Environment International 185 (2024) 108544

Contents lists available at ScienceDirect

Environment International

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

Full length article

Modeling arsenic in European topsoils with a coupled semiparametric (GAMLSS-RF) model for censored data

Arthur Nicolaus Fendrich [a,b,c,*], Elise Van Eynde [a], Dimitrios M. Stasinopoulos [d], Robert A. Rigby [d], Felipe Yunta Mezquita [a], Panos Panagos [a]

[a] European Commission, Joint Research Centre (JRC), Ispra, VA, Italy
[b] Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ-UPSACLAY, 91190 Gif sur Yvette, France
[c] Université Paris-Saclay, INRAE, AgroParisTech, UMR SAD-APT, 91120 Palaiseau, France
[d] School of Computing and Mathematical Sciences, University of Greenwich, Greenwich, UK

ARTICLE INFO

Keywords:
Arsenic
GAMLSS
Random forest
Soil contamination
Statistical modeling
Trace element

ABSTRACT

Arsenic (As) is a versatile heavy metalloid trace element extensively used in industrial applications. As is carcinogen, poses health risks through both inhalation and ingestion, and is associated with an increased risk of liver, kidney, lung, and bladder tumors. In the agricultural context, the repeated application of arsenical products leads to elevated soil concentrations, which are also affected by environmental and management variables. Since exposure to As poses risks, effective assessment tools to support environmental and health policies are needed. However, the most comprehensive soil As data available, the Land Use/Cover Area frame statistical Survey (LUCAS) database, contains severe limitations due to high detection limits. Although within International Organization for Standardization standards, the detection limits preclude the adoption of standard methodologies for data analysis. The present work focused on developing a new method to model As contamination in European soils using LUCAS soil samples. We introduce the GAMLSS-RF model, a novel approach that couples Random Forests with Generalized Additive Models for Location, Scale, and Shape. The semiparametric model can capture non-linear interactions among input variables while accommodating censored and non-censored observations and can be calibrated to include information from other campaign databases. After fitting and validating a spatial model, we produced European-scale As concentration maps at a 250 m spatial resolution and evaluated the patterns against reference values (i.e., two action levels and a background concentration). We found a significant variability of As concentration across the continent, with lower concentrations in Northern countries and higher concentrations in Portugal, Spain, Austria, France and Belgium. By overcoming limitations in existing databases and methodologies, the present approach provides an alternative way to handle highly censored data. The model also consists of a valuable probabilistic tool for assessing As contamination risks in soils, contributing to informed policy-making for environmental and health protection.

1. Introduction

Arsenic (As) is a versatile heavy metalloid trace element used in the production of semiconductors, batteries, paints, wood preservatives (Flora, 2015), plant defoliants, agricultural pesticides, and herbicides (Adriano, 1986), among others. As is the 53rd most abundant element among the 92 that occur naturally in the Earth's crust (Reimann et al., 2009), with a median global total concentration in soils estimated to be 5 mg kg^{-1} (Reimann and De Caritat, 1998), and an average of 7.2 mg kg^{-1} (Adriano, 1986). Despite being a non-essential element for humans (Medunić et al., 2019), the medical use of As dates back to the time of Hippocrates (Klaassen, 2013), and its use as a poison is reported to have happened in Roman times (Reimann et al., 2009). For its toxicity, As was one of the first chemical elements identified as a cause of cancer in the 19th century (Smith et al., 2002). As is the only known carcinogen that presents risks to humans by both inhalation and ingestion (McLaren et al., 2006), and currently, it is understood that exposure to As relates to the development of vascular diseases and to increased risk of liver, kidney, lungs, and bladder tumors (Palma-Lara et al., 2020). Human exposure to As can be detected through blood, hair, and urine samples.

* Corresponding author at: Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ-UPSACLAY, 91190 Gif sur Yvette, France.
E-mail address: arthur.fendrich@lsce.ipsl.fr (A.N. Fendrich).

https://doi.org/10.1016/j.envint.2024.108544
Received 12 December 2023; Received in revised form 21 February 2024; Accepted 28 February 2024
Available online 1 March 2024

While concentrations of 0.1 to 0.5 mg kg^{-1} in hair samples may indicate chronic As poisoning, the acute ingestion of 100 to 300 mg can be fatal after one to four days (Ratnaike, 2006). Worldwide, It is estimated that 226 million people are exposed to As contamination from drinking-water or food. Asia, with 174.1 million people at risk, accounts for most of this global exposure (Murcott, 2012).

In agricultural areas with repeated application of organic or inorganic arsenical products, very high concentrations of As can be detected due to the continuous accumulation of their residuals (Adriano, 1986). Examples of such products include As detected in manure (Adamse et al., 2017), herbicides based on dimethylarsinic acid and pesticides based on sodium arsenite (Saxe et al., 1964). Once in the soil environment, As molecules can react and become sorbed onto the solid phase of the soil, be uptaken by plants, be volatilized back into the atmosphere, or leach (McLaren et al., 2006). Experiments after the continuous application of As pesticides in the United States showed that As did not leach below the 20 cm depth (Veneman et al., 1983), but field studies in Denmark found traces of As contamination in up to 2.5-meter depth (Lund and Fobian, 1991), indicating a different behavior. The mobility and availability of As in soils are affected by environmental factors - such as pH, soil texture, clay minerals, metal (hydr)oxides, and redox potential -, and by management practices, such as the application of phosphorus products, the use of plowing, and the adoption of cover crops (Adriano, 1986). Other than the application of agricultural products, As accumulation in the soil can also result from the redeposition of atmospheric As particles, the contamination of surrounding areas by mining and smelting activities, the deposition of ashes after coal combustion, the disposal of urban and industrial wastes, or the spread by irrigation (McLaren et al., 2006).

Apart from the exposure to high doses, contact with As can also be a problem at lower concentrations. In the case of soil contamination, some potentially harmful activities include direct contact with the skin and hand-to-mouth ingestion through recreation, landscaping, and gardening (Klaassen, 2013; Venteris et al., 2014). According to the Integrated Risk Information System of the United States Environmental Protection Agency, the estimated increased cancer risk due to oral exposure to inorganic As equals 1.5 per mg kg^{-1} day^{-1}, being higher than that of insecticide toxaphene, 1.1 per mg kg^{-1} day^{-1}, and similar to the fungicide hexachlorobenzene, 1.6 per mg kg^{-1} day^{-1} (USEPA, 2023a). Consequently, the generic screening level for total inorganic As in residential soils is recommended to be as low as 0.68 mg kg^{-1}, indicating that sites exceeding such threshold may require further investigation of their carcinogenic potential (USEPA, 2023b).

Due to the high toxicity of As and its low generic threshold, an information tool to assist the development of environmental, health and soil policies must be able to estimate the risk of contamination for multiple reference threshold values. However, implementing this idea can face further complications. For instance, the Land Use/Cover Area frame statistical Survey (LUCAS) topsoil database, the largest and most comprehensive soil sampling campaign across the European Union (EU), collected information on As concentrations at more than 20,000 locations (Orgiazzi et al., 2018). However, the analytical procedures adopted do not allow a proper quantification of the values below the Limit of Quantification (LOQ) of 2.84 mg kg^{-1} (Tóth et al., 2016). With such a high LOQ, the LUCAS samples can be divided into two groups: i) the non-censored observations, for which we know the exact measured As concentration, and ii) the censored observations, for which we can only know that the measurement is inferior to 2.84 mg kg^{-1} (i.e., the interval where the measurement belongs). One way to potentially overcome this issue could be by incorporating soil samples from other campaigns, such as the Forum of European Geological Surveys (FOREGS) database (Salminen et al., 2005), the Geochemical Mapping of Agricultural Soils (GEMAS) database (Fabian et al., 2014), or national soil monitoring systems, such as the Réseau de Mesures de la Qualité des Sols from France (Marchant et al., 2017). However, because these observations were made on different dates, often years apart, using different and non-harmonized sampling and analytical procedures, combining databases would demand strong assumptions and a very extensive harmonization step.

Another potential way to take advantage of the large number of observations in the LUCAS database without the need for strong pre-processing assumptions is by developing methods to handle the particular data characteristics. For instance, the method used by Tóth et al. (2016) to generate European As maps does not mention how the censored observations were handled, which raises concern about the reliability of the spatial patterns obtained. Additionally, the method assumes a linear dependence of the As concentration on the spatial covariates, while more modern approaches suggest that the relationships among variables may contain complex high-order interactions (Van Eynde et al., 2023; Helfenstein et al., 2022; Ballabio et al., 2021). Ideally, a proper method for the LUCAS As data would take both limitations into account while preserving the strengths of the original dataset, such as the applicability at a continental scale.

In the present work, we generate maps of As concentration at the European scale based on the LUCAS 2009 database using a novel approach. We do so by presenting a new model in Section 2.3, which consists of coupling Random Forests (RF) to the Generalized Additive Models for Location, Scale and Shape (GAMLSS) framework. The model is named GAMLSS-RF after its components. The proposed semi-parametric approach models the censored and non-censored parts in a coupled manner, allowing the reconstruction of missing information by borrowing information from the other observations. The model selection process is presented in Section 2.4, and the resulting chosen model is given in Section 3.2. After thorough model calibration and validation procedures (see Section 2.5), we produce maps of As concentration across most EU member states at a 250 m spatial resolution and evaluate the exceedance probabilities concerning two limits of action and a threshold selected as representative of the background concentrations (see Section 3.3). Then, we discuss in Section 4 the policy implications of the results obtained. Conclusions are given in Section 5. A list of the abbreviations used in this work is provided in SM13.

2. Materials and methods

2.1. Soil samples and LUCAS topsoil survey

The As observations used in the present study come from the LUCAS topsoil survey, the largest periodic survey to collect topsoil information across Europe (Orgiazzi et al., 2018). The LUCAS database contains over 20,000 topsoil samples taken in European countries (SM7) and discloses information about soils' physical, chemical, and biological properties for different land use types in the years 2009 (plus 2012 for Bulgaria and Romania), 2015, 2018, and 2022 (EC, 2023a). Beyond the general topsoil information, the 21,682 soil samples of the LUCAS 2009/2012 survey were also analyzed for heavy metals and metalloids quantification and other elements, including Sb, As, Cd, Co, Cr, Cu, Fe, Pb, Hg, Mg, Mn, Ni, V, Zn.

