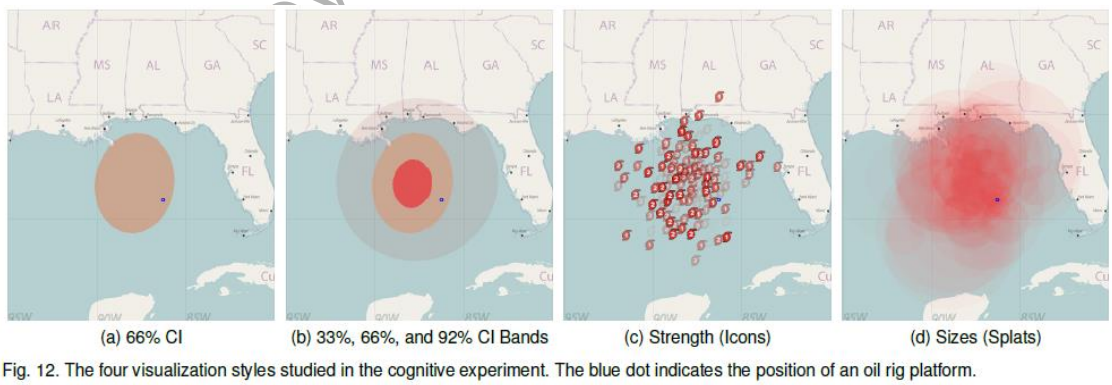


Fig. 12. The four visualization styles studied in the cognitive experiment. The blue dot indicates the position of an oil rig platform.

**Figure 121: 4 representations of the same thing, including the subset map (swarms of possibilities) in 3), which is the most informative about uncertainty, according to Liu et al., 2016.**

Since the interpolation work package is intended to produce informative maps, a public version will be designed in line with previous authors' recommendations, especially regarding color and text legends. For calculations in imprecise probability mode, an



Funded by the European Union

alternative display of spaghetti distributions or map subsets (sampling prediction ensembles) may be proposed.

## 6.2. Visualization of imprecise probabilities

In probability mode, the EEPH algorithm provides 4D data (X, Y, C, proba, or possi). A first natural display would be in the form of a tensor cube (Figure 122). As we have seen, this may speak to a scientist, but not to the public.

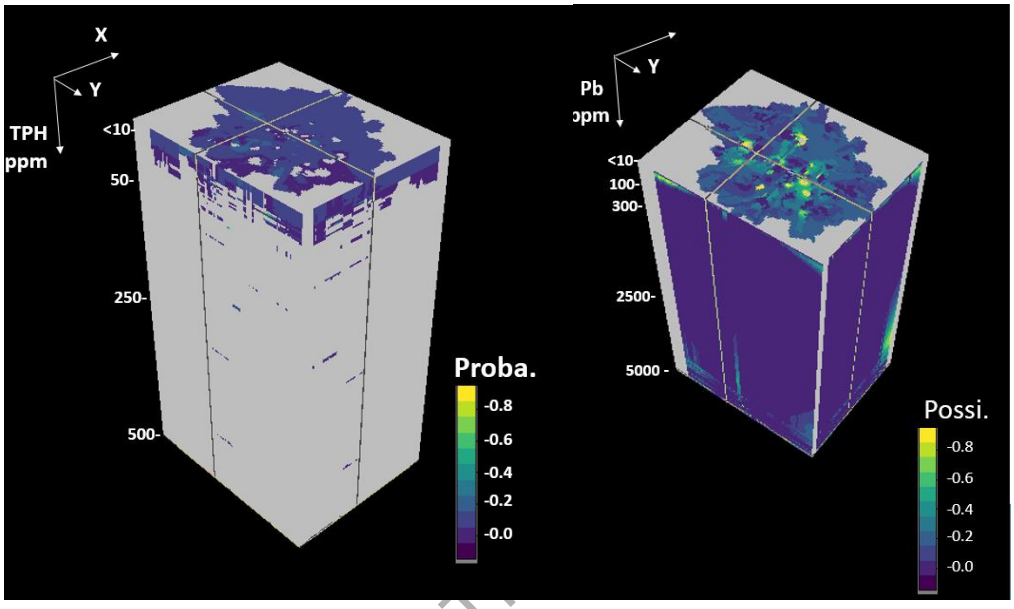


**Figure 122: Full probabilistic EEPH interpolation method for mapping total petroleum hydrocarbon (TPH) on the left and lead (Pb) on the right, possible results in the city of Toulouse**

*Note: the lines that cross are the slice position displayed on the cube view sides. Topsoil samples from Belbeze et al. (2019), n=139, 0–10cm, TPH analysis by multiple laboratories, LOQ: 10 mg/kg (8 samples), 20 mg/kg (57 samples). Pb analysis by multiple laboratories, LOQ: 10 mg/kg (only 3 samples below LOQ).*

Inspired by the work of Liu et al. (2017), we are proposing to create an animation of the various quantiles of the EEPHs produced and a 2D mapping test with glyphs conveying the remaining 2 or 3 dimensions on a representative sample.

A GIF can be used to animate the various quantiles presented in Figure 123. It has the merit of showing the extent of the possible, but may produce an anxiety-inducing effect by placing low probabilities on the same level as the others.





Funded by the European Union

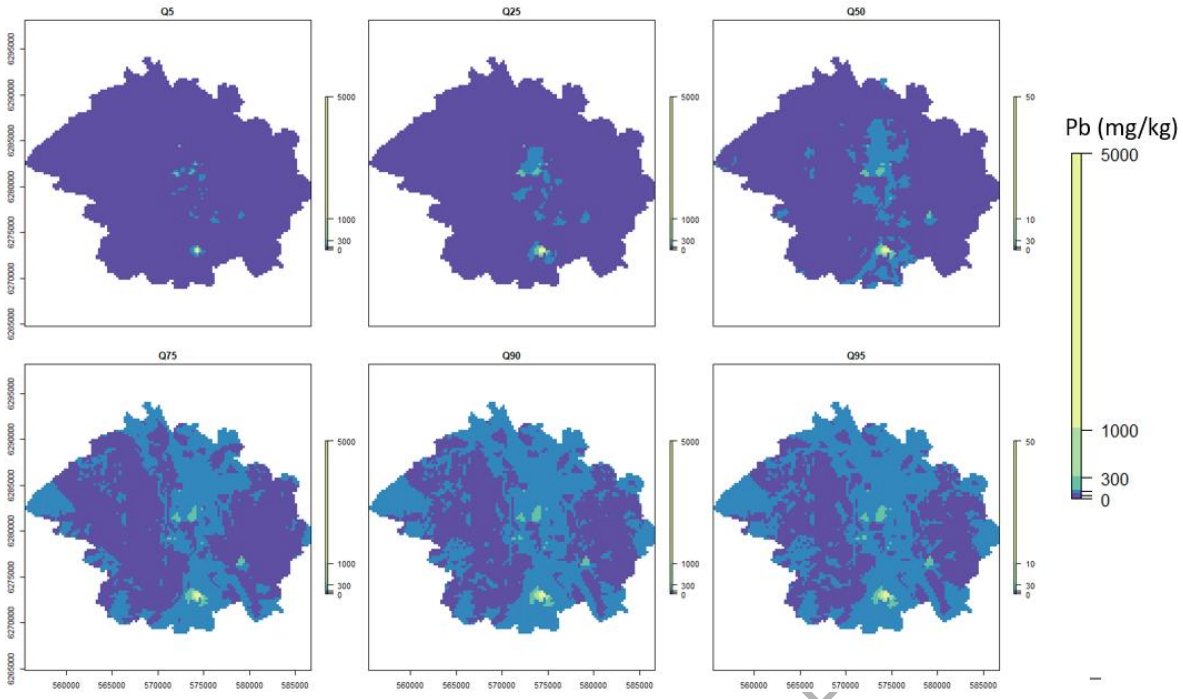


**Figure 123: Full probabilistic EEPH interpolation method for mapping lead (Pb) in the city of Toulouse. Some centiles.**

Topsoil samples from Belbeze et al. (2019), n=139, 0–10cm, Pb analysis by multiple laboratories, LOQ: 10 mg/kg (3 samples below).

To conduct ensemble prediction sampling, we first perform a spatial partitioning of the median contents into four distinct zones (these zones could be interpreted as low background, high background, low anomaly, high anomaly). We then take a random number of samples proportional to the surface area of the zones (Figure 124). We then create a map showing circles of increasing size with the concentration and a color based on the percentile of realization, which highlights the entire probabilistic calculation rather than just the expected value (**Error! Reference source not found.**).

It should also be noted that this calculation rightly raises questions about lead levels in industrial zones to the south of the perimeter, as well as in little-known natural zones.





Funded by the European Union

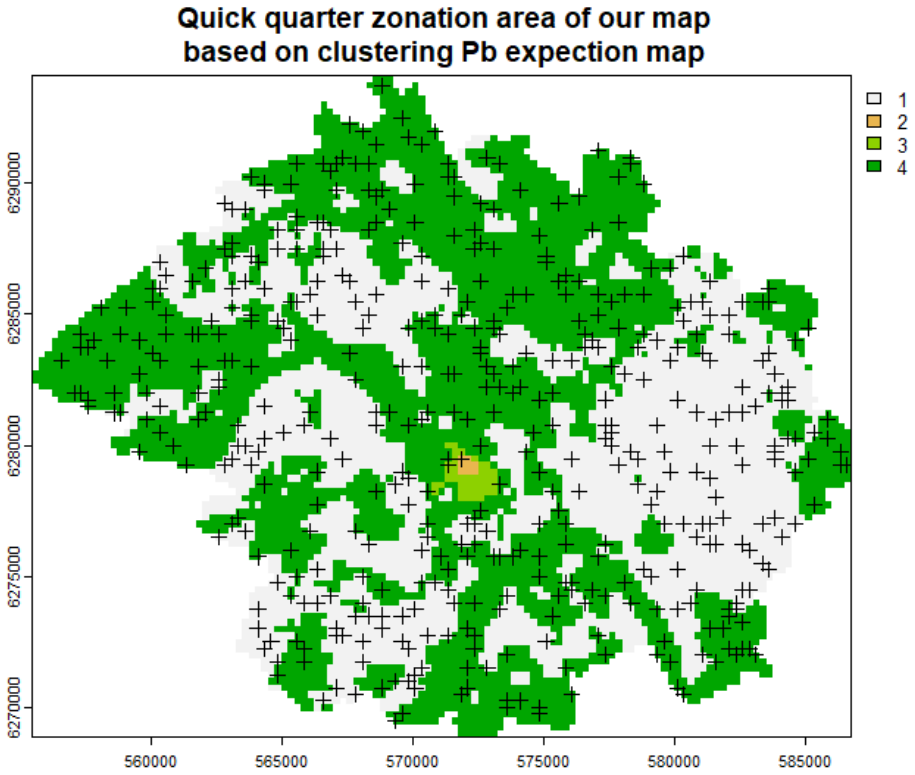


**Figure 124: Selection of representative samples forthe quick zonation of Pb expectation in the city of Toulouse**

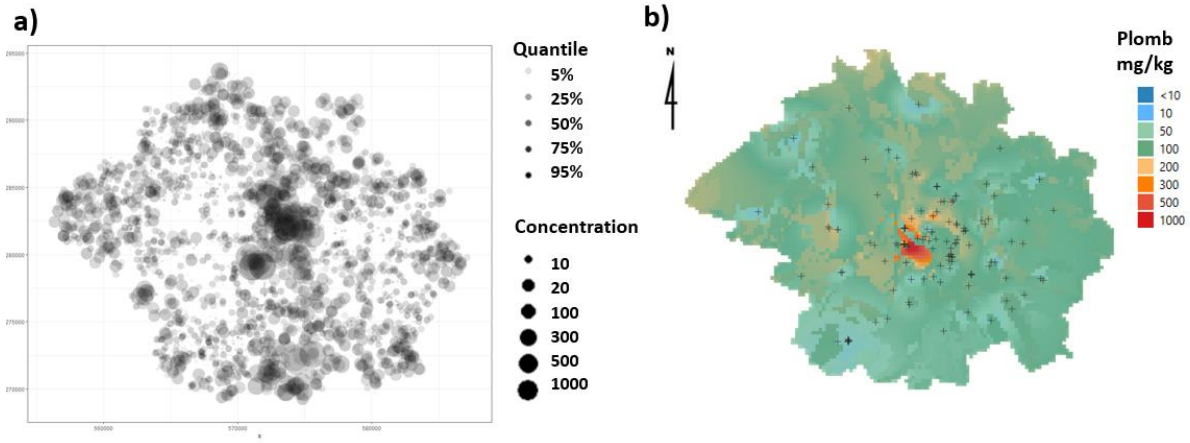


**Figure 125: on ensembles visualization (a) and (b), EEPH expectation for mapping lead (Pb) in the city of Toulouse**

Topsoil samples from Belbeze et al. (2019), n=139, 0–10cm, Pb analysis by multiple laboratories, LOQ: 10 mg/kg (3 samples).