Pseudototal concentration of metals and metalloids in LUCAS 2009/2012 soil samples were firstly obtained by using the aqua regia extractable fraction (HNO_3/HCl 1.5/4.5 v/v) and microwave-assisted digestion (140 °C, 35 min, 20 bar) (prEN16174) (Carmen-Ileana et al., 2014; Cristache et al., 2014), and then quantified by using inductively coupled plasma-optical emission spectrometry (ICP-OES). This analytical procedure differs from that used by the GEMAS topsoil database where a modified aqua regia extractable fraction (HNO_3/HCl/H_2O 1/1/1 v/v/v) and open digestion (95 °C, 60 min) was used for metals extraction and quantification was then carried out by using Inductively coupled Plasma quadrupole mass spectrometer (ICP-QMS). Due to these methodological differences, the limit of quantification (LOQ) of arsenic in the LUCAS soil samples (2.84 mg kg^{-1}) is significantly higher than that of the GEMAS database (0.05 mg kg^{-1}) (Tarvainen et al., 2013). However, the LOQ of arsenic from LUCAS database is similar to that obtained by testing wavelength dispersive X-ray fluorescence

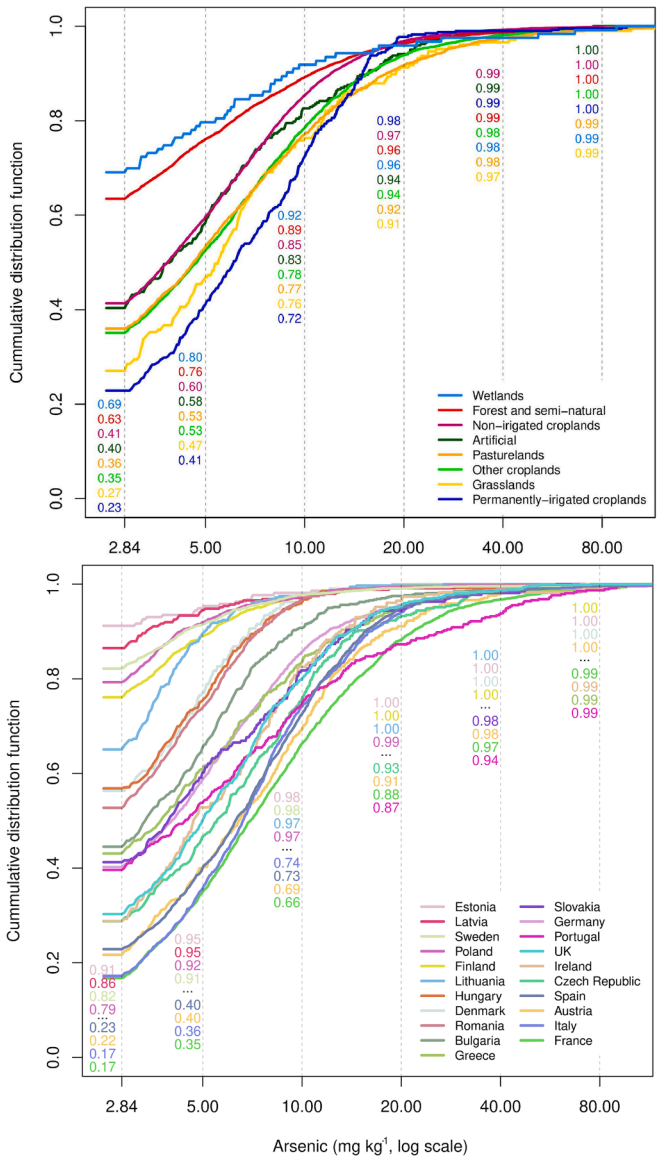

Fig. 1. Sample cumulative distribution of Arsenic stratified per land use (top) or country (bottom). The numbers in the top plot correspond to each land use class, while in the bottom plot correspond to the 4 top and bottom country classes.

spectrometry (XFS) on the GEMAS samples (i.e., 3.0 mg kg^{-1}), which led to a fraction of 25 % of the XFS observations below the detection limit on that database (Tarvainen et al., 2013).

2.2. Spatial covariates

Since the As observations of the LUCAS database are spatially explicit, our model covariates correspond to point attributes extracted from digital maps. The set of 17 variables used in the current work covers:

- 8 soil properties related to the As chemistry on soils based on LUCAS topsoil data published by Ballabio et al. (2016): pH, soil organic carbon content (SOC), cation exchange capacity (CEC), concentrations of phosphorus and calcium carbonate ($CaCO_3$), fractions of clay, sand and silt;
- 1 variable representing land cover, namely the normalized difference vegetation index (NDVI) (USGS, 2022);
- 2 landscape features namely terrain slope and elevation (DEM) (EEA, 2016);
- 2 climatic variables: annual average temperature and precipitation (Noce et al., 2020);
- and 4 indicators of anthropogenic activity: distance to mines (Lopes et al., 2018), distance to roads (OpenStreetMap, 2018), lights at night (Elvidge et al., 2017), and distance to coal, oil and gas (COG) industries (ResourceWatch, 2019).

Prior to any processing, all the distance variables were converted to the log-scale, and the datasets were spatially resampled to the target model spatial resolution of 250 m x 250 m.

2.3. Exploratory analysis and modeling

The LUCAS 2009/2012 database contains 21,682 samples, of which 329 do not have As data available. In the remaining 21,353 observations, 9,784 (i.e., 45.82 %) are below the LOQ of 2.84 mg kg^{-1}. Such a censored nature of these As observations has several implications to the exploratory analysis and modeling procedures. For instance, the commonly used distribution moments, such as the mean and variance, can not be calculated to characterize the data, and quantiles have to be used as an alternative. In that case, the only quantiles that can be obtained are those that exceed the fraction of observations below the detection limit for a given subset of the data. To deal with such restriction, the exploratory analysis in this work consisted of reporting the empirical cumulative function for the As concentration. The data was split into two different selections for exploratory purposes: by European country and by land use type.

Another implication of having a high proportion of censored As data is that the adoption of common simplifications found in the literature for similar cases, which include removing censored observations or replacing them with a fixed value within the interval they represent (Ballabio et al., 2019; Helsel, 1990), would have huge impacts on the results and can not be used without major drawbacks (i.e. losing important information) and biases. Such techniques are only less problematic when the fraction of censored observations is at most 10 % (Williams et al., 2020). These properties also mean that most methods that successfully handle similar problems do not support the use of left-censored data and, therefore, can not be used for the LUCAS As data. These approaches include quantile RFs for the spatial distribution of Zn in topsoils (Van Eynde et al., 2023) or soil pH (Helfenstein et al., 2022), regression-kriging for heavy metals (Rodríguez-Lado et al., 2008), deep neural networks for the Hg content in the topsoil (Ballabio et al., 2021), Gaussian process regression for chemical properties, such as N, P and the C/N ratio (Ballabio et al., 2019), among others. In this sense, a proper method for our data would allow the use of left-censored positive data while still capturing the high-dimensional interactions between variables that proved successful in similar contexts.

To address these issues, the proposed GAMLSS-RF model couples a RF model (Breiman, 2001) to the semiparametric regression GAMLSS framework (Stasinopoulos et al., 2018; Rigby and Stasinopoulos, 2005). In GAMLSS, the response variable can be assumed to have any parametric distribution, and all distribution parameters (i.e., location [e.g., mean], scale, and shape) can vary according to parametric or nonparametric functions of the explanatory variables. Because GAMLSS do not have the same distributional limitations as other statistical frameworks, e.g., Linear Models or Generalized Linear Models, standard distributions can be properly modified to capture relevant properties of the data, such as skewness, heavy tails, bimodality, truncation, and (left-, right- or interval-)censoring. Parameter estimation in GAMLSS is achieved through iterative procedures to maximize the (penalized) log-likelihood. These procedures contain a backfitting component, which allows the incorporation of several nonparametric techniques, such as neural networks, Multivariate Adaptive Regression Splines, and RFs (Stasinopoulos et al., 2017).

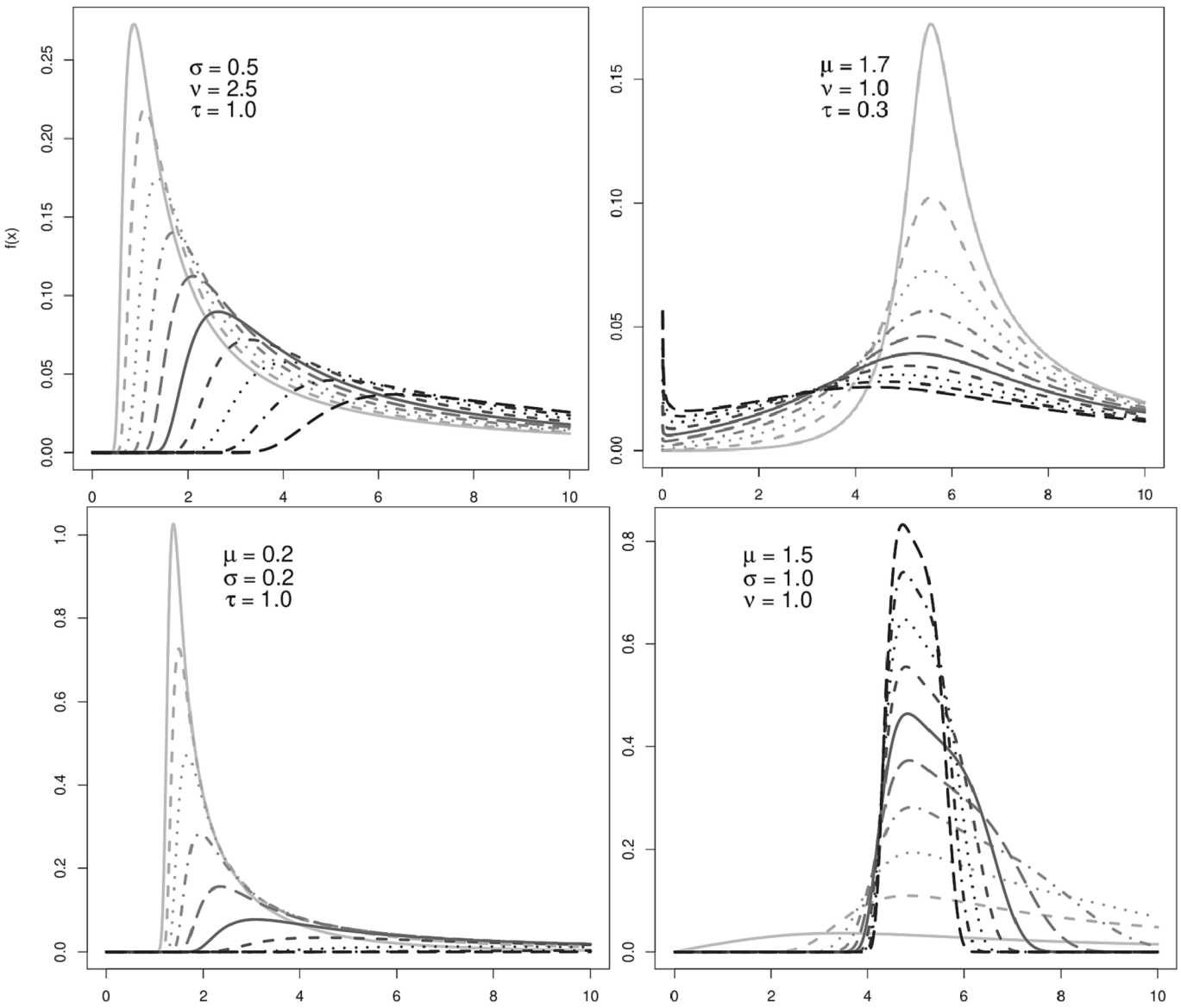

Fig. 2. The marginal effect of the μ (top left), σ (top right), ν (bottom left), and τ (bottom right) parameter on the probability distribution function of the log-transformed sinh-arcsinh (logSHASHo) distribution. The parameter μ varied from −1 to 1, σ from 0.1 to 0.7, ν from 1.5 to 5, and τ from 0.5 to 10, and a gradient from gray to black indicates an increase in the parameter.