Funded by
the European Union

# 7. Initial applications

## 7.1. Arsenic in France: all available soil data

There are several maps of arsenic levels in surface soil in France, all produced by different organizations, with different datasets and covariates sampled for various purposes. Figure 126 shows selected examples. The variety of legends used to assign a range for arsenic in French soil is of particular interest. These differ not only in the choices made by their authors, but also in the epistemic assumptions of the various models used. Without direct access to the rasters of these models, it is not possible to scale their content for scientific comparison.

A common thread in all these studies is the major role played by soil geogeny, S mineralization, and mining districts in explaining arsenic levels.

- Very low arsenic concentrations in the sands of Landes (1), Sologne (2) and the northern Vosges area (3).
- Low concentrations in the Paris Basin and in Quaternary terrain in general (4).
- High arsenic concentrations in Vosges (5), Limousin (6), and Cevennes (7) due to surface mineralization.
- High arsenic concentrations in Lorraine (8) due to its mineralurgical and coal mining past.
- Low concentrations in southern vineyard areas (9) where arsenic-based pesticides have been used. It is assumed that the sampling depth of 0–30 cm mitigates this impact (Marchand et al., 2017).
- Low concentrations in northern France (10). This seems contradictory with its coal mining past. The grid size used may miss kilometer-scale mining anomalies (Marchant et al., 2017). This is also the opinion of the author and the map in (Figure 97 d), adds mineralization data to the RMQS data, thus providing a more accurate map of France.

For ISLANDR, if we want to identify all contamination and extract the diffuse background, we will need to include all available data on several scales in order to miss as few anomalies as possible.





Funded by the European Union

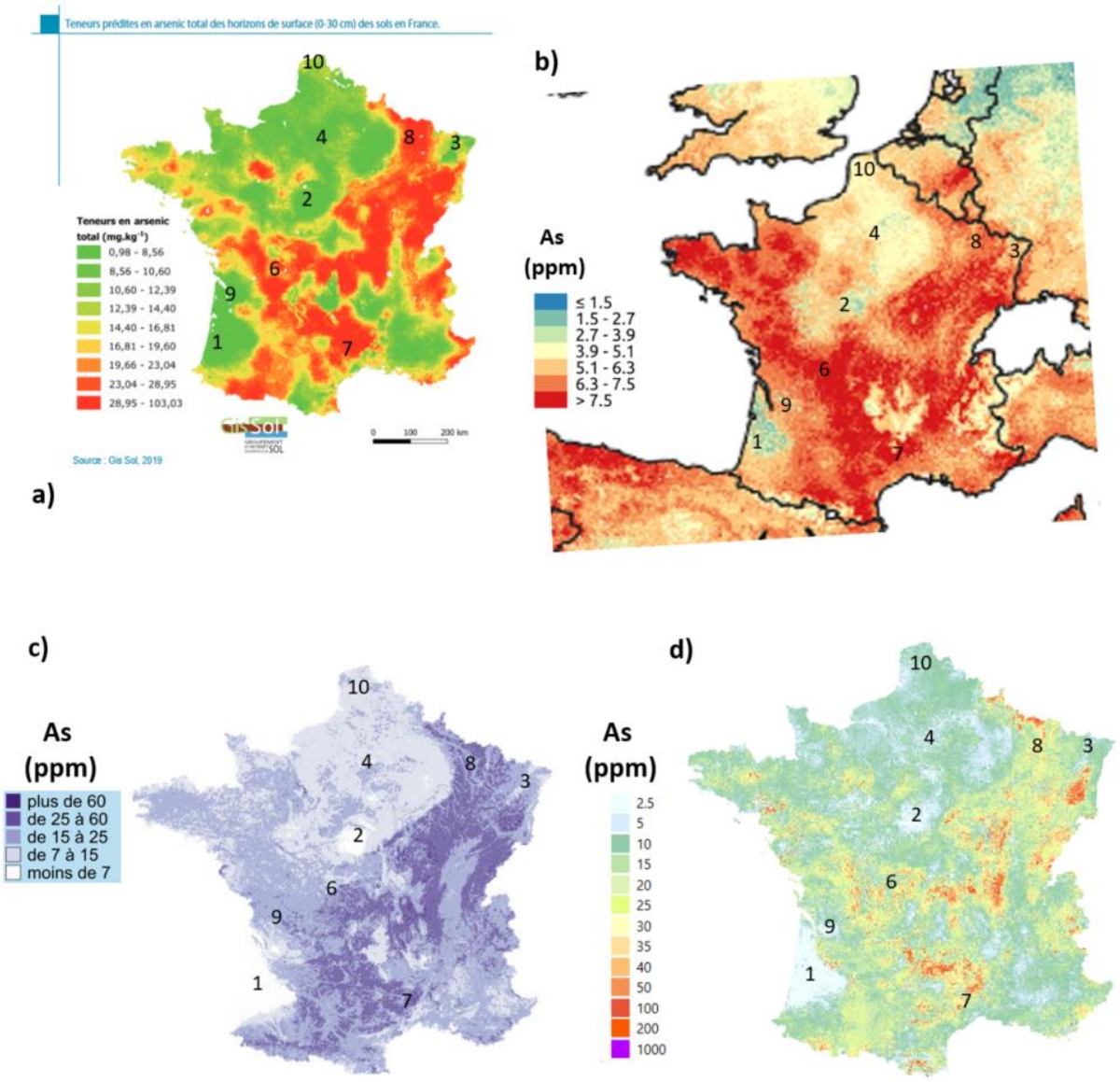


**Figure 126: Four maps of As in France**

*With a) As in soil from Marchant et al. (2017). 0–30 cm with 25 core on a 20 m x 20 m loose grid, RMQS regular survey with rate of 1/250 km². Hydrofluoric acid extraction. Trans-Gaussian linear mixed model with covariates. b) As in soil from LUCAS from Fendrich et al. (2017). 0.20 cm with 5 core on a 4 m × 4 m cross, soil samples with mean rate of 1/200 km². Aqua regia extraction. Coupled semi-parametric GAMLSS-RF model with covariates. c) As in soil from Antoni and Jamet (2021). Same set as Marchant et al. (2017). External drift Kriging with covariates (KED). d) As in soil from the author (2022, unpublished). Same set as Marchant et al. (2017) with additional mining exploration soil dataset, 1/km². Hydrofluoric acid extraction. Coupled RF – trans-Gaussian kriging*

Data from the mining and towns inventory (BdSolU) does not fall within the scope of the FAIR policy and can only be disclosed when necessary and as little as possible. This is primarily due to the presence of private data in the databases. Nevertheless, it is possible to apply EEPH to a pool of 38,000 data points (**Error! Reference source not found.**Table 32) and see the results.





Funded by the European Union

**Table 32: Available data collected by BRGM and used for mapping with EEPH**

| Kind of soil sampled | Nb. | Min. | Q25 | Median | Mean | Q75 | Max. |
|---|---|---|---|---|---|---|---|
| Agricultural – Public RMQS data | 2153 | 0.4 | 8 | 12 | 17.7 | 19.5 | 412 |
| Cities – BDSOLU – restricted access | 2614 | 1 | 7.6 | 11.1 | 18.3 | 17 | 1670 |
| Mineralizations – BRGM – non public | 35174 | 1 | 12 | 28 | 90.3 | 65 | 55736 |

The resulting map (Figure 127) shows various kilometer-scale anomalies superimposed on the major trends already defined on previous maps. The biggest anomalies are due to mineralization data and, to a lesser extent, to urban pollution. A map of this kind is ideal for detecting ISLANDR anomalies on a national scale.

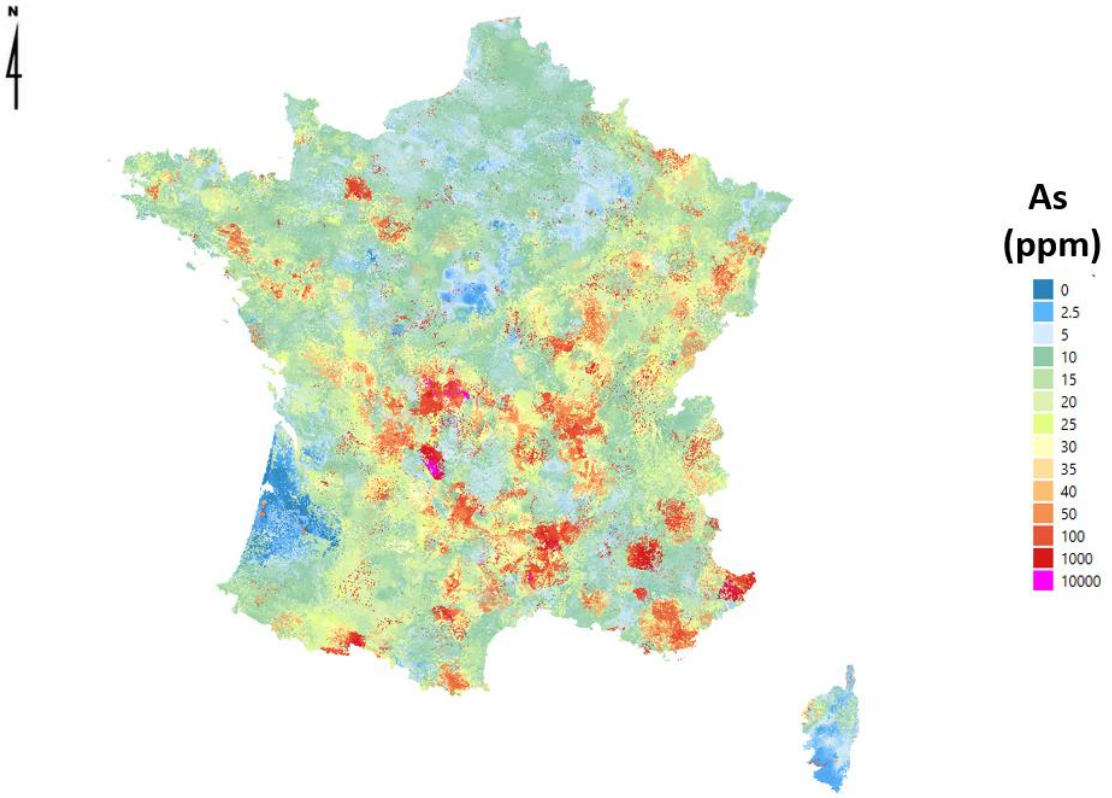


**Figure 127: As expectation in French soil, all available soil data. EEPH with LANU-optimized covariable**





Funded by the European Union

## 7.2. ITA3: City of Toulouse

Examination of the spatial contents of this ITA shows complex distributions with sparse, imprecise, and poorly distributed data (SIC, i.e., sparse, imprecise, clustered). This results in complex histograms with multimodal trends, punctuated by anomalies and outliers (Figure 129) and data that is particularly difficult to map (Belbeze et al., 2019). An experiment conducted by Belbeze et al. (2023) showed that each method can produce a different map, some of which can unnecessarily panic the general public. EEPH calculations produced the maps in Figure130 and 131 Using construction, these maps show the full range of values found in surface soils, with an amplification of anomalies.

It should be noted that in 2024, a local group informed the classified facilities authorities of lead pollution from a chimney in ITA3 in the Minimes neighborhood and on the banks of the canal (Figure 128). This zone is clearly shown on the EEPH map produced, which could prove to be a powerful decision-making tool. The stakes for health are high: the contamination plume covers 12,000 people; 322 blood tests are currently being performed, and the yards of 89 homes will be decontaminated.