The second component of the GAMLSS-RF is the nonparametric RF model. Standard RFs consist of a learning method combining many tree-based models (Breiman, 2001), which can capture high-order interactions in the data by partitioning the feature space into disjoint regions (Hastie et al., 2009). The method contains two sources of randomness. The first consists of the different samples with replacement taken and used to construct each tree, and the second is the random subset of the explanatory variables from which a variable is chosen to partition the feature space in each step (Fawagrehet al., 2014), with both procedures aiming at increasing robustness to noise and reducing overfitting and the variance of predictions (Hastie et al., 2009). The main advantages of RFs compared to standard parametric smoothers are their higher predictive performance and ability to capture complex multidimensional relationships, although at the cost of having harder interpretability or explainability (Aria et al., 2021; 2023).

2.4. Model selection

In GAMLSS-RF, RFs can be used as nonparametric learners for one or more of the distribution parameters, so we searched for the best possible model with several steps:

1) First, we divided the dataset of 21,353 observations into training and validation datasets with 12,811 and 8,542 entries (i.e., approx. 60 %, 40 %), respectively. Then, we expanded the set of 23 statistical distributions for positive continuous values available in the *gamlss* R implementation to include 31 distributions for real continuous data that were exponentially-transformed to modify its range to the positive continuous line (Rigby et al., 2019). Then, we modified the probability density functions of the resulting set of 54 distributions to handle censored variables (Stasinopoulos et al., 2017).
2) Next, we selected the best marginal statistical distribution among the 54 options by comparing quantitatively and qualitatively their results when fitted to the training data without any predictors. Several distributions failed to converge to a solution. For the successful ones, the quantitative and qualitative criteria used were their deviance (i.e., minus two times the log-likelihood) and a visual residual analysis, respectively.

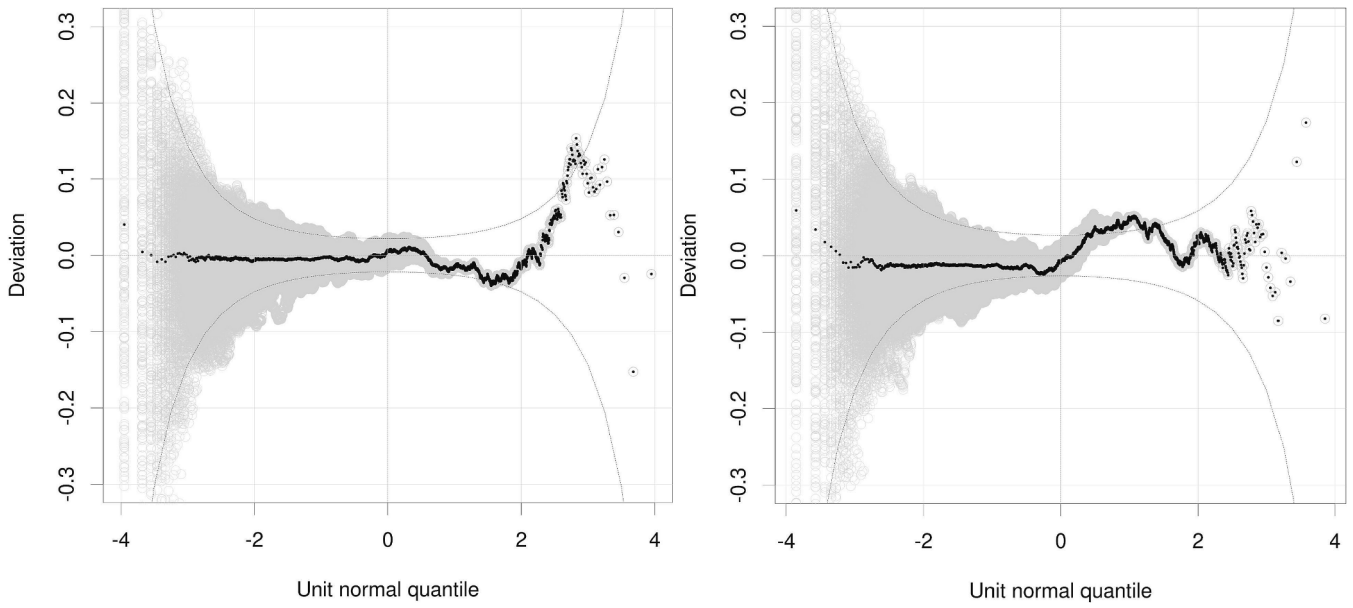

Fig. 3. Worm plot of the normalized randomized quantile residuals for the training dataset (left) and the validation dataset (right).

3) For the best marginal distribution selected in (2), a standard RF with default hyperparameters and 200 trees was added as a potential predictor for each distribution parameter (i.e., μ, σ, ν and τ - see Section 3.2) and their possible combinations. The results were also evaluated quantitatively and qualitatively using the training dataset. After finding the best model structure, the resulting GAMLSS-RF model was used to fine-tune two RF hyperparameters: the number of trees (*num.trees*) and the number of variables considered in each split (*mtry*). The first hyperparameter was allowed to vary from 10 to 250 in steps of 10, and the second, from 1 to 17 in steps of 2.
4) Since the results of (3) selected a RF learner only for the first distribution parameter μ, a linear model was tested for the remaining parameters. To do so, we first reassessed the choice marginal distribution of (1) and then used a stepwise strategy by adding linear terms to the distribution parameter σ in a *forward* manner. A limit of three variables was defined and the Bayesian Information Criterion (BIC) was used to compare models in every step.
5) The result of (4) did not provide an adequate fit to the training dataset. In particular, the multiple worm plot of the residuals against the distance to mines was inadequate. To mitigate this issue, the last step consisted of manually fine-tuning the GAMLSS model and two RF hyperparameters. For the GAMLSS model, a linear term for the distance to mines variable was added to ν and τ to improve the distance to mines multiple worm plot diagnostics (see Section 2.5). For the RF, the fraction of observations used to grow each tree (sample.fraction) and the fraction of the samples used to select tree splits (*honesty.fraction*) were manually fine-tuned by iteratively subtracting 0.05 from the default values and visually assessing the impact on the residuals calculated using the training dataset.

2.5. Residual diagnostics and predictive assessment

The model diagnostics were made through residual analysis. The standard raw residuals (i.e., defined by the difference between model predictions and observations) could not be used since they are not well-defined for censored observations and do not generalize to other distributions than the Gaussian. A possible alternative in this case is the normalized randomized quantile residuals (NRQR). NRQRs result from a probability integral transform of the As values given their fitted distribution, with an additional randomization procedure for the censored observations. Consequently, if a GAMLSS model is adequate for the response variable being analyzed, then its NRQRs have an approximate standard normal distribution, which can be assessed using, for example, detrended quantile–quantile plots (also called worm plots) (Stasinopoulos et al., 2017). Worm plots can be used to evaluate the overall model accuracy (i.e., single worm plot) or the model accuracy for different ranges of the explanatory variables (i.e., multiple worm plots).

In the residual analysis, we evaluated model adequacy visually by plotting the single and multiple worm plots of the NRQRs from the fitted model using the training and validation datasets, with approximate 95 % intervals. While the single worm plot allowed us to evaluate the overall model performance, the multiple worm plots enabled the investigation of possible systematic prediction biases. The NRQRs of the training dataset were also used for a spatial autocorrelation analysis, where a variogram was constructed based on three different models (i.e., Matern, Spherical and Gaussian), and the nugget-to-sill ratio was calculated. The NRQRs and the worm plots of the validation dataset allowed us to check the model's adequacy for extrapolation. Since NRQRs contain a random component due to the censored observations, we did 250 repetitions in each case and calculated summary statistics.

Besides, to improve our understanding of the model's internal behavior, we calculated each explanatory variable's importance in three ways. The first method corresponds to the change in deviance that resulted from randomly permuting the values of each covariate. A total of 250 repetitions per variable were performed. Such a method was adopted for its popularity, but the model extrapolation that it tends to induce likely limits its power (Hooker et al., 2021). For this reason, the second importance measure was calculated for each variable by the change in deviance resulting from refitting the chosen GAMLSS-RF model after removing them from the set of explanatory variables. In this case, 250 pseudorandom number generators were used, resulting in different fitted RF models. The third importance measure was calculated by generating the isolated effect of each variable (i.e., the Accumulated Local Effects (ALE) plot (Apley and Zhu, 2020), and using the range used as a measure of practical importance. The permutation and leave-one-out methods evaluate each variable's *statistical importance*, while the ALE plot range evaluates their *practical importance*. For improved comparison, the results were divided by the maximum absolute effect, which constrained the absolute values to the (0, 1) interval.