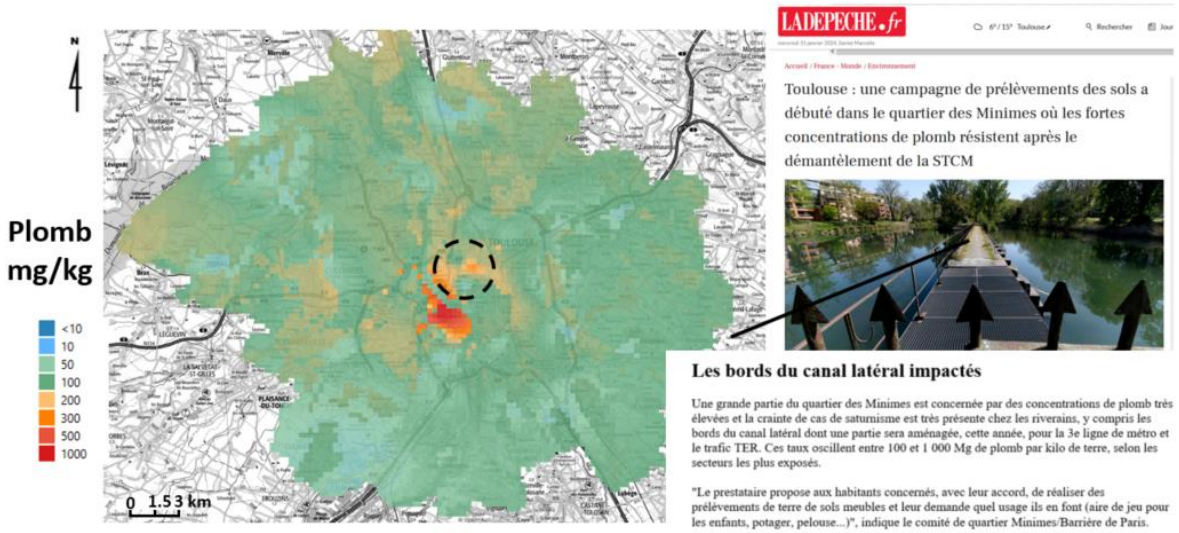


**Figure 128: EEPH lead enhanced maps and French newspaper clips of the discovery of an environmental issue.**

*Note: in ITA3, normal background for lead is 100 mg/kg, sanitary threshold is 300 kg*





Funded by the European Union

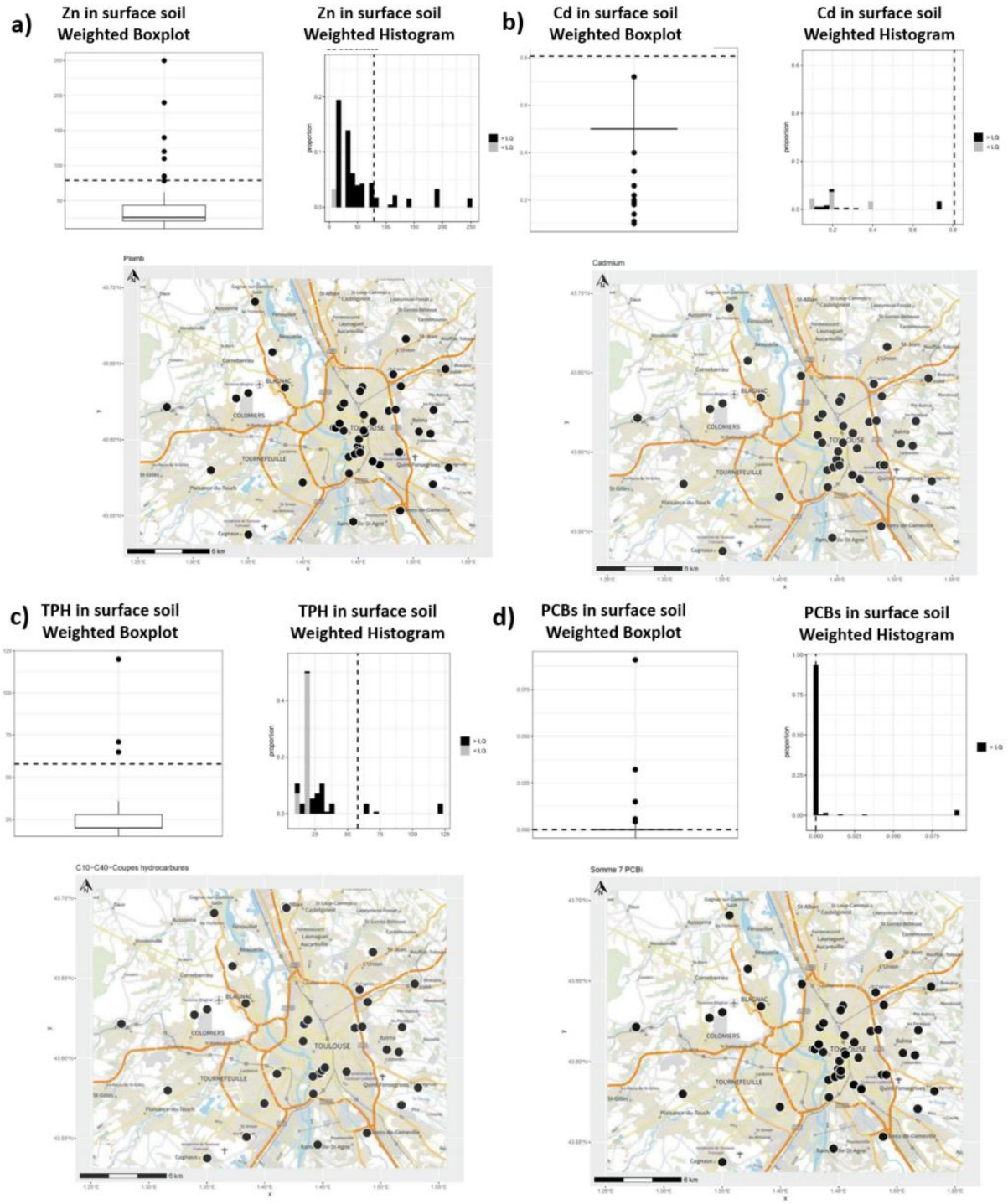


**Figure 129: Weighted histogram and box plot and localization for ITA3 surface soil selected pollutant content**

*with a) topsoil samples from Belbeze et al. (2019), n=138, 0–10 cm, As analysis by multiple laboratories, LOQ: 1 mg/kg (5 samples); b) topsoil samples from Belbeze et al. (2019), n=138, 0–10 cm, Cd analysis by multiple laboratories, LOQ: 0.1 mg/kg (5 samples), 0.2 mg/kg (16 samples), 0.4 mg/kg (10 samples), 0.5 mg/kg (64 samples), 0.6 mg/kg (1 sample), with c) topsoil samples from Belbeze et al. (2019), n=139, 0–10 cm, TPH (C10C40) analysis by multiple laboratories, LOQ: 10 mg/kg (8 samples), 20 mg/kg (57 samples) and d) topsoil samples from Belbeze et al. (2019), n=112, 0–10 cm, PCB sum (7) analysis by multiple laboratories, LOQ: 0.007 mg/kg (14 samples), 0.0078 mg/kg (1 sample),0.01 mg/kg (57 samples), 0.02 mg/kg (1 sample), 0.07 mg/kg (30 samples).*





Funded by the European Union

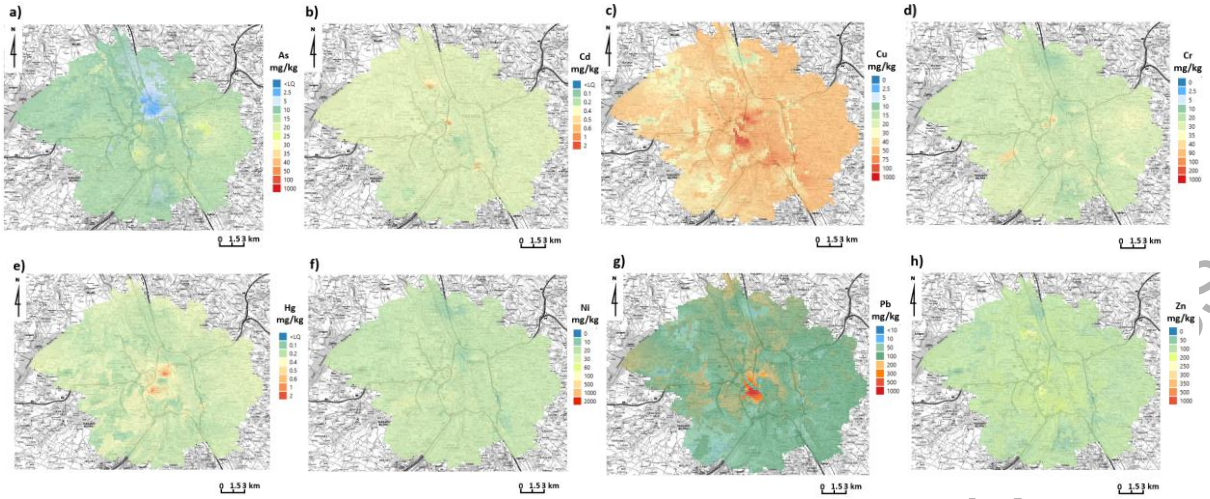


**Figure 130: EEPH expectation with optimized LANU covariate for ITA3 surface soil metal(oids) content**

*with a) topsoil samples from Belbeze et al. (2019), n=138, 0–10 cm, As analysis by multiple laboratories, LOQ: 1 mg/kg (5 samples); b) topsoil samples from Belbeze et al. (2019), n=138, 0–10 cm, Cd analysis by multiple laboratories, LOQ: 0.1 mg/kg (5 samples), 0.2 mg/kg (16 samples), 0.4 mg/kg (10 samples), 0.5 mg/kg (64 samples), 0.6 mg/kg (1 sample); c) topsoil samples from Belbeze et al. (2019), n=138, 0–10 cm, Cu analysis by multiple laboratories, LOQ: 1 mg/kg; d) topsoil samples from Belbeze et al. (2019), n=138, 0–10 cm, Cr analysis by multiple laboratories, LOQ: 1 mg/kg; e) topsoil samples from Belbeze et al. (2019), n=138, 0–10 cm, Hg analysis by multiple laboratories, LOQ: 0.05 mg/kg (16 samples), 0.10 mg/kg (57 samples); f) topsoil samples from Belbeze et al. (2019), n=139, 0-10 cm, Ni analysis by multiple laboratories, LOQ: 1 mg/kg; g) topsoil samples from Belbeze et al. (2019), n=138, 0–10 cm, Pb analysis by multiple laboratories, LOQ: 10 mg/kg (3 samples); and h) topsoil samples from Belbeze et al. (2019), n=138, 0–10 cm, Zn analysis by multiple laboratories, LOQ: 10 mg/kg .*





Funded by the European Union

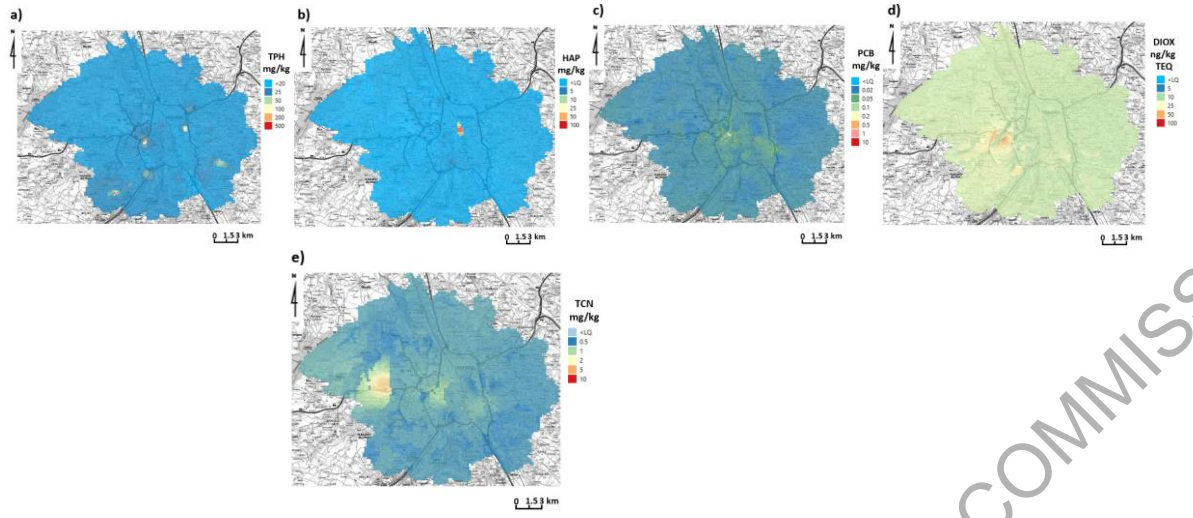


**Figure 131: EEPH expectation with optimized LANU covariate for ITA3 organics compounds.**

*With a) topsoil samples from Belbeze et al. (2019), n=139, 0–10 cm, TPH (C10C40) analysis by multiple laboratories, LOQ: 10 mg/kg (8 samples), 20 mg/kg (57 samples); b) topsoil samples from Belbeze et al. (2019), n=138, 0–10cm, PAH sum (16) analysis by multiple laboratories, LOQ: 0.01 mg/kg (22 samples), 0.16 mg/kg (1 sample), 0.32 mg/kg (16 samples), 0.48 mg/kg (2 samples) 0.80 mg/kg (17 samples), 0.94 mg/kg (1 sample); c) topsoil samples from Belbeze et al. (2019), n=112, 0–10cm, PCB sum (7) analysis by multiple laboratories, LOQ: 0.007 mg/kg (14 samples), 0.0078 mg/kg (1 sample),0.01 mg/kg (57 samples), 0.02 mg/kg (1 sample), 0.07 mg/kg (30 samples); d) topsoil samples from Belbeze et al. (2019), n=38, 0-10 cm, dioxin toxic equivalent (DIOX) analysis by multiple laboratories, LOQ: 2 ng/kg TEQ (6 samples), 3 ng/kg TEQ (4 samples), 4 ng/kg TEQ (3 samples), 5 ng/kg TEQ (3 samples), 6 ng/kg TEQ (3 samples); e) topsoil samples from Belbeze et al. (2019), n=108, 0-10 cm, Total cyanide (TCN) analysis by multiple laboratories, LOQ: 0.1 mg/kg (27 samples), 0.5 mg/kg (11 samples), 1 mg/kg (16 samples).*





Funded by the European Union

# Annex 1 Extended Bibliography

**Aberle, M. G., Robertson, J., & Hoogewerff, J. A. (2023).** Voronoi Natural Neighbours Tessellation: An interpolation and grid agnostic approach to forensic soil provenancing. Forensic Chemistry, 35, 100522. https://doi.org/10.1016/j.forc.2023.100522

**Abraj, M., Wang, Y.-G., & Thompson, M. H. (2022).** A new mixture copula model for spatially correlated multiple variables with an environmental application. Scientific Reports, 12(1), 13867. https://doi.org/10.1038/s41598-022-18007-z

**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). Journal of Geochemical Exploration, 93(1), 21–34. https://doi.org/10.1016/j.gexplo.2006.07.006

**Albert, C. G., & Rath, K. (2020).** Gaussian Process Regression for Data Fulfilling Linear Differential Equations with Localized Sources. Entropy, 22(2), 152. https://doi.org/10.3390/e22020152

**Ali, M., Son, L. H., Khan, M., & Tung, N. T. (2018).** Segmentation of dental X-ray images in medical imaging using neutrosophic orthogonal matrices. Expert Systems with Applications, 91, 434–441. https://doi.org/10.1016/j.eswa.2017.09.027

**Andrade, R., Silva, S. H. G., Weindorf, D. C., Chakraborty, S., Faria, W. M., Guilherme, L. R. G., & Curi, N. (2021).** Micronutrients prediction via pXRF spectrometry in Brazil: Influence of weathering degree. Geoderma Regional, 27, e00431. https://doi.org/10.1016/j.geodrs.2021.e00431

**Anter, A. M., & Hassenian, A. E. (2019).** CT liver tumor segmentation hybrid approach using neutrosophic sets, fast fuzzy c-means and adaptive watershed algorithm. Artificial Intelligence in Medicine, 97, 105–117. https://doi.org/10.1016/j.artmed.2018.11.007

**Antoni, V., & Jamet, C. (2021).** Arsenic et mercure dans les sols: Les zones exposées en France. SDES, Datalab Essentiel, 4.