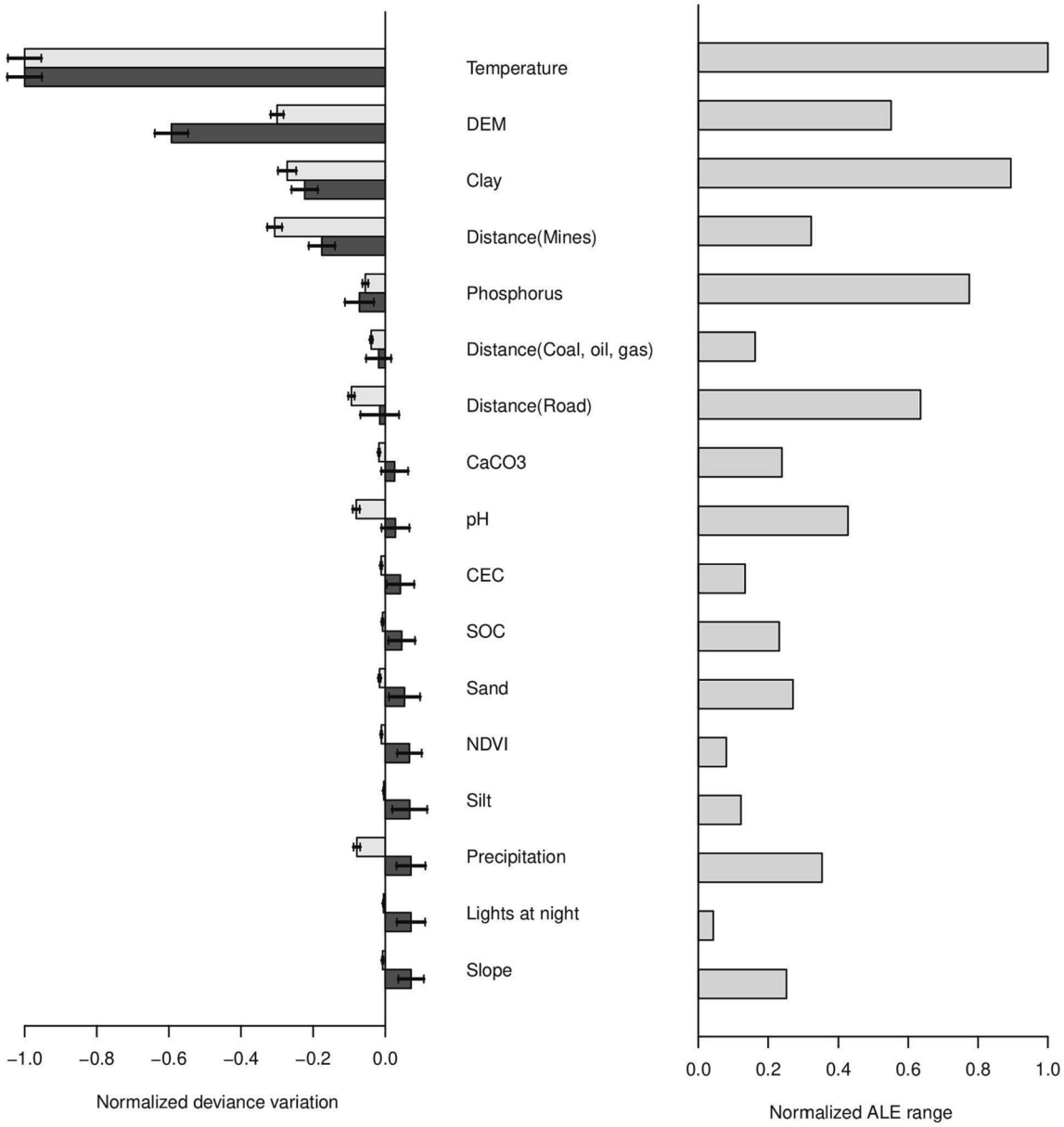

Fig. 4. Scaled importance measures: *statistical measures* based on the deviance variation (left) and a *practical measure* based on the Accumulated Local Effects (ALE) range (right). In the left plot, the feature permutation measure is shown in light gray and the leave-one-out measure, in dark gray. The uncertainty bars refer to the standard deviation of 250 repetitions of the method.

2.6. Development of European maps of arsenic concentrations

As mentioned in Section 2.4, a new statistical distribution was generated by modifying the probability density function (PDF) of the best distribution found to account for the censoring on the interval (0, 2.84]. Such a modification solely affects the values within the two extremes of the censoring interval, where the continuous PDF from the original distribution is replaced by a mass point equal to the integral over the interval, while the other parts of the distribution do not change. The definite integral is not invertible, which, in practice, means that we don't have enough information to reconstruct the distribution of the lower As values with certainty. Despite this limitation, we adopted the additional assumption that the best reconstruction for the left tail of the censored PDF is the original PDF itself. Such an assumption is notably strong, but natural since the adopted approach couples all parts of the distribution, meaning that modifications made to its right tail also affect its left tail, and vice-versa. Therefore, we assumed that a properly fitted model, which should be necessarily well-adjusted to the 11,569 non-censored observations and the 9,784 censored observations simultaneously, should contain enough information to extrapolate on the missing range of values.

With the additional assumption for the reconstruction of the left tail, we generated maps describing the estimated median of As and evaluated the spatial patterns across European countries and against two action levels and one background concentration. The action levels taken were those reported by Tarvainen et al. (2013): the limit of good soil status of 20 mg kg^{-1} in Norway, which is also the maximum tolerable concentration in agricultural soils of Germany (Reimann and De Caritat, 1998), and the threshold of 45 mg kg^{-1} defined on Belgium's 1995 Soil Remediation Act (bottom). The background concentration adopted was that estimated by Taylor and McLennan (1995) in the upper continental crust: 1.5 mg kg^{-1}. The adequacy of these limits is discussed in the discussion section. Per-country average values were calculated by sampling from their pixels' predicted distribution. This procedure was repeated 500 times, and the average and standard deviation were

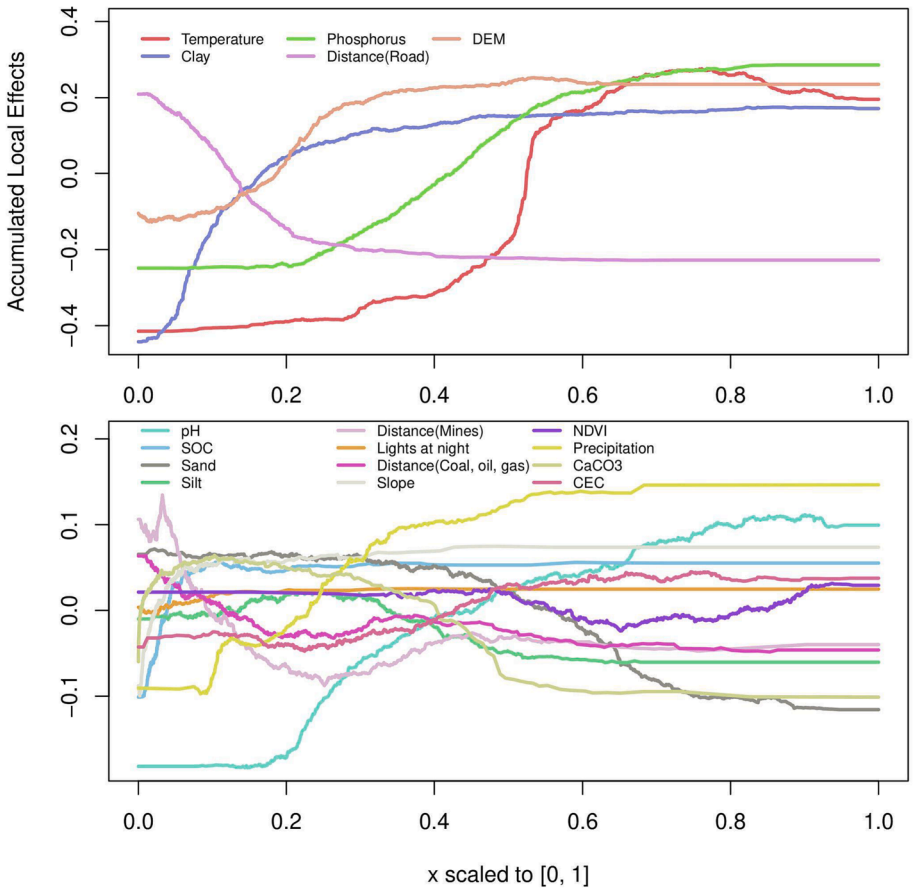

Fig. 5. Accumulated Local Effects (ALE) on the μ parameter (i.e. the log of the median As concentration) for the 5 explanatory variables of higher practical importance (top) and the other variables (bottom). The distance variables are displayed in their original scale, not log-transformed.

calculated. Cyprus and Malta are not included in the model results due to the lack of observations.

Due to systematic database differences, the model obtained in Section 2.4 (i.e., the 'fitted LUCAS model') naturally does not perform well against the GEMAS dataset (SM9). However, since GEMAS is a valuable source of information for As in soils, we generated an alternative model version called the 'calibrated GEMAS model'. The calibrated GEMAS model was obtained by first filtering the GEMAS dataset to keep only the observations in cropland areas, whose samples were taken in the 20 cm of the topsoil (Tarvainen et al., 2013). This step attempted to reduce the divergences between GEMAS and LUCAS data. Next, the NRQRs of the GEMAS observations were extracted using the fitted LUCAS model, and a best distribution (i.e., the generalized t distribution, according to the log-likelihood criterion) was fitted to them. Following the derivation in Stasinopoulos et al., (2017, p.441–442), this information can be combined with the fitted LUCAS model to derive a new, calibrated, fitted model consistent with the GEMAS observations. All results presented in the current work refer to the 'fitted LUCAS model', unless mentioned otherwise. Calculations for the calibrated GEMAS model were made in a lower spatial resolution (i.e., 1000 m) for computational speed. The GAMLSS-RF model was implemented using the *gamlss* R package (Rigby and Stasinopoulos, 2005) and the *grf* package (Tibshirani et al., 2023).

3. Results

3.1. Exploratory analysis

Fig. 1 (top) presents the empirical cumulative distribution function for the As concentrations separated by land-use class. It shows that wetlands and the group of forests and semi-natural areas contain the largest shares of observations smaller than 2.84 mg kg^{-1}, 69 % and 63 %, respectively, while non-irrigated croplands, artificial areas, pasturelands, and other croplands follow a gradient from the higher to the lower shares. Grasslands and permanently irrigated croplands present the lowest shares among all uses, at 27 % and 23 %, respectively. With a few changes in the ordering between uses, 91 % to 98 % of the observations in each class are lower than 20 mg kg^{-1}. In all cases, 97 % to 99 % of the observations are lower than 40 mg kg^{-1}. The stratification per country of Fig. 1 (bottom) shows a large variability. While in Estonia, Latvia, and Sweden, the shares of censored observations are 91 %, 86 %, and 82 %, respectively, these values equal 22 % for Austria and 17 % for both Italy and France, indicating a strong geographical trend of As concentration in the EU. The same three countries contain 6 %, 9 %, and 12 % of the observed As exceeding 20 mg kg^{-1} (in Italy, Austria, and France, respectively).

3.2. GAMLSS-RF modeling

The procedure to select the GAMLSS distribution yielded better results with the log-transformed sinh-arcsinh distribution (Jones and Pewsey, 2009), denoted logSHASH, after step (2) and its original parameterization (hereinafter referred to as logSHASHo) after step (4). This distribution is described by four parameters: μ, σ, ν, and τ. These parameters marginally control the location, shape, and scale of (non-censored) logSHASHo according to the patterns displayed in Fig. 2. Depending on the combination of parameters, the logSHASHo distribution can become more or less heavy-tailed and be uni or bimodal, indicating that a high degree of flexibility can be achieved. For some combinations of parameters, its mean (or expected value) is properly defined, but for others, the integral may diverge, while the median is always properly defined.