**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. International Journal of General Systems, 44(5), 601–614. https://doi.org/10.1080/03081079.2014.980242

**Ballabio, C., Jones, A., & Panagos, P. (2024).** Cadmium in topsoils of the European Union – An analysis based on LUCAS topsoil database. Science of The Total Environment, 912, 168710. https://doi.org/10.1016/j.scitotenv.2023.168710





Funded by the European Union

**Barbu, T. (2019).** Novel Diffusion-Based Models for Image Restoration and Interpolation. Springer International Publishing. https://doi.org/10.1007/978-3-319-93006-0

**Bárdossy, G., & Fodor, J. (2004).** Evaluation of uncertainties and risks in geology: New mathematical approaches for their handling. Springer.

**Bartkute, V., & Sakalauskas, L. (2008).** Experimental probabilistic hypersurface construction by Gaussian fields. Institute of Mathematics and Informatics.

**Baudrit, C., Dubois, D., & Guyonnet, D. (2006).** Joint Propagation and Exploitation of Probabilistic and Possibilistic Information in Risk Assessment. IEEE Transactions on Fuzzy Systems, 14(5), 593–608. https://doi.org/10.1109/TFUZZ.2006.876720

**Beauzamy, B. (2004).** Méthodes probabilistes pour l'étude des phénomènes réels. Société de calcul mathématique.

**Beauzamy, B. (2010).** Nouvelles méthodes probabilistes pour l'évaluation des risques. Société de calcul mathématique.

**Beauzamy, B., & Bradat, M. (2009).** Probabilistic method for epidemiology , Algorithmes et optimisation (p. 35) [Bonnes pratiques]. Societé de Calcul Mathematique S.A. http://www.scmsa.eu/archives/SCM_Bonnes_pratiques_epidemio_2009b.pdf

**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. European Journal of Soil Science, 69(5), 757–770. https://doi.org/10.1111/ejss.12687

**Behrens, T., Viscarra Rossel, R. A., Kerry, R., MacMillan, R., Schmidt, K., Lee, J., Scholten, T., & Zhu, A.-X. (2019).** The relevant range of scales for multi-scale contextual spatial modelling. Scientific Reports, 9(1), 14800. https://doi.org/10.1038/s41598-019-51395-3

**Belbeze, S. (2018).** Expertise technique concernant la représentation surfacique des concentrations en nitrates du projet ERMES 2016 (Rapport Final Confidentiel No. RP-69502-FR; p. 16). BRGM.

**Belbeze, S. (2023).** Expertise sur le rapport de plan de gestion et analyse des risques résiduels. Atoll de Hao, archipel des Tuamotu, Polynésie française (Confidentiel No. RP-72412-FR; p. 48). BRGM.

**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; p. 347). BRGM.

**Bera, T., & Mahapatra, N. K. (2017).** On Neutrosophic Soft Linear Spaces. Fuzzy Information and Engineering, 9(3), 299–324. https://doi.org/10.1016/j.fiae.2017.09.004





Funded by the European Union

**Berton, G. (2018).** Comparison between two interpolation methods: Kriging and EPH. Journal of Physics: Conference Series, 1141, 012130. https://doi.org/10.1088/1742-6596/1141/1/012130

**Boisvert, J. B., Manchuk, J. G., & Deutsch, C. V. (2009).** Kriging in the Presence of Locally Varying Anisotropy Using Non-Euclidean Distances. Mathematical Geosciences, 41(5), 585–601. https://doi.org/10.1007/s11004-009-9229-1

**Boogaart, K.G., & Albert, C. G. (2019).** Gaussian Processes for Data Fulfilling Linear Differential Equations. The 39th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, 5. https://doi.org/10.3390/proceedings2019033005

**Bostrom, A., Morss, R., Lazo, J. K., Demuth, J., & Lazrus, H. (2018).** Eyeing the storm: How residents of coastal Florida see hurricane forecasts and warnings. International Journal of Disaster Risk Reduction, 30, 105–119. https://doi.org/10.1016/j.ijdrr.2018.02.027

**Breiman, L. (2001).** Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

**Brent, R. P. (2002).** Algorithms for minimization without derivatives (Unabridged republication of the work publ. by Prentice-Hall ... 1973). Dover Publications.

**Buhmann, M. D. (2000).** Radial basis functions. Acta Numerica, 9, 1–38. https://doi.org/10.1017/S0962492900000015

**Chen, J., Chen, Z., Zhang, C., & Jeff Wu, C. F. (2022).** APIK: Active Physics-Informed Kriging Model with Partial Differential Equations. SIAM/ASA Journal on Uncertainty Quantification, 10(1), 481–506. https://doi.org/10.1137/20M1389285

**Cheong, L., Bleisch, S., Kealy, A., Tolhurst, K., Wilkening, T., & Duckham, M. (2016).** Evaluating the impact of visualization of wildfire hazard upon decision-making under uncertainty. International Journal of Geographical Information Science, 30(7), 1377–1404. https://doi.org/10.1080/13658816.2015.1131829

**Chilès, J.P., & Delfiner, O. (2013).** J.-P. Chilès, P. Delfiner: Geostatistics: Modeling Spatial Uncertainty: 2nd Edition. Wiley, 2012. Mathematical Geosciences, 45(3), 377–380. https://doi.org/10.1007/s11004-012-9429-y

**Cicchella, D., De Vivo, B., Lima, A., Albanese, S., & Fedele, L. (2008).** Urban geochemical mapping in the Campania region (Italy). Geochemistry: Exploration, Environment, Analysis, 8(1), 19–29. https://doi.org/10.1144/1467-7873/07-147





Funded by the European Union

**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). Environmental Earth Sciences, 75(19), 1325. https://doi.org/10.1007/s12665-016-6093-4

**Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., & Zucker, S. W. (2005).** Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proceedings of the National Academy of Sciences, 102(21), 7426–7431. https://doi.org/10.1073/pnas.0500334102

**Cook, D., & Laa, U. (2024).** Interactively exploring high-dimensional data and models in R. https://dicook.github.io/mulgar_book/

**Cook, D., Laa, U., & Valencia, G. (2018).** Dynamical projections for the visualization of PDFSense data. The European Physical Journal C, 78(9), 742. https://doi.org/10.1140/epjc/s10052-018-6205-2

**Dahlberg, E. C. (1975).** Relative effectiveness of geologists and computers in mapping potential hydrocarbon exploration targets. Journal of the International Association for Mathematical Geology, 7(5–6), 373–394. https://doi.org/10.1007/BF02080496

**De Kemp, E. A. (2021).** Spatial agents for geological surface modelling. Geoscientific Model Development, 14(11), 6661–6680. https://doi.org/10.5194/gmd-14-6661-2021

**Demetriades, A., Johnson, C., & Birke, M. (2018).** Urban Geochemical Mapping: The EuroGeoSurveys Geochemistry Expert Group's URGE project (Journal of Geochemical Exploration).

**Dempster, A. P. (2008).** The Dempster–Shafer calculus for statisticians. International Journal of Approximate Reasoning, 48(2), 365–377. https://doi.org/10.1016/j.ijar.2007.03.004

**Deutsch, C. V., & Journel, A. G. (Eds.). (1998).** GSLIB: Geostatistical software library and user's guide. Buch (2. ed). Oxford Univ. Press.

**Dhariwal, P., & Nichol, A. (2021).** Diffusion Models Beat GANs on Image Synthesis (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2105.05233

**Dubois, D., & Guyonnet, D. (2011).** Risk-informed decision-making in the presence of epistemic uncertainty. International Journal of General Systems, 40(2), 145–167. https://doi.org/10.1080/03081079.2010.506179

**Dubois, D., & Prade, H. (1985).** Théorie des Possibilités. Applications à la Représentation des Connaissances en Informatique (Masson).





Funded by the European Union

**Dubois, D., & Prade, H. (1988).** Possibility Theory: An Approach to Computerized Processing of Uncertainty. Springer US. https://doi.org/10.1007/978-1-4684-5287-7

**Dubois, G., & Galmarini, S. (2005).** Introduction to the Spatial Interpolation Comparison (SIC) 2004 Exercise and Presentation of the Datasets. Applied GIS, 1(2). https://doi.org/10.2104/ag050009

**Ducommun, C., Duvigneau,C., & Vidal-Beaudet, L. (2023).** Cartographie des sols urbains: Éléments de méthode. Revue Etude et Gestion des Sols, 30, 127–144.

**Dürr, H. H., Meybeck, M., & Dürr, S. H. (2005).** Lithologic composition of the Earth's continental surfaces derived from a new digital map emphasizing riverine material transfer. Global Biogeochemical Cycles, 19(4), 2005GB002515. https://doi.org/10.1029/2005GB002515

**Dutta, S., Ganguli, R., & Samanta, B. (2005).** Investigation of two Neural Network Methods in an Automatic Mapping Exercise. Applied GIS, 1(2), 19.

**Einstein, A. (1905).** Über die von der molecular-kinetischen Theorie der Wärme geforderte Bewegung von in rhuenden Flüssigkeiten suspendierten Teilchen. Ann. Phys., 17, 560.

**Fang, K.K.B. (2005).** Multi-Dimension and Real-Time Interpolation (Dirac-Monte Carlo Method) (p. 16) [Un-published]. FANG, INC.,. http://www.fanginc.com/rdic/sic1.pdf

**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. Environment International, 185, 108544. https://doi.org/10.1016/j.envint.2024.108544

**Fisher, R. A. (1992).** Statistical Methods for Research Workers. In S. Kotz & N. L. Johnson (Eds.), Breakthroughs in Statistics (pp. 66–70). Springer New York. https://doi.org/10.1007/978-1-4612-4380-9_6

**Fournier, B., & Furrer, R. (2016).** Automatic Mapping in the Presence of Substitutive Errors: A Robust Kriging Approach. 614184 Bytes. https://doi.org/10.4225/03/58007202615D5

**Friedman, J. H. (2001).** Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5). https://doi.org/10.1214/aos/1013203451

**Friedman, J., Hastie, T., & Tibshirani, R. (2010).** Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1). https://doi.org/10.18637/jss.v033.i01





Funded by the European Union

**Gibin, M., Longley, P., & Atkinson, P. (2007).** Kernel Density Estimation and Percent Volume Contours in General Practice Catchment Area Analysis in Urban Areas. GIScience Research UK Conference (GISRUK), Maynooth - Ireland. https://api.semanticscholar.org/CorpusID:70413057

**Gnann, S. J., Allmendinger, M. C., Haslauer, C. P., & Bárdossy, A. (2018).** Improving copula-based spatial interpolation with secondary data. Spatial Statistics, 28, 105–127. https://doi.org/10.1016/j.spasta.2018.07.001

**Godan. F., Zeydina. O., Richet. Y., & Beauzamy, B. (2015).** Reactor Safety and Incomplete Information: Comparison of Extrapolation Methods for the Extension of Computational Codes. Paper 15377, 5.

**Grunsky, E.C., & de Caritat, P. (2017).** Advances in the use of geochemical data for mineral exploration. Proceedings of Exploration 17, 17, 451–456.

**Guyonnet, D., Ménard, Y., Baudrit, C. & Dubois, D. (2005).** HyRisk—Traitement Hybride des Incertitudes en Evaluation des Risques (public No. RP 53714; p. 46). BRGM.