The model selected as the best presented the following structure:

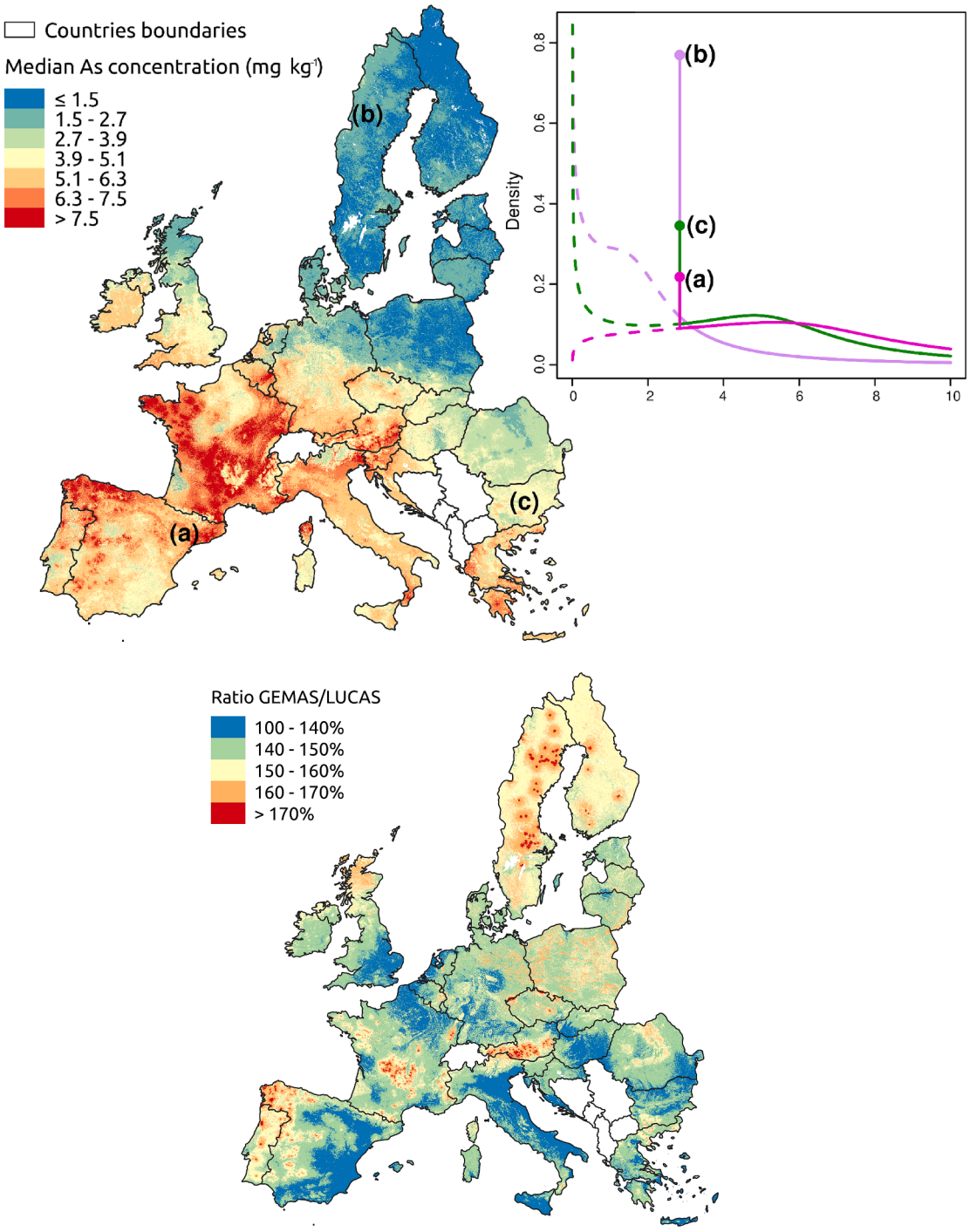

Fig. 6. Median arsenic concentrations in Europe: model predictions, along with three example points (top), and ratio between results using the GEMAS and the LUCAS database (bottom). In the top right plot, the continuous line is the probability density function for the censored logSHASHo distribution, while the dashed line shows the reconstructed left tail. The points on the vertical line show the corresponding fitted probabilities of being below the censor value 2.84 mg kg^{-1}. The bottom plot shows the ratio of predictions by the calibrated GEMAS model over the fitted LUCAS model.

$$
\begin{aligned}
y \;&\sim\; logSHASHo_c\;(\mu, \sigma, \nu, \tau) \\
\mu &= RF(\mathbf{X}) \\
\sigma &= exp[\,-0.613 - 0.473{\cdot}pH - 0.286{\cdot}Phosphorus - 1.466{\cdot}dist(mines)\,] \\
\nu &= 0.467 - 1.002{\cdot}dist(mines) \\
\tau &= exp[0.259 - 1.049{\cdot}dist(mines)\,]
\end{aligned}
$$

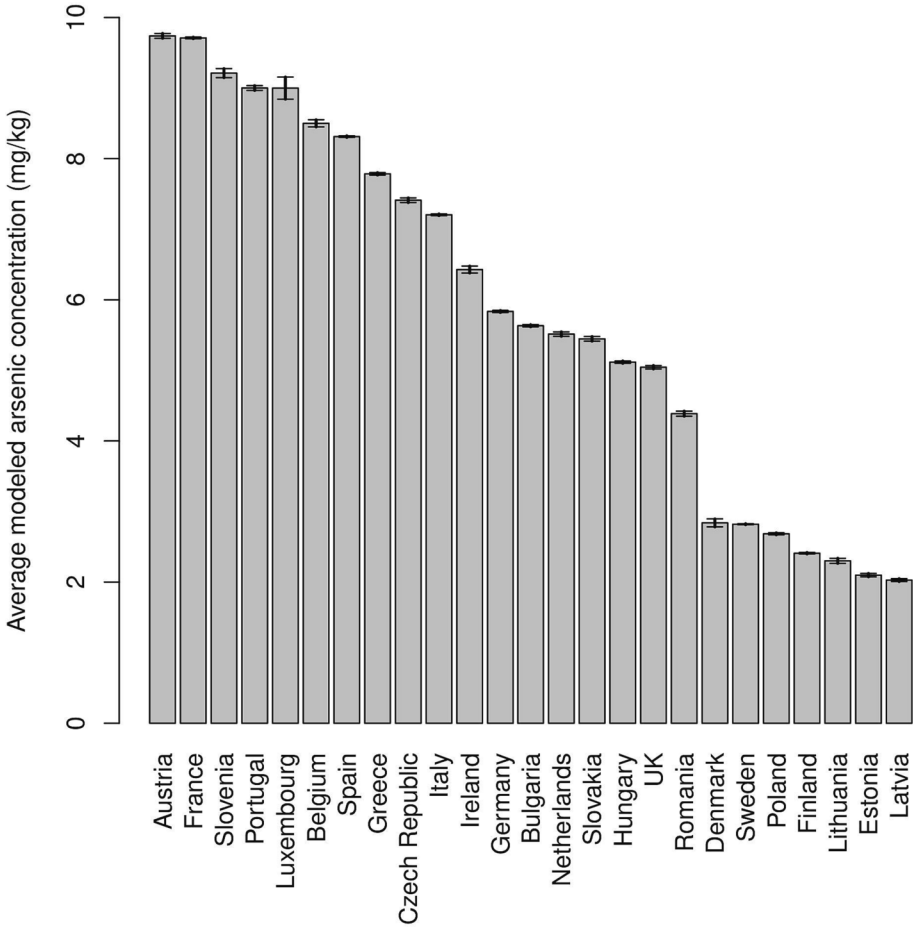

Fig. 7. Average modeled arsenic concentrations per European country. The uncertainty lines represent confidence intervals and are equal to two times the standard deviation of the 500 repetitions used to calculate the average.

with y denoting the As concentration in mg kg^{-1}; $logSHASHo_c$ being the censored version of the four-parameter logSHASHo distribution; $RF(\mathbf{X})$ denoting a RF learner including all explanatory variables, and selected parameters $num.trees = 160$, $mtry = 17$, $sample.fraction = 0.35$, and $honest.fraction = 0.35$; and *pH*, *Phosphorus* referring to the variables with the corresponding names, and *dist*(*mines*) referring to the log-transformed distance to mines variable, as detailed in Section 2.2.

The residual diagnostics displayed in Fig. 3 (left) indicate an adequate model for the training dataset, with most points falling within the approximate 95 % intervals. As a consequence of the number of censored observations, the left tail presents variability due to the 250 repetitions of the NRQRs (see Section 2.5), but no variability in the right tail. The assessment of spatial correlation with the nugget-to-sill ratio of the NRQRs for the Gaussian, Matern, and Spherical covariance functions presented median values of 0.70, 0.76 and 0.81, respectively, and averages ($\pm$ standard deviation) of 0.62 $\pm$ 0.23, 0.66 $\pm$ 0.24, and 0.76 $\pm$ 0.24, respectively, indicating low residual spatial correlation. The model was also found to be adequate according to the validation dataset (Fig. 3, right), as found by the small percentage of points outside the 95 % intervals. Such a result points to the adequacy of the model to predict outside the training dataset. The multiple worm plots split by the predicted median (SM10) points in the same direction overall, indicating a reasonable fit through most of the range of predicted values despite some deviations in the class of higher values.

The *statistical importance* measures of variables in Fig. 4 (left plot) indicate a strong influence of air temperature, distance to mines, terrain elevation and clay content. Fig. 4 (right plot) shows the *practical importance* measure of variables (i.e. the range of the ALE plot), indicating a strong influence of air temperature, clay content, phosphorus content, distance to roads and terrain elevation. The worm plot of the residuals from the validation dataset against each of these variables (SM1-5) indicates an overall good fit within the ranges of the explanatory variables, with the most serious violation happening for high Phosphorus content values (S2, top right). SM6 indicates that the linear *dist(mines)* terms were able to fix violations in the distance to mines multiple worm plots. Furthermore, Fig. 4 (left) shows that other variables had an ambiguous statistical impact in the model, varying according to the criteria used. This pattern indicates their lower statistical influence on the As model.

The ALE plots (Fig. 5, top) for the explanatory variables show an increasing relationship for temperature, terrain elevation, clay content, and phosphorus content, indicating that warmer and higher areas with clayey and phosphorus-rich soils tend to have higher As concentrations. Fig. 5 (top and bottom) also shows a decreasing relationship for the distance variables (i.e., from roads, mines and COG industries) indicating that As concentrations tend to be higher around areas of more intense human influence. A visualization of how the curves of Fig. 5 vary in space can be found in SM11 and SM12.

3.3. Arsenic in European soils

Fig. 6 (top) shows the median As concentrations calculated with the fitted LUCAS model at the 250 m spatial resolution for Europe. The values range from 1.1 to 64.6 mg kg^{-1}, with arithmetic and geometric means of 4.1 and 3.5 mg kg^{-1}, respectively. The map also contains three