**Hair, J. F. (Ed.). (2010).** Multivariate data analysis (7. ed). Pearson Prentice Hall.

**Haji, S. O., & Yousif, R. Z. (2019).** A Novel Neutrosophic Method for Automatic Seed Point Selection in Thyroid Nodule Images. BioMed Research International, 2019, 1–14. https://doi.org/10.1155/2019/7632308

**Hannart, A., & Naveau, P. (2018).** Probabilities of Causation of Climate Changes. Journal of Climate, 31(14), 5507–5524. https://doi.org/10.1175/JCLI-D-17-0304.1

**Haslauer, C. P., Heißerer, T., & Bárdossy, A. (2016).** Including land use information for the spatial estimation of groundwater quality parameters – 2. Interpolation methods, results, and comparison. Journal of Hydrology, 535, 699–709. https://doi.org/10.1016/j.jhydrol.2016.01.054

**Hawkins, D. M., Basak, S. C., & Mills, D. (2003).** Assessing Model Fit by Cross-Validation. Journal of Chemical Information and Computer Sciences, 43(2), 579–586. https://doi.org/10.1021/ci025626i

**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., … Crouch, J. (2021).** African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. Scientific Reports, 11(1), 6130. https://doi.org/10.1038/s41598-021-85639-y

**Hensman, J., Fusi, N., & Lawrence, N. D. (2013).** Gaussian Processes for Big Data (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1309.6835





Funded by the European Union

**Heuvelink, G. B. M., Kros, J., Reinds, G. J., & De Vries, W. (2016).** Geostatistical prediction and simulation of European soil property maps. Geoderma Regional, 7(2), 201–215. https://doi.org/10.1016/j.geodrs.2016.04.002
Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models (No. arXiv:2006.11239). arXiv. https://doi.org/10.48550/arXiv.2006.11239

**Hofierka, J. (2005).** Interpolation of Radioactivity Data Using Regularized Spline with Tension. Applied GIS, 1(2). https://doi.org/10.2104/ag050016

**Hong, S. (2010).** Multivariate Analysis of Diverse Data for improved Geostatistical Reservoir Modeling (Department of Civil Engineering, p. 203) [Doctoral Thesis]. University of Alberta. https://www.ccgalberta.com/ccgresources/theses/2010-s_hong-phd_thesis.pdf

**Huang, C. (2002).** Information diffusion technique and the small sample problem. International Journal of Information Technology & Decision Making, 01(02), 229–249. https://doi.org/10.1142/S0219622002000142

**Huang, C., & Moraga, C. (2005).** Extracting fuzzy if–then rules by using the information matrix technique. Journal of Computer and System Sciences, 70(1), 26–52. https://doi.org/10.1016/j.jcss.2004.05.001

**Huang, C., & Shi, Y. (2002).** Towards Efficient Fuzzy Information Processing (Vol. 99). Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-1785-0

**Huang, H., Liang, Z., Li, B., & Wang, D. (2019).** A new spatial precipitation interpolation method based on the information diffusion principle. Stochastic Environmental Research and Risk Assessment, 33(3), 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. Geographical Research, 17, 71–77.

**Hullman, J., Resnick, P., & Adar, E. (2015).** Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. PLOS ONE, 10(11), e0142444. https://doi.org/10.1371/journal.pone.0142444

**Ingamells, C. O., & Pitard, F. F. (1986).** Applied geochemical analysis. J. Wiley and sons.
Ingram, B., Csat, L., & Evans, D. (2016). Fast Spatial Interpolation using Sparse Gaussian Processes. 722803 Bytes. https://doi.org/10.4225/03/580070B4D62D9

**Jenny, H. (1994).** Factors of soil formation: A system of quantitative pedology. Dover.

**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





Funded by the European Union

processing techniques on European agricultural soil data. Journal of Geochemical Exploration, 186, 143–157. https://doi.org/10.1016/j.gexplo.2017.11.011

**Joslyn, S., & Savelli, S. (2021).** Visualizing Uncertainty for Non-Expert End Users: The Challenge of the Deterministic Construal Error. Frontiers in Computer Science, 2, 590232. https://doi.org/10.3389/fcomp.2020.590232

**Journel, A. G. (2002).** Combining Knowledge from Diverse Sources: An Alternative to Traditional Data Independence Hypotheses. Mathematical Geology, 34(5), 573–596. https://doi.org/10.1023/A:1016047012594

**Kale, A., Nguyen, F., Kay, M., & Hullman, J. (2019).** Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. IEEE Transactions on Visualization and Computer Graphics, 25(1), 892–902. https://doi.org/10.1109/TVCG.2018.2864909

**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. International Journal of Mining, Reclamation and Environment, 34(1), 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. Applied Mathematical Modelling, 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 of Conferences, 170, 08002. https://doi.org/10.1051/epjconf/201817008002

**Kinkeldey, C., MacEachren, A. M., Riveiro, M., & Schiewe, J. (2017).** Evaluating the effect of visually represented geodata uncertainty on decision-making: Systematic review, lessons learned, and recommendations. Cartography and Geographic Information Science, 44(1), 1–21. https://doi.org/10.1080/15230406.2015.1089792

**Koenderink, J. J. (1984).** The structure of images. Biological Cybernetics, 50(5), 363–370. https://doi.org/10.1007/BF00336961

**Kolmogoroff, A. (1931).** Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. Mathematische Annalen, 104(1), 415–458. https://doi.org/10.1007/BF01457949

**Kolmogorov, A. N., & Morrison, N. (1956a).** Foundation of the theory of probability. Chelsea.

**Kolmogorov, A. N., & Morrison, N. (1956b).** Foundation of the theory of probability. Chelsea.





Funded by the European Union

**Koundal, D., Gupta, S., & Singh, S. (2018).** Neutrosophic Based Nakagami Total Variation Method for Speckle Suppression in Thyroid Ultrasound Images. IRBM, 39(1), 43–53. https://doi.org/10.1016/j.irbm.2017.11.003

**Krishnan, S. (2008).** The Tau Model for Data Redundancy and Information Combination in Earth Sciences: Theory and Application. Mathematical Geosciences, 40(6), 705–727. https://doi.org/10.1007/s11004-008-9165-5

**Krishnan, S., Boucher, A., & Journel, A. G. (2005).** Evaluating Information Redundancy Through the Tau Model. In O. Leuangthong & C. V. Deutsch (Eds.), Geostatistics Banff 2004 (Vol. 14, pp. 1037–1046). Springer Netherlands. https://doi.org/10.1007/978-1-4020-3610-1_108

**Kruskal, J. B. (1964).** Nonmetric Multidimensional Scaling: A Numerical Method. Psychometrika, 29(2), 115–129. https://doi.org/10.1007/BF02289694

**Kuijper, A., & Florack, L., M., J. (2001).** The application of catastrophe theory to image analysis (Technical Report No. UU–CS–2001–23; , Dept. of Computer Science). Utrecht University. ftp://ftp. cs.uu.nl/pub/RUU/CS/techreps/CS- 2001/2001- 23.pdf

**Kuijper, A., & Florack, L., M., J. (2002).** The relevance of non-generic events in scale space models. 7th European Conf. Computer Vision, Copenhagen, Denmark.
Lajaunie, C. (1996). Documentation of the mixed support kriging programs (p. 21). Ecole nationale des mines de Paris. https://www.geovariances.com/wp-content/uploads/2017/01/mixed-support-kriging-programs-ensmp-1996.pdf

**Lemons, D. S., & Gythiel, A. (1997).** Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146 , 530–533 (1908)]. American Journal of Physics, 65(11), 1079–1081. https://doi.org/10.1119/1.18725

**Levine, N. (2010).** Crimestat iii: A spatial statistics program for the analysis of crime incident locations (version 3.3) (Ned Levine & Associates/Washington, p. 874). National Institute of Justice.

**Li, J. (2016).** Probability density evolution method: Background, significance and recent developments. Probabilistic Engineering Mechanics, 44, 111–117. https://doi.org/10.1016/j.probengmech.2015.09.013

**Li, J., & Chen, J. bing. (2009).** Stochastic dynamics of structures. J. Wiley & Sons.

**Li, J., & Chen, J. (2006).** The probability density evolution method for dynamic response analysis of non-linear stochastic structures. International Journal for Numerical Methods in Engineering, 65(6), 882–903. https://doi.org/10.1002/nme.1479





Funded by the European Union

**Li, J., & Chen, J. (2008).** The principle of preservation of probability and the generalized density evolution equation. Structural Safety, 30(1), 65–77. https://doi.org/10.1016/j.strusafe.2006.08.001

**Li, J., & Chen, J. B. (2004).** Probability density evolution method for dynamic response analysis of structures with uncertain parameters. Computational Mechanics, 34(5), 400–409. https://doi.org/10.1007/s00466-004-0583-8

**Lifshitz, L. M., & Pizer, S. M. (1990).** A multiresolution hierarchical approach to image segmentation based on intensity extrema. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(6), 529–540. https://doi.org/10.1109/34.56189

**Lin, Y.-P., Chu, H.-J., Wu, C.-F., Chang, T.-K., & Chen, C.-Y. (2010).** Hotspot Analysis of Spatial Environmental Pollutants Using Kernel Density Estimation and Geostatistical Techniques. International Journal of Environmental Research and Public Health, 8(1), 75–88. https://doi.org/10.3390/ijerph8010075

**Lindeberg, T. (1997).** Scale-space theory in computer vision (3. printing). Kluwer Acad. Publ.

**Lipp, A., & Vermeesch, P. (2023).** Short communication: The Wasserstein distance as a dissimilarity metric for comparing detrital age spectra and other geological distributions. Geochronology, 5(1), 263–270. https://doi.org/10.5194/gchron-5-263-2023

**Liu, L., Boone, A. P., Ruginski, I. T., Padilla, L., Hegarty, M., Creem-Regehr, S. H., Thompson, W. B., Yuksel, C., & House, D. H. (2017).** Uncertainty Visualization by Representative Sampling from Prediction Ensembles. IEEE Transactions on Visualization and Computer Graphics, 23(9), 2165–2178. https://doi.org/10.1109/TVCG.2016.2607204

**Lophaven, S., Nielsen, H. B., & Sondergaard, J. (2005).** Automatic Mapping of Monitoring Data. Applied GIS, 1(2). https://doi.org/10.2104/ag050013

**Loquin, K., & Dubois, D. (2010).** Kriging and Epistemic Uncertainty: A Critical Discussion. In R. Jeansoulin, O. Papini, H. Prade, & S. Schockaert (Eds.), Methods for Handling Imperfect Spatial Information (Vol. 256, pp. 269–305). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-14755-5_11

**Maaten, L. van der, & Hinton, G. (2008).** Visualizing Data using t-SNE. Journal of Machine Learning Research, 9(86), 2579–2605.

**Mako, Z. (2005).** Approximation with diffusion-neural-network. Proceedings, 589–600.
Malone, B. P., Minasny, B., & McBratney, A. B. (2017). Using R for Digital Soil Mapping. Springer International Publishing. https://doi.org/10.1007/978-3-319-44327-0





Funded by the European Union

**Marchant, B. P., Saby, N. P. A., & Arrouays, D. (2017).** A survey of topsoil arsenic and mercury concentrations across France. Chemosphere, 181, 635–644. https://doi.org/10.1016/j.chemosphere.2017.04.106

**Mastrandrea, M. D., Mach, K. J., Plattner, G.-K., Edenhofer, O., Stocker, T. F., Field, C. B., Ebi, K. L., & Matschoss, P. R. (2011).** The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups. Climatic Change, 108(4), 675–691. https://doi.org/10.1007/s10584-011-0178-6

**McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003).** On digital soil mapping. Geoderma, 117(1–2), 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 secteur minier de Salsigne (Aude) (public No. RP-7067-FR; p. 113). BRGM.

**Moon, K. R., Van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Van Den Elzen, A., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., & Krishnaswamy, S. (2017).** Visualizing Structure and Transitions for Biological Data Exploration. https://doi.org/10.1101/120378

**Mulder, K. J., Lickiss, M., Black, A., Charlton-Perez, A. J., McCloy, R., & Young, J. S. (2020).** Designing environmental uncertainty information for experts and non-experts: Does data presentation affect users' decisions and interpretations? Meteorological Applications, 27(1), e1821. https://doi.org/10.1002/met.1821

**Négrel, P., 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. Applied Geochemistry, 111, 104425. https://doi.org/10.1016/j.apgeochem.2019.104425

**Négrel, P., Sadeghi, M., Ladenberger, A., Reimann, C., & Birke, M. (2015a).** Geochemical fingerprinting and source discrimination of agricultural soils at continental scale. Chemical Geology, 396, 1–15. https://doi.org/10.1016/j.chemgeo.2014.12.004

**Négrel, P., Sadeghi, M., Ladenberger, A., Reimann, C., & Birke, M. (2015b).** Geochemical fingerprinting and source discrimination of agricultural soils at continental scale. Chemical Geology, 396, 1–15. https://doi.org/10.1016/j.chemgeo.2014.12.004

**Niles-Weed, J., & Rigollet, P. (2019).** Estimation of Wasserstein distances in the Spiked Transport Model (No. arXiv:1909.07513). arXiv. https://doi.org/10.48550/arXiv.1909.07513





Funded by the European Union

**O'Sullivan, D., & Wong, D. W. S. (2007).** A Surface-Based Approach to Measuring Spatial Segregation. Geographical Analysis, 39(2), 147–168. https://doi.org/10.1111/j.1538-4632.2007.00699.x

**Özyurt, F., Sert, E., Avci, E., & Dogantekin, E. (2019).** Brain tumor detection based on Convolutional Neural Network with neutrosophic expert maximum fuzzy sure entropy. Measurement, 147, 106830. https://doi.org/10.1016/j.measurement.2019.07.058

**Padilla, L. M., Ruginski, I. T., & Creem-Regehr, S. H. (2017).** Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. Cognitive Research: Principles and Implications, 2(1), 40. https://doi.org/10.1186/s41235-017-0076-1

**Palaseanu-Lovejoy, M. (2016).** Bayesian Automating Fitting Functions for Spatial Predictions. 1054203 Bytes. https://doi.org/10.4225/03/580070F90569D

**Panagos, P., Meusburger, K., Ballabio, C., Borrelli, P., & Alewell, C. (2014).** Soil erodibility in Europe: A high-resolution dataset based on LUCAS. Science of The Total Environment, 479–480, 189–200. https://doi.org/10.1016/j.scitotenv.2014.02.010

**Pannecooke, L. (2021).** Combinaison de la géostatistique et des simulations à base physique – application à la caractérisation de panaches de contaminants (Doctoral thesis No. tel-03135798 , version 1; Ingénierie de l'environnement., p. 187). Mines Paritch. https://pastel.hal.science/tel-03135798v1/document

**Parzen, E. (1962).** On Estimation of a Probability Density Function and Mode. The Annals of Mathematical Statistics, 33(3), 1065–1076. https://doi.org/10.1214/aoms/1177704472

**Pearl, J. (2000).** Causality: Models, reasoning, and inference. Cambridge university press.

**Pebesma, E. J. (2016).** Mapping Radioactivity from Monitoring Data: Automating the Classical Geostatistical Approach. 1096714 Bytes. https://doi.org/10.4225/03/58006F1B9331A

**Pelz, M.-T., Schartau, M., Somes, C. J., Lampe, V., & Slawig, T. (2023).** A diffusion-based kernel density estimator (diffKDE, version 1) with optimal bandwidth approximation for the analysis of data in geoscience and ecological research. https://doi.org/10.5194/gmd-2023-17

**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





Funded by the European Union

**Pitard, F. F. (Ed.). (2019).** Theory of sampling and sampling practice (Third edition). CRC Press, Taylor & Francis Group.

**Planck, M. (1917).** Uber einen Satz der statistichen Dynamik und eine Erweiterung inder Quantumtheori. Sitzungberichte Preuss, 324–341.

**Polyakova, E. I., & Journel, A. G. (2007).** The Nu Expression for Probabilistic Data Integration. Mathematical Geology, 39(8), 715–733. https://doi.org/10.1007/s11004-007-9117-5

**Pozdnoukhov, A. (2016).** Support Vector Regression for Automated Robust Spatial Mapping of Natural Radioactivity. 299626 Bytes. https://doi.org/10.4225/03/5800648078133

**Pyrcz, M. J., & Deutsch, C. V. (2014).** Geostatistical reservoir modeling (2nd ed). Oxford university press.

**Quinlan, J.R. (1993).** Combining Instance-Based and Model-Based Learning. Elsevier. https://doi.org/10.1016/C2009-0-27798-1

**R Core Team. (2022).** R: A Language and Environment for Statistical Computing [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

**Ranftl, S. (2022).** A connection between probability, physics and neural networks. https://doi.org/10.48550/ARXIV.2209.12737

**Reimann, C., Birke, M., & Demetriades, A. (2014).** Chemistry of Europe's agricultural soils: Part A : Methodology and interpretation of the GEMAS data set. Bundesanstalt für Geowissenschaftem und Rohstoffe.

**Reimann, C., Birke, M., Filzmoser, P., & O'Connor, P. (2014).** Chemistry of Europe's Agricultural Soils, Part B : General Background Information and Further Analysis of the GEMAS Data Set. Bundesanstalt für Geowissenschaften und Rohstoffe.

**Reimann, C., & Filzmoser, P. (2000).** Normal and lognormal data distribution in geochemistry: Death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environmental Geology, 39(9), 1001–1014. https://doi.org/10.1007/s002549900081

**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. Environmental Pollution, 182, 15–27. https://doi.org/10.1016/j.envpol.2013.06.041





Funded by the European Union

**Rigol-Sanchez, J. P. (2016).** Spatial Interpolation of Natural Radiation Levels with Prior Information using Back-propagation Artificial Neural Networks. 575071 Bytes. https://doi.org/10.4225/03/58006CBDC9DCC

**Roux, W., Yi, G., & Gandikota, I. (2020).** A spatial kernel approach for topology optimization. Computer Methods in Applied Mechanics and Engineering, 361, 112794. https://doi.org/10.1016/j.cma.2019.112794

**Salminen, R., De Vos, W., & Tarvainen, T. (2005).** Geochemical atlas of Europe [Map]. Geological survey of Finland.

**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. Journal of Geochemical Exploration, 240, 107031. https://doi.org/10.1016/j.gexplo.2022.107031

**Saveliev, A. A., Romanov, A. V., & Mukharamova, S. S. (2005).** Automated Mapping using Multilevel B-Splines. Applied GIS, 1(2). https://doi.org/10.2104/ag050017

**Savelieva, E. (2005).** Using Ordinary Kriging to Model Radioactive Contamination Data. Applied GIS, 1(2). https://doi.org/10.2104/ag050010

**Scott, D. W. (1992).** Multivariate density estimation: Theory, practice, and visualization. Wiley. https://doi.org/10.1002/9780470316849

**Sego, L.H. & Wilson, J.E. (2007).** Accounting for false negatives in hotspot detection (No. PNNL-16812). Pacific Northwest National Laboratory. https://digital.library.unt.edu/ark:/67531/metadc897079/m2/1/high_res_d/946674.pdf

**Sert, E. & Avci, D. (2019).** Brain tumor segmentation using neutrosophic expert maximum fuzzy-sure entropy and other approaches. Biomedical Signal Processing and Control, 47, 276–287. https://doi.org/10.1016/j.bspc.2018.08.025

**Shi, S., Zhou, P. & Lü, Z. (2021).** A density-based topology optimization method using radial basis function and its design variable reduction. Structural and Multidisciplinary Optimization, 64(4), 2149–2163. https://doi.org/10.1007/s00158-021-02972-6

**Silva, V. & Tenenbaum, J. (2002).** Global Versus Local Methods in Nonlinear Dimensionality Reduction. Neural Information Processing Systems. http://papers.nips.cc/paper/2141-global-versus-local-methods-in-nonlinear-dimensionality-reduction.pdf

**Silverman, B. W. (2018).** Density Estimation for Statistics and Data Analysis. Routledge.
Sinclair, A. J., & Blackwell, G. H. (2002). Applied Mineral Inventory Estimation (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511545993





Funded by the European Union

**Singer, D.A. (1972).** ELIPGRID: A Fortran IV Program for Calculating the Probability of Success in Locating Elliptical Targets with Square, Rectangular and Hexagonal Grids (Version Programs 4:1–16) [Computer software].

**Skøien, J. O., Blöschl, G., Laaha, G., Pebesma, E., Parajka, J., & Viglione, A. (2014).** rtop: An R package for interpolation of data with a variable spatial support, with an example from river networks. Computers & Geosciences, 67, 180–190. https://doi.org/10.1016/j.cageo.2014.02.009

**Smarandache, F. (2006).** Neutrosophic set—A generalization of the intuitionistic fuzzy set. 2006 IEEE International Conference on Granular Computing, 38–42. https://doi.org/10.1109/GRC.2006.1635754

**Sohl-Dickstein, J., Maheswaranathan,N., & Ganguli, S. (2015).** Deep Unsupervised Learning using Nonequilibrium Thermodynamics. Proceedings of the 32nd International Conference on Machine Learning, 2256--2265. https://proceedings.mlr.press/v37/sohl-dickstein15.html

**Stratonovič, R. L. (1981).** Topics in the theory of Random noise. 1: General theory of random processes, nonlinear transformations of signals and noise (Rev. Engl. ed., 3. print). Gordon & Breach.

**Székely, G. J., & Rizzo, M. L. (2013).** Energy statistics: A class of statistics based on distances. Journal of Statistical Planning and Inference, 143(8), 1249–1272. https://doi.org/10.1016/j.jspi.2013.03.018

**Tak, S., Toet, A., & Van Erp, J. (2014).** The Perception of Visual UncertaintyRepresentation by Non-Experts. IEEE Transactions on Visualization and Computer Graphics, 20(6), 935–943. https://doi.org/10.1109/TVCG.2013.247

**Tarvainen, T., Albanese, S., Birke, M., Poňavič, M., & Reimann, C. (2013).** Arsenic in agricultural and grazing land soils of Europe. Applied Geochemistry, 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(2). https://doi.org/10.2104/ag050019

**Torgerson, W. S. (1952).** Multidimensional Scaling: I. Theory and Method. Psychometrika, 17(4), 401–419. https://doi.org/10.1007/BF02288916

**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. Science of The Total Environment, 565, 1054–1062. https://doi.org/10.1016/j.scitotenv.2016.05.115





Funded by the European Union

**Tóth, G., Jones, A., & Montanarella, L. (2013).** LUCAS topsoil survey: Methodology, data and results. Publications Office. https://data.europa.eu/doi/10.2788/97922

**Van Eynde, E., Fendrich, A. N., Ballabio, C., & Panagos, P. (2023).** Spatial assessment of topsoil zinc concentrations in Europe. Science of The Total Environment, 892, 164512. https://doi.org/10.1016/j.scitotenv.2023.164512

**Verstraete, J., & Radziszewska, W. (2021).** Rulebase construction using variables with data-dependent domains. Information Sciences, 570, 52–69. https://doi.org/10.1016/j.ins.2021.04.037

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

**Wady, S. H., Yousif, R. Z., & Hasan, H. R. (2020).** A Novel Intelligent System for Brain Tumor Diagnosis Based on a Composite Neutrosophic-Slantlet Transform Domain for Statistical Texture Feature Extraction. BioMed Research International, 2020(1), 8125392. https://doi.org/10.1155/2020/8125392

**Wand, M. P., & Jones, M. C. (1995).** Kernel smoothing (1. ed). Chapman & Hall.

**Wasserstein, L. (1969**). Markov processes on countable space products describing large systems of automata. Probl. Pered. Inform., 5(3), 64–72.

**Webster, K. N., Department of Mathematics, Imperial College London, 180 Queens Gate, London SW7 2AZ, UK, & FeedForward Ltd., London, UK. (2023).** Low-rank kernel approximation of Lyapunov functions using neural networks. Journal of Computational Dynamics, 10(1), 152–174. https://doi.org/10.3934/jcd.2022026

**Webster, R., & Oliver, M. A. (2007).** Geostatistics for Environmental Scientists (1st ed.). Wiley. https://doi.org/10.1002/9780470517277

**Whittle, P. (1958).** On the Smoothing of Probability Density Functions. Journal of the Royal Statistical Society Series B: Statistical Methodology, 20(2), 334–343. https://doi.org/10.1111/j.2517-6161.1958.tb00298.x

**Wiener, N. (1923).** Differential-Space. Journal of Mathematics and Physics, 2(1–4), 131–174. https://doi.org/10.1002/sapm192321131

**Williams, D. (2001).** Weighing the odds: A course in probability and statistics. Cambridge university press.

**Witt, J. K., Clegg, B. A., Blalock, L. D., & Warden, A. C. (2021).** The Impact of Familiarity on Visualizations of Spatial Uncertainty. Proceedings of the Human Factors and





Funded by the European Union

Ergonomics Society Annual Meeting, 65(1), 596–600. https://doi.org/10.1177/1071181321651208

**Witt, J. K., Clegg, B. A., Wickens, C. D., Smith, C. A. P., Laitin, E. L., & Warden, A. C. (2020).** Dynamic Ensembles versus Cones of Uncertainty: Visualizations to Support Understanding of Uncertainty in Hurricane Forecasts. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64(1), 1644–1648. https://doi.org/10.1177/1071181320641399

**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 and Management, 39(2), 900–916. https://doi.org/10.1111/sum.12874

**Xie, Z., & Yan, J. (2008).** Kernel Density Estimation of traffic accidents in a network space. Computers, Environment and Urban Systems, 32(5), 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. Science of The Total Environment, 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 Y. Shi, G. D. Van Albada, J. Dongarra, & P. M. A. Sloot (Eds.), Computational Science – ICCS 2007 (Vol. 4489, pp. 1069–1076). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72588-6_171

**Zadeh, L. A. (1999).** Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems, 100, 9–34. https://doi.org/10.1016/S0165-0114(99)80004-9

**Zeydina, O., & Beauzamy, B. (2013).** Probabilistic information transfer. Société de calcul mathématique.

**Zhou C., Wan Q., Huang S., & Chen D. (2000).** A GIS-based Approach to Flood Risk Zonation. Acta Geographica Sinica, 55, 15.