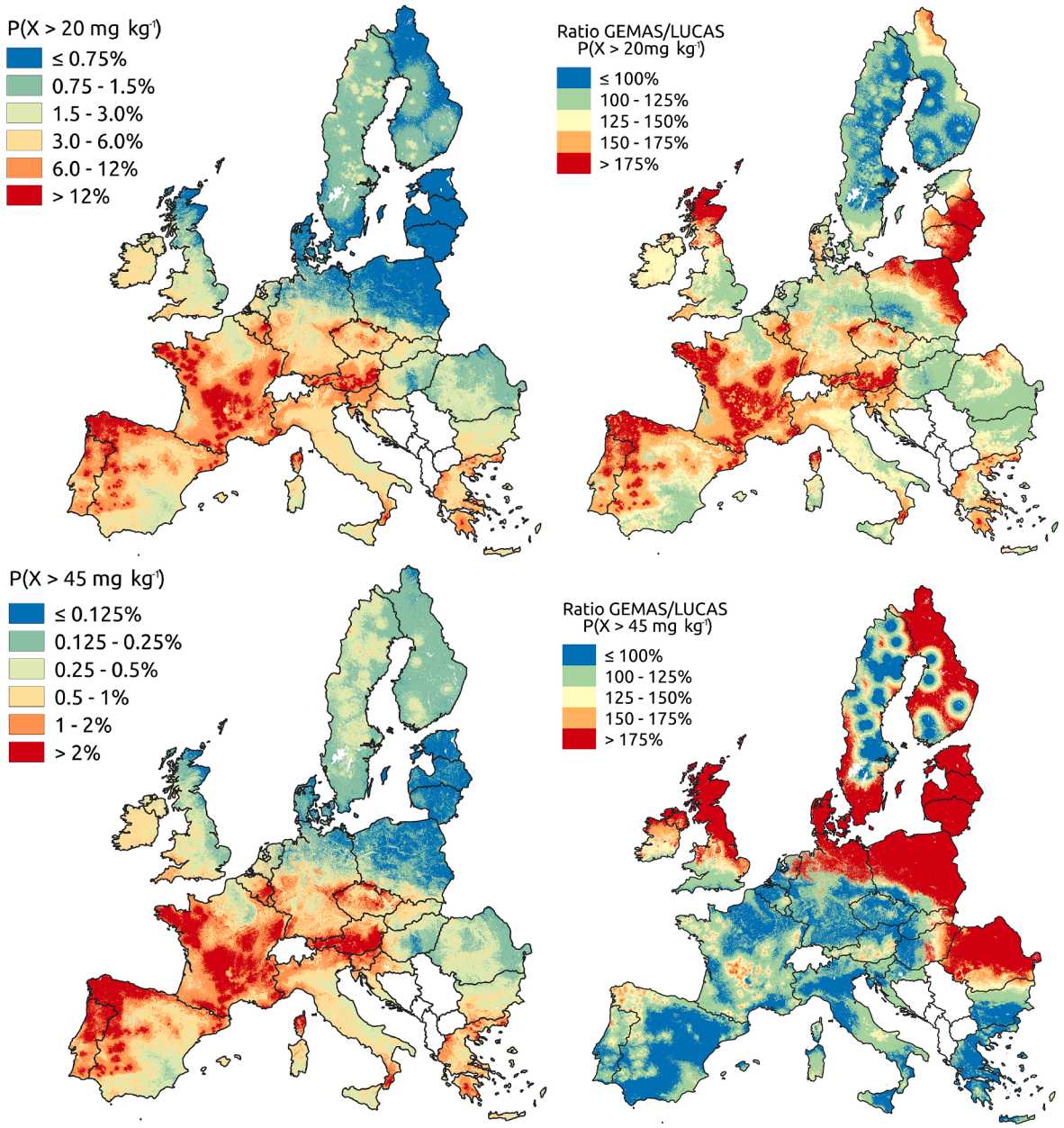

Fig. 8. Exceedance probabilities for two limits of action in Europe: 20 mg kg^{-1} (top) and 45 mg kg^{-1} (bottom). Results are predictions from the fitted LUCAS model (left), and the ratio of predictions from the calibrated GEMAS model to those of the LUCAS model (right).

points, (a), (b), and (c), in Spain, Sweden and Bulgaria, respectively. The corresponding estimated probability density function for these points have different shapes and reconstructed left tails. In the three cases, the logSHASHo distributions show a heavy right-tail. In Fig. 6, bottom, the ratio between the predictions from the calibrated GEMAS model over the fitted LUCAS model are presented. In all pixels, the calibrated GEMAS model (see SM8), predicts higher values than the fitted LUCAS model, with the ratio ranging from 131 to 254 %. Such a difference is more evident in Sweden and Finland, where the fitted LUCAS model calculates generally low values, but also happens in Austria and northwest Spain, where the fitted LUCAS model calculated high As concentrations. Per land use, average median predictions of 4.32, 4.94, 5.02 and 5.35 mg kg^{-1} are calculated for arable land, pasture, other agricultural areas and permanent crops, respectively.

The average value per country (Fig. 7) shows that Latvia, Estonia, Lithuania, Finland and Poland present the lowest averages, equal to 2.03, 2.10, 2.30, 2.41 and 2.68 mg kg^{-1}, respectively. Among the countries with the highest values, Luxembourg, Portugal, Slovenia, France and Austria present averages of 9.00, 9.00, 9.21, 9.71 and 9.74 mg kg^{-1}, respectively.

Fig. 8 shows the probability of pixels exceeding two soil As action levels in European countries, 20 mg kg^{-1} (top) and 45 mg kg^{-1} (bottom), obtained from the fitted distribution of As for each pixel. Austria, France, Spain, Portugal, and Belgium contain several locations where the chance of exceeding the Norwegian/German and Belgian thresholds (i.e. 20 mg kg^{-1} and 45 mg kg^{-1}, respectively) surpasses 12 % and 2 %, respectively. Germany, the Czech Republic, Slovenia, Italy and Greece also display a similar pattern, but to a more limited extent. The highest probability calculated for exceeding the first and second thresholds was 77.1 % and 58.1 %, respectively, with the two pixels belonging to France. A comparison with the calibrated GEMAS model indicates that the LUCAS dataset may be underestimating the risk against the 20 mg kg^{-1} threshold in Portugal, Scotland (in the United Kingdom - UK), France, Spain, Poland, Lithuania and Latvia. For the 45 mg kg^{-1}

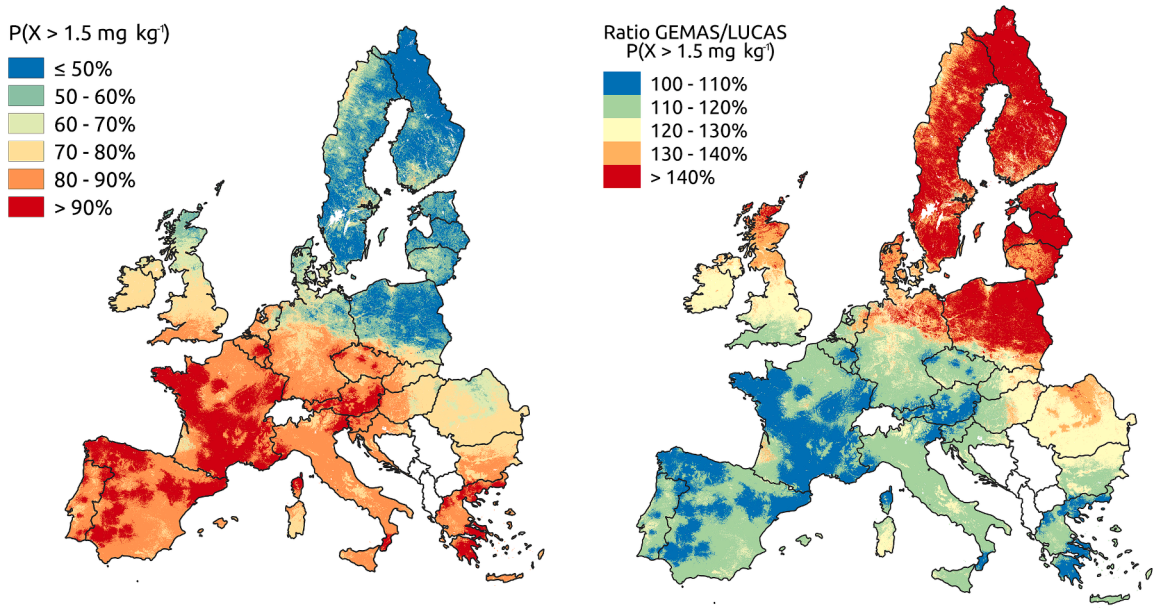

Fig. 9. Exceedance probabilities for the background concentration of 1.5 mg kg^{-1}. Results are predictions from the fitted LUCAS model (left), and the ratio of predictions with the calibrated GEMAS model (right).

threshold, the comparison with GEMAS indicates that the fitted LUCAS model may be overestimating the risk for most of Western Europe, but largely underestimating for the UK, Romania, Germany, Poland, Lithuania, Latvia and Estonia. In Sweden and Finland, both over and underestimation may be occurring.

The probability of exceedance concerning the background concentration of 1.5 mg kg^{-1} (Fig. 9) shows a high chance that the As concentrations from the fitted LUCAS model are higher than to the natural occurrence levels. In several locations in Poland, Sweden, Finland, Latvia, Estonia, and Lithuania, such a chance does not generally exceed 50 %. However, the comparison against the GEMAS dataset indicates that the risks may be underestimated in these countries.

4. Discussion

4.1. GAMLSS-RF approach and the drivers of as concentration in Europe

The diagnostics made on the residuals from the fitted LUCAS model indicated a good fit and ability to extrapolate beyond the training dataset, with very few violations to the 95 % pointwise confidence intervals (Fig. 3). Such a result was reinforced by the mostly flat worm plots obtained against the validation dataset (SM1–6) for different ranges of the five variables with the most practical influence on the model. Model residuals also indicated low spatial correlation due to a high nugget-to-sill ratio, which indicates a non-violation of the assumption of independence between observations. Comparison against the GEMAS samples indicated that the values represented in our fitted LUCAS model are possibly underestimated, but comparison must be taken with care due to methodological differences. The generation of maps of median values and exceedance probabilities was only possible due to the assumption that the left tail could be reconstructed (Fig. 6, right). Although necessary to overcome the data limitation problems described earlier, this assumption is strong and can be seen as a limitation of the modeling approach itself. However, In comparison to other works that mapped Arsenic in Europe, the GAMLSS-RF model advances other assumptions beyond the incorporation of censored observations. For instance, the inclusion of explanatory variables produces more detailed results than the kriging interpolations performed by Tarvainen et al. (2013), and the RF model is able to capture non-linear relationships, therefore extending the linear assumption of Tóth et al. (2016) and Rodríguez-Lado et al. (2008). Furthermore, applying a learning technique to censored data adds to recent efforts to improve As contamination mapping. Such efforts include, for example, the detection of As concentrations using hyperspectral data using RFs (Agrawal and Petersen, 2021), the use of several machine learning algorithms to estimate As concentrations from drone imagery (Jia et al., 2021), the estimation of background As concentrations using support vector machines (Wu et al., 2016), the prediction of sustainable As mitigation techniques using Naïve Bayes classifier (Singh et al., 2022), among others.

The assessment of the importance of the input variables for the parameter μ was made using three different metrics, and pointed towards the high importance of edaphoclimatic factors and indicators of human influence. In decreasing order, temperature, clay content, phosphorus content, distance to roads and terrain elevation were found to be the most practically influential features (i.e., with higher *practical importance*). These variables were followed by soil pH, annual precipitation and distance to mines, although the statistical effect of these variables was usually not unanimous across all metrics used (Fig. 4). The linear models for the other distribution parameters indicated an effect of soil pH, phosphorus content and distance to mines on the shape of the distribution. The linear models coefficients suggest that all variables tend to marginally decrease σ, ν and τ, leading to different patterns, as described in Fig. 2.

For the human-related factors, the one-dimensional ALE plots for the model variables (Fig. 5) shows that As concentrations tend to be higher in areas surrounding the existence of mines and roads. As is known to be found in metal ores (McLaren et al., 2006), and the results may be capturing a pattern of As accumulation in the soil as a result of human pollution, for example from the release of dusts and effluents (Thornton and Farago, 1997). Among the edaphoclimatic variables, the reasoning behind the temperature effect on As concentration may reflect its impact on solubilization and sorption rates, as well as the uptake by roots and leaves (Horswell and Speir, 2006). Besides, clay-sized particles include metal (Fe, Al, Mn) (hydr)oxides, which are the most important adsorbents for As in soils (Voegelin et al., 2007). The relationship obtained for the phosphorus content may relate to the fact that this element reacts similarly to As in the soil environment (Adriano, 1986), to the highly complex interactions between their availability in soils (Jing et al., 2022), and to the previous application of agricultural products (Jayasumana et al., 2015).

4.2. As contamination assessment and policy implications

The country-averaged As concentration of Fig. 7 points to the existence of three groups of countries: with lower (< 4 mg kg^{-1}), medium (4 - 7 mg kg^{-1}), and higher (> 7 mg kg^{-1}) As concentrations. The group of low values is geographically clustered, with the spatial distribution of Fig. 6 displaying a clear difference between the As concentrations in Northern Europe and the other regions. These findings visually coincide with previous modeling efforts, such as those by Tóth et al. (2016), Tarvainen et al. (2013) and Rodríguez-Lado et al. (2008). This North-South differentiation between topsoil As concentration has been explained by the natural difference between Southern Europe's older and more fine-textured soils and Northern Europe's younger and more coarse-textured soils (Tarvainen et al., 2013). As noted by Tarvainen et al. (2013), the spatial pattern coincides with the areas covered by glacial ice in the last glacial period. A similar case is observed, for instance, in the concentration of Zn in European topsoils (Van Eynde et al., 2023). Besides, the visualization of the practical importance of model variables (SM11 and SM12) suggests different conditions affecting As concentration across countries. In the Northern countries mentioned above, the status of most variables with high practical importance in the model (e.g., temperature, clay content, and soil pH) leads to predominantly lower As concentrations. On the other hand, different dynamics are observed in the countries with the highest As concentrations. For example, the high As median concentrations in regions of Central France (Fig. 6) are correlated with a particular combination of precipitation and temperature, soil phosphorus content, distance to mines, and terrain elevation. It is worth mentioning, however, that this analysis ignores interactions between variables, which are presented in our model but not in the visualizations of SM11 and SM12.

While the results of the comparison against the background concentration (Fig. 9) indicate that most of the As found may come from human contamination, the comparison against exceedance probabilities (Fig. 8) indicate that most of Europe has a relatively small risk of exceeding 45 mg kg^{-1}. Higher risks are found in France, Austria, Spain and Portugal, as well as smaller contamination areas in Belgium, Germany, Italy and the Czech Republic. Since the highest threshold adopted exceeds the 40 mg kg^{-1} usually used to detect harm for crop plants (Sheppard, 1992), these regions must take extra care for adverse effects that include inhibited metabolic processes and death (Mahimairaja et al., 2005). It must be noted that the thresholds adopted in Fig. 8 are not a consensus, and some regionality exists in the regulations. For instance, in Finland, where As concentrations are generally lower than in other countries (Fig. 7), the threshold for assessing contamination and remediation needs is 5 mg kg^{-1}, and the limit for ecological risks ranges between 50 and 100 mg kg^{-1} (FME, 2007). In Sweden and Denmark, the screening values for residential use are 15 and 20 mg kg^{-1}, respectively, with the second value also being the threshold for Austria (Carlon, 2007). Slovakia, Germany, and the Czech Republic adopt screening values of 30, 50, and 65 mg kg^{-1}, respectively (Carlon, 2007). In a slightly different context, the European Chemicals Agency (ECHA) evaluated the toxicity of As against terrestrial organisms, and the mean values of the 10 % effect concentration (i.e., EC10) ranged from 5.0 to 142.8 mg (kg dry weight of soil)$^{-1}$, depending on the species under consideration (ECHA, 2023). Since EC10 values correspond to the concentrations at which 10 % of the organisms present are significantly negatively affected (Corn, 1993), the range presented is expected to be lower than other commonly used indicators, such as the LC10 and LC50 values (i.e., the concentrations at which 10 and 50 % of the organisms die, respectively). The different references, together with the increased variability when including GEMAS observation in the analyses, suggests a relatively large uncertainty concerning the true risks of As concentration in European soils.

Similarly to the contamination levels, the definition of background concentrations also varies. Beyond the 1.5 mg kg^{-1} adopted, the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation from the ECHA defined the predicted no-effect concentration as 0.7 mg kg^{-1} (Reimann et al., 2018). In Finland, the Ministry of the Environment defines natural concentrations as 1.0 [0.1, 2.5] mg kg^{-1} (FME, 2007), and the background concentration in German soils were calculated to vary spatially from the interval [0, 5] to > 25 mg kg^{-1} (BGR, n.d.). In Sweden, sediment data from the Baltic Sea indicated a median pre-industrial concentration of 12.4 mg kg^{-1}, exceeding the 10 mg kg^{-1} recommended by the National Environmental Protection Agency (Shahabi-Ghahfarokhi et al., 2021). Additionally, natural background concentrations in Poland were reported to vary between 0.8 and 9.1 mg kg^{-1}, 2.76 to 16.0 mg kg^{-1} in the Czech Republic, and equal to 15 mg kg^{-1} in Austria (Sakala et al., 2011). In this sense, the large spatial variation of As concentration across Europe (Fig. 6) led Tarvainen et al. (2013) to state that "it is clearly not possible to define one background value for the whole continent".

Concerning policy developments, the European Commission proposed in 2021 the Zero Pollution Action Plan (ZPAP) to improve soil quality and reduce diffuse contamination, including improvements to air and water quality. The overarching objective of ZPAP is to create a toxic-free environment by reducing soil pollution to levels considered no longer harmful for health and ecosystems. In ZPAP, the goals of better preventing, remedying, monitoring and reporting on soil pollution are pursued by monitoring the current state of diffuse pollution in soils. In this sense, the present work contributes to establishing baselines of pollution by As, therefore aligning with the objectives of the EU Soil Observatory of searching for better uses of the LUCAS soil survey, and promoting modeling assessments to develop baseline maps of metals in the soil environment (Panagos et al., 2022b). In addition, the European Commission recently adopted new rules to increase food safety by reducing the presence of As in food products (i.e., Commission Regulation n^{o} 2023/465 of 3 March 2023). With most of the food coming from soils, this regulation exemplifies how policy measures could benefit from more knowledge of baseline indicators of heavy metal occurrence in European lands. As, along with Hg, Cd and Pb, has a high priority for the dangers it poses (Fuller et al., 2022). This and other legal efforts could help prevent industrial abuse in the application of As-based products.

Furthermore, the proposed Soil Monitoring Law has three main objectives: i) "a solid and coherent monitoring framework for all soils across the EU", ii) "making sustainable soil management the norm in the EU", and iii) "requesting Member States to identify potentially contaminated sites and contributing to a toxic free environment by 2050" (EC, 2023b). In this context, the proposed As map of the European Union is a baseline that contributes to the estimation of diffuse soil contamination. The present work aligns with past efforts to map other elements in soil, such as Cu (Ballabio et al., 2018), Hg (Ballabio et al., 2021), and Zn (Van Eynde et al., 2023), and the development of a high resolution As dataset as well as the investigation of the main natural and anthropogenic variables correlated with increased As concentration contribute to a better understanding of soil contamination in the EU.

5. Conclusions

In this work, a model called GAMLSS-RF was proposed as an alternative to mapping As concentrations in Europe while dealing with data censoring issues that appear in the LUCAS database. GAMLSS-RF allowed modeling highly nonlinear interactions among variables while establishing a coupled model for the left and right parts of data (i.e. below and above the LUCAS detection limit of 2.84 mg kg^{-1}, respectively) in such a way that an additional assumption leads to the reconstruction of the unobserved left tail. Before the fitting procedure, the observations were split into training and validation datasets, and the analysis of residuals showed a consistent performance of the fitted model against all datasets. An interpretation of the statistical importance of model variables showed a reasonable behavior, with edaphoclimatic and human-related variables playing a relevant role in the prediction of

As concentrations.

Compared to other existing approaches to map As at the European level, the present work contains a higher spatial resolution and presents more adequate modeling assumptions, thus advancing towards more realistic spatial representations. The approach also allows the incorporation of observation from external data sources, which helps to understand the uncertainties of the analysis developed. The results indicated a high spatial variability of As concentrations in Europe, and countries such as Portugal, Belgium, Austria, France and Spain present a non-negligible risk of exceeding even the highest limit of action considered in the analysis (i.e., 45 mg kg^{-1}). Results also indicated a high chance of human-related contamination of As in the whole of Europe, but the background concentration adopted is highly uncertain (i.e., 1.5 mg kg^{-1}) and other threshold values could be checked for the whole EU when consolidated. The proposed GAMLSS-RF approach can be adopted by researchers facing similar limitations in other contexts, and the findings presented in this work can help support future assessments of soil health and pollution at a continental level, as well as its ecotoxicological implications.

6. Data availability

The datasets generated are available in the European Soil Data Centre 2.0 (ESDAC) (ESDAC, 2023; Panagos et al., 2022a).

CRediT authorship contribution statement

Arthur Nicolaus Fendrich: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Elise Van Eynde:** Writing – original draft, Methodology, Investigation, Data curation, Conceptualization, Formal analysis. **Dimitrios M. Stasinopoulos:** Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Robert A. Rigby:** Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Felipe Yunta Mezquita:** Writing – original draft, Visualization, Supervision, Investigation, Conceptualization. **Panos Panagos:** Writing – original draft, Visualization, Supervision, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

The data that has been used is confidential.

Acknowledgments

This work was supported by the European Commission Joint Research Centre and the CLAND project under the Collaborative Doctoral Partnership Agreement No. 35403.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2024.108544.

References

Adamse, P., Van Der Fels-Klerx, H.I., De Jong, J., 2017. Cadmium, lead, mercury and arsenic in animal feed and feed materials – trend analysis of monitoring results. Food Additives & Contaminants: Part A 34 (8), 1298–1311. https://doi.org/10.1080/19440049.2017.1300686.

Adriano, D.C., 1986. Trace elements in the terrestrial environment. In Springer eBooks. https://doi.org/10.1007/978-1-4757-1907-9.

Agrawal, A., Petersen, M.R., 2021. Detecting arsenic contamination using satellite imagery and machine learning. Toxics 9 (12), 333. https://doi.org/10.3390/toxics9120333.

Apley, D.W., Zhu, J., 2020. Visualizing the effects of predictor variables in black box supervised learning models. J. R. Stat. Soc. Ser. B Stat Methodol. 82 (4), 1059–1086. https://doi.org/10.1111/rssb.12377.

Aria, M., Cuccurullo, C., Gnasso, A., 2021. A comparison among interpretative proposals for random forests. Machine Learning with Applications 6, 100094. https://doi.org/10.1016/j.mlwa.2021.100094.

Aria, M., Gnasso, A., Iorio, C., Pandolfo, G., 2023. Explainable ensemble trees. Comput. Stat. https://doi.org/10.1007/s00180-022-01312-6.