Funded by the European Union

# Appendix 2 Algorithm scripts

```
#########################################################################
###############
# ================================
# Install R iisdia dependencies
# ================================


#Prerequisite is an R > v3.1 and Rtools installed

# I povide here the ipak function which install and load multiple R packages.

ipak <- function(pkg){
   new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
   if (length(new.pkg))
      install.packages(new.pkg, dependencies = TRUE)
   sapply(pkg, require, character.only = TRUE)
}

# You can then install on this computer iidia package dependancies  mainly for anomaly analysis

packages <-  ("raster","RANN", "pbapply", "data.table", "geocmeans", "terra" ,"robustHD", "sf",
"RDP", "FuzzyMCDM", "RankAggreg" , "RColorBrewer")
ipak(packages)


#if we have to re-compile from the source, you must also install developper R packages

packages_dvlp <-  ("devtools", "roxygen2", "usethis", "Rcpp","RcppArmadillo")
ipak(packages_dvlp )

#once these package have been set you can cut internet connection all will function locally

#interpolation and anomaly detection ISLANDR package#
# a specific package iisdia with the useful demo functions has been compiled for this test to a zip
format.

# ================================
# the GEMAS Demo
# ================================

library("iisdia") #the WP1.2 algoritm

#interpolation under R ecosystem ####

library("raster")
library("RANN")
```





Funded by the European Union

```
library(pbapply)
library(data.table)


#colors for maps#######################
library(RColorBrewer)

# Local I moran index raster and SGCFM clustering #############
library("geocmeans")
library(terra)

#######  Singularity index #################
library(robustHD)

###### automatic fractal methods#######
library(sf)
library(RDP)

###### Meta ranking of anomaly#################
library("FuzzyMCDM")
library ("RankAggreg")


#################################################################################
#####
# ================================
# Apply isidia treatment on GEMAS Data
# ================================

### load packages

library("raster")
library("RANN")
library(pbapply)
library(data.table)

#library("iisdia")


###load datas

data  <-  read.table("D:/Documents/belbeze/OneDrive -
BRGM/GEMAS/GEMAS_Ap_test.csv",header=TRUE,sep = ";",na.strings="NA")

#Notes :
#North and south europe are geo diffent so data are separated in TYPE2 field  "ApS" and "ApN"
#Countries fields in COUNTRY field.

#first reproject coodinates
```





Funded by the European Union

```
df$coords <- data.frame(longitude = data2$GPS_LONG, latitude = data2$GPS_LAT)

names(df) = c("longitude", "latitude")

df$coords_sf <- st_as_sf(data.frame( data2[, c("GPS_LONG","GPS_LAT")]), coords = c("GPS_LONG",
"GPS_LAT"), crs = 4326)

df2  <- st_transform(df , 3035)


df3 = st_coordinates(df2)

data2[, "XLAEA"] <-df3$X
data2[, "YLAEA"] <- df3$Y

data2 <- data2[ !is.na(data2$XLAEA) ,]

######then a north south field is necessary for european datas

df2 = st_as_sf(data.frame( data[, c("XLAEA","YLAEA")]), coords = c("XLAEA","YLAEA"), crs =3035)

suture = st_read( "D:/Documents/belbeze/OneDrive -
BRGM/GEMAS/SIG/Suture_nord_polygone_3035.shp" )

st_crs(suture) = st_crs(df2)
good_points <- st_filter(df2, suture)

who_is_north = data.frame( st_coordinates(good_points))

plot(good_points)

data2 = data
xy = data [, c("XLAEA","YLAEA")]
names(xy) = c("x","y")


data2[,"TYPE3"] = "ApS"

#aquicker  way
#data2[ (xy$x %in% who_is_north$X) & (xy$y %in% who_is_north$Y),"TYPE3" ] =  "ApN"

for (ii in 1:dim(who_is_north)[1] ){
                    data2[ xy$x == who_is_north[ii,]$X & xy$y == who_is_north[ii,]$Y, "TYPE3"] = "ApN"


}

#test if it is well done

sum( data2$TYPE3=="ApN")
sum( data2$TYPE3=="ApS")
```





Funded by
the European Union

```
#now xe fill a country fields for the dataset

test_pays = st_read("D:/Documents/belbeze/OneDrive - BRGM/GEMAS/SIG/Europe_merged.shp")
test_pays_3035 = st_transform(test_pays, crs =3035)

# which points fall inside the first polygon?
st_intersects(test_pays_3035, df2)[[1]]

data2[ ,"COUNTRY"]<-"UNK"

# which points fall inside the ii t polygon?
ii=1

for( ii in 1:dim(test_pays_3035)[1] ) {
        a = st_intersects(test_pays_3035, df2)[[ii]]
        if( length(a)> 0) data2[a ,"COUNTRY"]<-test_pays_3035$GID_0[ii]

}

# Now we can start trhe iisdia process

#extract coords

xy = data [, c("XLAEA","YLAEA")]

###load basic rasters for interpolation 1kmx1km

#interpolation grid and perimeter

rr <- raster("D:/Documents/belbeze/OneDrive - BRGM/GEMAS/peri_euro_3035.tif")

# a goof covarariate  1kmx1km

geole_disc <- raster("D:/Documents/belbeze/OneDrive - BRGM/GEMAS/Europe_geole_3035.tif")

# these no learning datas about lithologic code 11. so set it to 10.
 geole_disc[geole_disc[]==11 ] <- 10

plot(geole_disc)


#if not done, prepare north/south europe fiels and country fiels


#select analyte from gemas
### choose an analyte

analyte = "Cd"
LQ = 0.005 # pour Cd
```





Funded by the European Union

```
##lests start monitoring the procedure

t_start = Sys.time()

#####perform a quick neutral interpolation for anaomaly analysis


obs  <- data  [   , analyte  ]
xy =  data[   ,c("XLAEA", "YLAEA")]
names(xy) <- c("x","y")

#set an auto  discretization for obs for probabilistic  interpolation
# Of course a manual one could be better

range(obs)
hist(obs, breaks = "Freedman-Diaconis")
breaks_dec = hist(obs, breaks = "Freedman-Diaconis")$breaks

summary(obs)

#build the censoring indicator
# with only one LQ
#censored = ifelse(obs<=LQ,1,0)
#sum(censored)

# with multiple LQ
#with LQ1 < LQ2 < LQ3
#censored = rep(0, length(obs))
#censored[  obs== LQ1] <-1
#censored[  obs== LQ2] <-1
#censored[  obs== LQ2] <-1

params_tous  =  data[ ,c("XLAEA", "YLAEA" )]
names(params_tous ) = c("x","y" )

param_stack   = as.data.frame( geole_disc, xy = TRUE)[, c("x","y")]
hdis = 10000

#build if necessarry declustering weight

#w1 = Poids_des_clusters_spaciaux_2D_sy(A[,c("x", "y")] ,  hdis = 10000 )

#build if needed censored or uncertainty fields.

#we ask only the median  not quantile or proba tensor for this demo

t1 <- Sys.time()
test  = iisdia_interpolate_dvlp (  A = as.matrix(params_tous)   ,  obs= obs  ,  censored = censored,
breaks_dec = breaks_dec ,
  param_stack = as.matrix(param_stack)  , Mode_Voisinage = 1 ,kmax = as.integer(25) ,
```





Funded by the European Union

```
 Mode_Sortie = 0 ) #radius = radius
t2 <- Sys.time()
t2-t1
#Time difference of 5.179866 mins

#make a raster of ou results

#test set

Enterpolated_Neutral_means = mask( rasterFromXYZ(test [,c(1,2,3)] ),rr  )

plot(log10(Enterpolated_Neutral_means ))

#####perform a base interpolation for further publication
#Sample the covariate to add to data

data2

data2[  , "geole"] <-extract(geole_disc, xy)
data = data[!is.na( data$geole),]

obs  <- data  [   , analyte  ]
xy =  data[   ,c("XLAEA", "YLAEA")]

xy =param_tous[, c("x","y")]

names(xy) <- c("x","y")

censored = ifelse(obs<=LQ,1,0)


#Set the vector giving the breakpoints of contiguous intervals (ie sets), also called bins, that span
the range of the variable obs values (discretisation).
#The breaks are the lower limit of each bin
range(obs)
hist(obs, breaks = "Freedman-Diaconis")
breaks_dec = hist(obs, breaks = "Freedman-Diaconis")$breaks

#build parameter
params_tous  =  data[ ,c("XLAEA", "YLAEA", "geole")]

params_tous  =  data[ ,c("XLAEA", "YLAEA", "geole")]
names(params_tous ) = c("x","y","geole")
param_stack   = as.data.frame( geole_disc, xy = TRUE)
names(param_stack) = c("x","y","geole")
hdis = 10000

#we ask only the median and not quantile or proba tensor for this demo
t1 <- Sys.time()
test  = iisdia_interpolate_dvlp (  A = as.matrix(params_tous)   ,  obs= obs  ,censored=censored ,
breaks_dec = breaks_dec ,
```





Funded by the European Union

```
  param_stack = as.matrix(param_stack)  , Mode_Voisinage = 1 ,kmax = as.integer(25) ,
 Mode_Sortie = 0 ) #radius = radius
t2 <- Sys.time()
t2-t1
#Time difference of 1.671623 mins


Eventually_published_covar_means = mask( rasterFromXYZ(test [,c(1,2,3)] ),rr  )

plot(log10(Eventually_published_covar_means ))

#save the two interpolated tiff for further use in anomaly detection

path_grappe = "D:/Documents/belbeze/OneDrive - BRGM/GEMAS/"
rf <- writeRaster( Enterpolated_Neutral_means ,
filename=paste0(path_grappe, "GEMAS_NEUTRAL_", analyte,"_1kmx1km_3035.tif") ,
format="GTiff", overwrite=TRUE)

rf <- writeRaster( Eventually_published_covar_means  ,
filename=paste0(path_grappe, "GEMAS_COVAR_", analyte,"_1kmx1km_3035.tif") , format="GTiff",
overwrite=TRUE)

library(RColorBrewer)
cols <- brewer.pal(11, "Spectral")
plot(log10(Enterpolated_Neutral_means) , col= rev(cols), main = paste0("LUCAS EEPH ",
analyte,"Expectation") )

plot(log10(Eventually_published_covar_means) , col= rev(cols), main = paste0("LUCAS EEPH ",
analyte,"Expectation") )


#####perform a reimann quick threshold analysis (Q95, Q98, TIF) per north and south europe


x =  data2[,analyte]
data2 [ ,paste0( analyte ,"_REI") ] <- Apply_reimanns_Thresholds( x )


x = data [ data $TYPE2=="ApS",analyte]
data [ data $TYPE2=="ApS",paste0( analyte ,"_REI")] <- Apply_reimanns_Thresholds( x )

x = data [ data $TYPE2=="ApN",analyte]
data [ data $TYPE2=="ApN",paste0( analyte ,"_REI") ] <- Apply_reimanns_Thresholds( x )

#####perform a zero probability bands analysis per country

 #probability bands

for (pays in data $COUNTRY ) {
          x = data [ data$COUNTRY==pays ,analyte]
      data [ data $COUNTRY==pays ,paste0( analyte ,"_pb")] <- Apply_Probability_bands ( x )
```





Funded by
the European Union

```
}


#######  Lets prepare North south raster for methods that are sensitive to multimodality
##################

r_teneurs = raster(paste0(path_grappe, "GEMAS_NEUTRAL_", analyte,"_1kmx1km_3035.tif") )
#r_teneurs = raster(paste0(path_grappe, "GEMAS_COVAR_", analyte,"_1kmx1km_3035.tif") )

t1 <- Sys.time()
Val = cut_NS_an_european_raster (r_teneurs  ,
  Suture_polygon_path = "D:/Documents/belbeze/OneDrive -
BRGM/GEMAS/SIG/Suture_nord_polygone_3035.shp" )
t2 <- Sys.time()
t2-t1
#Time difference of 0.2072151 secs

North_europe_r = Val[[1]]
South_europe_r = Val[[2]]
plot (North_europe_r)
plot (South_europe_r)

then save

writeRaster(North_europe_r, paste0(path_grappe, "NGEMAS_NEUTRAL_",
analyte,"_1kmx1km_3035.tif"), overwrite=TRUE)
writeRaster(South_europe_r, paste0(path_grappe, "SGEMAS_NEUTRAL_",
analyte,"_1kmx1km_3035.tif"), overwrite=TRUE)


#######  Lets Build a Local I moran index raster #################

library("geocmeans")
library(terra)


r = rast(paste0(path_grappe, "NGEMAS_NEUTRAL_", analyte,"_1kmx1km_3035.tif") )
w <- matrix(1, nrow = 3, ncol = 3)
testN = calc_local_moran_raster(r , w)

r = rast(paste0(path_grappe, "SGEMAS_NEUTRAL_", analyte,"_1kmx1km_3035.tif") )
w <- matrix(1, nrow = 3, ncol = 3)
testS = calc_local_moran_raster(r , w)

# in terra it is a bit complex than overlay

z <- c(testN,testS)

mygreedyfun <- function(x ){
        if( is.na(x[[1]])& !is.na(x[[2]]) ) val <- x[[2]]
```





Funded by
the European Union

```
        if( is.na(x[[2]])& !is.na(x[[1]]) ) val <- x[[1]]
        if( is.na(x[[2]])& is.na(x[[1]]) ) val <- NA
return(val)
}

test = app(z, mygreedyfun)

#plot(log(test +1))

writeRaster(test, paste0(path_grappe, "GEMAS_IMORAN_", analyte,"_1kmx1km_3035.tif"),
overwrite=TRUE)

#fill data with i moran value

data[, paste0(analyte,"_moran" ) ]<-  extract(test, xy, ID=FALSE)


#######  Lets Build a Local clustering  index ##################

nclust = 26

r = rast(paste0(path_grappe, "GEMAS_NEUTRAL_", analyte,"_1kmx1km_3035.tif") )

data_gcm =list(  r  [[1]])
w1 <- matrix(1, nrow = 3, ncol = 3)
t1 <- Sys.time()

SGFCM_result <- SGFCMeans(data_gcm , k = nclust , m = 1.5, standardize = TRUE,
                  lag_method = "mean",
                  window = w1, alpha = 0.9, beta = 0.5, maxiter = 5000 ,
                  seed = 789, tol = 0.001, verbose = FALSE, init = "kpp")
 cluster_r <- SGFCM_result$rasters$Groups
t2 <- Sys.time()
t2-t1

writeRaster( cluster_r, paste0(path_grappe, "GEMAS_CLUST26_", analyte,"_1kmx1km_3035.tif"),
overwrite=TRUE)

#detach(package:terra, unload = TRUE)

#lets re-load to work on a ::raster base
library(raster)
library(sp)

r_teneurs = raster::raster(paste0(path_grappe, "GEMAS_NEUTRAL_",
analyte,"_1kmx1km_3035.tif") )

nclust = 26

cluster_r <- raster::raster( paste0(path_grappe, "GEMAS_CLUST26_",
analyte,"_1kmx1km_3035.tif"))
```





Funded by the European Union

```
## build interesting spatial cluster statistics
surfaces_t  = cluster_stat_sum( nclust=26 , cluster_map =  cluster_r   , conc_map = r_teneurs )

##what does it looks like ?
surfaces_t

 plot( log(surfaces_t$Med+1) ~ log(surfaces_t$s+1), log="xy", type="l")

#lets cut approax 1 x grid length, 2 x grid length, 3 x grid length according to theory of laplace.
too long to explain.

mes_cuts = c(30000, 24000, 10400)

#as.raster(r_teneurs)


 data[, paste0(analyte,"_clust")]  <- cluster_index_extract ( xy = as.matrix(xy) ,cluster_map =
cluster_r   ,
conc_map =    r_teneurs  , surfaces_t    , mes_cuts , lowborder = 0.2 )


##########################
# Estimating singularity index

library(raster)
library(robustHD)

r_teneurs = raster(paste0(path_grappe, "GEMAS_NEUTRAL_", analyte,"_1kmx1km_3035.tif") )

res(r_teneurs)

#[1] 10000 10000

#define a least five window for calculus

maille=10000 #m
wsize_km = c( 20, 40, 60,80,100,140,180,200, 300)
wsize_pix <- wsize_km * 1000 /maille
wsize_pix [( (wsize_pix %% 2) == 0) ] = wsize_pix [( (wsize_pix %% 2) == 0) ] + 1
wsize_pix

#calculate raster of Singular index
t1 <- Sys.time()
SinguIndexmp = Calculate_Singl (r=r_teneurs, wsize_pix)
t2 <- Sys.time()
t2-t1
#Time difference of 2.830892 mins
```





Funded by the European Union

```
writeRaster(SinguIndexmp, paste0(path_grappe, "GEMAS_SINGI_", analyte,"_1kmx1km_3035.tif"),
overwrite=TRUE)

plot(SinguIndexmp)

#sample this singular index

data[, paste0(analyte, "_singI" )  ]<-  extract(SinguIndexmp, xy)

##########################
# Estimating Fractal thresolds

library(sf)
library(RDP)


#preparating north south raster

#r_teneurs = raster(paste0(path_grappe, "GEMAS_NEUTRAL_", analyte,"_1kmx1km_3035.tif") )
r_teneurs = raster(paste0(path_grappe, "GEMAS_COVAR_", analyte,"_1kmx1km_3035.tif") )

t1 <- Sys.time()
Val = cut_NS_an_european_raster (r_teneurs  ,
  Suture_polygon_path = "D:/Documents/belbeze/OneDrive -
BRGM/GEMAS/SIG/Suture_nord_polygone_3035.shp" )
t2 <- Sys.time()
t2-t1
#Time difference of 0.2072151 secs

North_europe_r = Val[[1]]
South_europe_r = Val[[2]]
plot (North_europe_r)
plot (South_europe_r)


breaks_dec_loc  = hist(North_europe_r, breaks = "Freedman-Diaconis")$breaks

testCN = data.frame( analyse_CA( r= North_europe_r,  seuils = breaks_dec_loc   ) )


breaks_dec_loc  = hist(South_europe_r, breaks = "Freedman-Diaconis")$breaks

testCS = data.frame( analyse_CA( r= South_europe_r,  seuils = breaks_dec_loc    ) )

fractal_inflexionsN  <- estimate_fractal_inflexion( x = testCN[, c("cl_min", "npix_equal_above_cl")] ,
plot=TRUE)$ cl_min

fractal_inflexionsS  <- estimate_fractal_inflexion( x = testCS[, c("cl_min", "npix_equal_above_cl")] ,
plot=TRUE)$ cl_min

#then update data
```





Funded by
the European Union

```
rcl = build_rcl_from_Threshold_list(threshold_l = fractal_inflexionsN )

data[ data$TYPE2=="ApN",paste0(analyte, "_FRA") ] <- df_re_classify(data[
data$TYPE2=="ApN",analyte ] , rcl, dolowest = TRUE, lowborder = 0 , highborder = 4 ,  NAval =NA)

rcl = build_rcl_from_Threshold_list(threshold_l = fractal_inflexionsS )
data[ data$TYPE2=="ApS",paste0(analyte, "_FRA")] <- df_re_classify(data[
data$TYPE2=="ApS",analyte] , rcl, dolowest = TRUE, lowborder = 0 , highborder = 4  ,  NAval =NA)

names(data)

######## In ITA mode calculate the Nemerow index

#load large scale survey as gemas
# r_teneurs = raster(paste0(path_grappe, "GEMAS_COVAR_", analyte,"_1kmx1km_3035.tif") )

 #data[, paste0(analyte,"_G")] = raster::extract(r_teneurs , xy)
# data[, paste0(analyte,"_neme" ) ] <- 0
 #data[, paste0(analyte,"_neme" ) ] <- data[,  analyte  ] / data[, paste0(analyte,"_G" ) ]

########Add moss data and pm2.5 data

moss_r = raster(paste0("D:/Documents/belbeze/OneDrive -
BRGM/GEMAS/Moss_PM25_continuous/",tolower(analyte),
"_neutralincerteph221324_10kmx10km.tiff" ))

pm25_r = raster("D:/Documents/belbeze/OneDrive -
BRGM/GEMAS/Moss_PM25_continuous/PM2.5_NeutralIncertEPH221324.tif")

# add moss and dust to data

data[, paste0(analyte,"_M")] = raster::extract(moss_r , xy)
data[, "PM2.5_M"] = raster::extract(pm25_r , xy)

########Save the detection matrix results for the given analyte


#data  <-  read.table("D:/Documents/belbeze/OneDrive -
BRGM/GEMAS/GEMAS_Ap_test.csv",header=TRUE,sep = ";",na.strings="NA") #stringsAsFactors
=FALSE)

detection_matrix = data[, c("XLAEA","YLAEA", "ID" ,"COUNTRY","TYPE2", analyte, paste0(analyte,
c("_REI","_pb","_FRA", "_singl", "_moran", "_clust", "_M")), "PM2.5_M" ) ]

write.table (detection_matrix , paste0(path_grappe, analyte, "_detection_matrix.csv" ), sep=";")

###############test

#detection_matrix =  read.table( paste0(path_grappe, analyte, "_detection_matrix.csv"),  sep=";")
```





Funded by
the European Union

```
fuzzy_detection_matrix <- fuzzify_detection_matrix ( detection_matrix, mode="lucas")

# save for safety

write.table ( fuzzy_detection_matrix , paste0(path_grappe, analyte, "_fuzzy_detection_matrix.csv"
), sep=";")


##################################Build decision matrix

#The decision matrix (m x n) with the values of the m alternatives, for the n criteria.

#following criterias will be ranked

mes_citeres =  paste0(analyte, c("_REI_", "_pb_", "_FRA_",  "_singl_",  "_moran_",  "_clust_")) #toward
6 criteria ranking

#and build the matrix

decision_matrix = build_anomaly_decision_matrix( fuzzy_detection_matrix, mes_citeres =
mes_citeres)

# A vector of length n, containing the weights for the criteria. The sum of the weights has to be 1.
#ranks established on islandr tests
#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)

n_criteres = 6
mes_criteres_w = c(0.3,0.38,0.45,
0.005,0.01,0.03,
0.05,0.04,0.1,
```





Funded by the European Union

```
0.17,0.24,0.3,
0.02,0.03,0.1,
0.17,0.3,0.4)

#check that the sum pf ranking criteria is 1
#and there is 6 ranked criteria
sum(mes_criteres_w [seq(2, length(mes_criteres_w), 3)])
length(mes_criteres_w)/3

#on ita site
#n_criteres = 7
#mes_criteres_w = c(0.3,          0.25,     0.34,
#0.005, 0.01,     0.03,
#0.03,   0.04,     0.05,
#0.17,   0.23,     0.3,
#0.02,   0.03,     0.08,
#0.17,   0.25,     0.35,
#0.15,   0.2,      0.25)#tghe last one for _neme_

# Create a vector of length n. Each component is either cb(i)='max' if the i-th criterion is benefit or
cb(i)='min' if the i-th criterion is a cost.
cb <- rep("max",n_criteres )

#beware for ITA there are 7 criteria


#install.packages("FuzzyMCDM")
#install.packages("RankAggreg")

library("FuzzyMCDM")
library("RankAggreg")
#set default parameter

lambda <- 0.5
v <- 0.5

#run MULticriteria algorithms

MMoora = FuzzyMMOORA(decision=decision_matrix , weights = mes_criteres_w, cb)
TopsisL = FuzzyTOPSISLinear(decision=decision_matrix , weights = mes_criteres_w, cb)
TopsisV = FuzzyTOPSISVector(decision=decision_matrix , weights = mes_criteres_w, cb)
Vikor = FuzzyVIKOR(decision=decision_matrix , weights = mes_criteres_w, cb, v)
Waspas = FuzzyWASPAS( decision=decision_matrix , weights = mes_criteres_w, cb,lambda)

#Meta-Ranking
MetaR = MMoora[,8]+TopsisV[,6]+TopsisL[,3]+Vikor[,14]+Waspas[,5]


t1 <- Sys.time()
#Ranking Aggregated
#    ra = rbind(MMoora[,8],TopsisV[,6],TopsisL[,3],Vikor[,14],Waspas[,5])
```





Funded by
the European Union

```
#   if(nrow(decision_matrix)<=10) {
#     RA = RankAggreg::BruteAggreg(ra, nrow(decision_matrix ), distance="Spearman")
#   } else {
#     RA = RankAggreg::RankAggreg(ra, nrow(decision_matrix ), method = "GA", distance =
"Spearman", verbose=FALSE, maxIter = 3000)
#  }
# Did not converge after 3000  iterations and it is not surprise with environment data.
t2 <- Sys.time()
t2-t1
#40.06 min


results = data.frame(Alternatives = 1:nrow(decision_matrix ), MMOORA = MMoora[,8],
TOPSISVector = TopsisV[,6],
                TOPSISLinear = TopsisL[,3], VIKOR = Vikor[,14], WASPAS = Waspas[,5],
                MetaRanking_Sum = rank(MetaR, ties.method= "first") ) #, MetaRanking_Aggreg =
RA$top.list)

#add meta ranking to our fuzzy detection matrix

fuzzy_detection_matrix[, paste0(analyte,"_MR_Sum") ] = results[ , "MetaRanking_Sum"]
#save the work

write.table ( fuzzy_detection_matrix , paste0(path_grappe, analyte, "_fuzzy_detection_matrix.csv"
), sep=";")

##lests end monitoring the procedure

t_end = Sys.time()
t_end  - t_start # all these produced in less time than a KED elemental kriging !
#Time difference of 13.79  mins

#Interpreting the results

#5-item Likert scale for anomaly de detection as reported by the fuzzy_detection_matrix
#Virtually certain VC
#Very likely VL
#Likely L
#About as likely as not ALN
#Unlikely U

#Dust and moss indicatore have a #5-item Likert scale for abondance
#low L
#medium low ML
#medium M
#medium High MH
#High H
```





Funded by the European Union