Ballabio, C., Panagos, P., Montanarella, L., 2016. Mapping topsoil physical properties at european scale using the LUCAS database. Geoderma 261, 110–123. https://doi.org/10.1016/j.geoderma.2015.07.006.

Ballabio, C., Panagos, P., Lugato, E., Huang, J., Orgiazzi, A., Jones, A., Fernández-Ugalde, O., Borrelli, P., Montanarella, L., 2018. Copper distribution in european topsoils: an assessment based on LUCAS soil survey. Sci. Total Environ. 636, 282–298. https://doi.org/10.1016/j.scitotenv.2018.04.268.

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., Panagos, P., 2019. Mapping LUCAS topsoil chemical properties at european scale using gaussian process regression. Geoderma 355, 113912. https://doi.org/10.1016/j.geoderma.2019.113912.

Ballabio, C., Jiskra, M., Osterwalder, S., Borrelli, P., Montanarella, L., Panagos, P., 2021. A spatial assessment of mercury content in the European Union topsoil. Sci. Total Environ. 769, 144755 https://doi.org/10.1016/j.scitotenv.2020.144755.

BGR. Karte der Hintergrundwerte für Arsen. (n.d.). Retrieved 14 October, 2023, from https://www.bgr.bund.de/DE/Themen/Boden/Bilder/Bod_HGW_KarteAs_g.html.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/a:1010933404324.

Carlon, C. (2007). Derivation methods of soil screening values in Europe. A review and evaluation of national procedures towards harmonisation. JRC Scientific and Technical Reports. Retrieved 12 October, 2023, from https://esdac.jrc.ec.europa.eu/content/derivation-methods-soil-screening-values-europe-review-and-evaluation-national-procedures.

Carmen-Ileana, C., Comero, S., Giovanni, L., Isabelle, F., Agustin, A.R., Gergely, T., Bernd, G., 2014. Comparative study on open system digestion vs. microwave-assisted digestion methods for trace element analysis in agricultural soils. JRC publications. Office. https://doi.org/10.2788/79443.

Corn, M., 1993. Handbook of hazardous materials. Elsevier Academic Press. https://doi.org/10.1016/C2009-0-02346-0.

Cristache, C., et al., 2014. Comparative study on open system digestion vs. microwave assisted digestion methods for trace element analysis in agricultural soils. In: EUR 26636 EU. Technical Report – Joint Research Centre.

[EC] European Commission. LUCAS - ESDAC. (2023a). Retrieved September 4, 2023, from https://esdac.jrc.ec.europa.eu/projects/lucas.

EC. Soil health. (2023b, November 10). https://environment.ec.europa.eu/topics/soil-and-land/soil-health_en.

[ECHA] European Chemicals Agency. Registration dossier: Arsenic (2023). Retrieved 14 October, 2023, from https://echa.europa.eu/registration-dossier/-/registered-dossier/22366/6/4/1.

[EEA] European Environment Agency [EEA]. European Digital Elevation Model (EU-DEM). (2016, April). https://www.eea.europa.eu/en/datahub/datahubitem-view/d08852bc-7b5f-4835-a776-08362e2fbf4b#tab-metadata.

Elvidge, C.D., Baugh, K., Zhizhin, M., Hsu, F., Ghosh, T., 2017. VIIRS night-time lights. Int. J. Remote Sens. 38 (21), 5860–5879. https://doi.org/10.1080/01431161.2017.1342050.

European Soil Data Centre (ESDAC). (2023). https://esdac.jrc.ec.europa.eu/.

Fabian, C., Reimann, C., Fabian, K., Birke, M., Baritz, R., Haslinger, E., 2014. GEMAS: spatial distribution of the pH of european agricultural and grazing land soil. Appl. Geochem. 48, 207–216. https://doi.org/10.1016/j.apgeochem.2014.07.017.

Fawagreh, K., Gaber, M.M., Elyan, E., 2014. Random forests: from early developments to recent advancements. Systems Science & Control Engineering 2 (1), 602–609. https://doi.org/10.1080/21642583.2014.956265.

Finnish Ministry of the Environment [FME], 2007. Government decree on the assessment of soil contamination and remediation needs (214/2007, march 1, 2007). Retrieved 22 October, 2023, from. https://www.finlex.fi/en/laki/kaannokset/2007/en20070214.pdf.

Flora, S. J. (2015). Arsenic: Chemistry, Occurrence, and Exposure. In Handbook of Arsenic Toxicology (pp. 1–49). https://doi.org/10.1016/b978-0-12-418688-0.00001-0.

Fuller, R., Landrigan, P.J., Balakrishnan, K., Bathan, G., Blüml, S., Bräuer, M., Caravanos, J., Chiles, T., Cohen, A., Corra, L., Cropper, M., Ferraro, G., Hanna, J.L., Hanrahan, D., Hu, H., Hunter, D., Janata, G., Kupka, R., Lanphear, B.P., Yan, C., 2022. Pollution and health: a progress update. The Lancet Planetary Health 6 (6), e535–e547. https://doi.org/10.1016/s2542-5196(22)00090-0.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning. In Springer Series in Statistics. https://doi.org/10.1007/978-0-387-84858-7.

Helfenstein, A., Mulder, V.L., Heuvelink, G., Okx, J.P., 2022. Tier 4 maps of soil pH at 25 m resolution for the Netherlands. Geoderma 410, 115659. https://doi.org/10.1016/j.geoderma.2021.115659.

Helsel, D.R., 1990. Less than obvious - statistical treatment of data below the detection limit. Environ. Sci. Tech. 24 (12), 1766–1774. https://doi.org/10.1021/es00082a001.

Hooker, G., Mentch, L., Zhou, S., 2021. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. Stat. Comput. 31 (6) https://doi.org/10.1007/s11222-021-10057-z.
Horswell, J. & Speir, T. (2006). Arsenic phytotoxicity. In Naidu, R., Smith, E., Owens, G., Bhattacharya, P. & Nadebaum, P. (Eds.), Managing arsenic in the environment: from soil to human health (pp. 183-208). CSIRO. ISBN 0-643-06868-6.
Jayasumana, C., Fonseka, S., Fernando, P. U. a. I., Jayalath, K., Amarasinghe, M. D., Siribaddana, S., Gunatilake, S., & Paranagama, P. (2015). Phosphate fertilizer is a main source of arsenic in areas affected with chronic kidney disease of unknown etiology in Sri Lanka. SpringerPlus, 4(1). https://doi.org/10.1186/s40064-015-0868-z.
Jia, X., Cao, Y., O'Connor, D., Zhu, J., Tsang, D.C.W., Zou, B., Hou, D., 2021. Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field. Environ. Pollut. 270, 116281 https://doi.org/10.1016/j.envpol.2020.116281.
Jing, W., Liang, J., Björn, L.O., Li, J., Shu, W., Wang, Y., 2022. Phosphorus-arsenic interaction in the 'soil-plant-microbe' system and its influence on arsenic pollution. Sci. Total Environ. 802, 149796 https://doi.org/10.1016/j.scitotenv.2021.149796.
Jones, M.C., Pewsey, A., 2009. Sinh-arcsinh distributions. Biometrika 96 (4), 761–780. https://doi.org/10.1093/biomet/asp053.
Klaassen, C., 2013. Casarett & Doull's toxicology: the basic science of poisons. Eighth Edition, McGraw Hill Professional.
Lopes, C., Quental, L., Oliveira, D., Filipe, A., Pereira, A., 2018. INSPIRE data harmonisation of mineral resources: contribution of MINERALS4EU project armonización de datos de recursos minerales INSPIRE: contribución del proyecto MINERALS4EU. REVISTA MAPPING 27 (187), 56–63. https://www.europe-geology.eu/mineral-resources/.
Lund, U., Fobian, A., 1991. Pollution of two soils by arsenic, chromium and copper. Denmark. Geoderma 49 (1–2), 83–103. https://doi.org/10.1016/0016-7061(91)90093-9.
Mahimairaja, S., Bolan, N., Adriano, D. C., & Robinson, B. (2005). Arsenic Contamination and its Risk Management in Complex Environmental Settings. In Advances in Agronomy (pp. 1–82). https://doi.org/10.1016/s0065-2113(05)86001-8.
Marchant, B.P., Saby, N.P., 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.
McLaren, R. G., Megharaj, M., & Naidu, R. (2006). Fate of arsenic in the soil environment. In Naidu, R., Smith, E., Owens, G., Bhattacharya, P. & Nadebaum, P. (Eds.), Managing arsenic in the environment: from soil to human health (pp. 157–182). CSIRO. ISBN 0-643-06868-6.
Medunić, G., Fiket, Ž., & Ivanić, M. (2019). Arsenic contamination status in Europe, Australia, and other parts of the world. In Springer eBooks (pp. 183–233). https://doi.org/10.1007/978-981-13-8587-2_6.
Murcott, S., 2012. Arsenic contamination in the world: an international sourcebook 2012. IWA Publishing 11. https://doi.org/10.2166/9781780400396.
Noce, S., Caporaso, L., Santini, M., 2020. A new global dataset of bioclimatic indicators. Sci. Data 7 (1). https://doi.org/10.1038/s41597-020-00726-5.
OpenStreetMap. (2018). OpenStreetMap. https://www.openstreetmap.org/.
Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2017. LUCAS soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. 69 (1), 140–153. https://doi.org/10.1111/ejss.12499.
Palma-Lara, I., Martínez-Castillo, M., Quintana-Pérez, J.C., Arellano-Mendoza, M.G., Tamay-Cach, F., Valenzuela, O.L., García-Montalvo, E.A., Hernández-Zavala, A., 2020. Arsenic exposure: a public health problem leading to several cancers. Regul. Toxicol. Pharm. 110, 104539 https://doi.org/10.1016/j.yrtph.2019.104539.
Panagos, P., Van Liedekerke, M., Borrelli, P., Köninger, J., Ballabio, C., Orgiazzi, A., Lugato, E., Liakos, L., Hervás, J., Jones, A., Montanarella, L., 2022a. European soil data Centre 2.0: soil data and knowledge in support of the EU policies. Eur. J. Soil Sci. 73 (6) https://doi.org/10.1111/ejss.13315.
Panagos, P., Montanarella, L., Barbero, M., Schneegans, A., Aguglia, L., Jones, A., 2022b. Soil priorities in the European Union. Geoderma Reg. 29, e00510.
Ratnaike, R. N. (2006). Arsenic in health and disease. In Naidu, R., Smith, E., Owens, G., Bhattacharya, P. & Nadebaum, P. (Eds.), Managing arsenic in the environment: from soil to human health (pp. 288-309). CSIRO. ISBN 0-643-06868-6.
Reimann, C., De Caritat, P., 1998. Chemical elements in the environment. In Springer eBooks. https://doi.org/10.1007/978-3-642-72016-1.
Reimann, C., Matschullat, J., Birke, M., Salminen, R., 2009. Arsenic distribution in the environment: the effects of scale. Appl. Geochem. 24 (7), 1147–1167. https://doi.org/10.1016/j.apgeochem.2009.03.013.
Reimann, C., Fabian, K., Birke, M., Filzmoser, P., Demetriades, A., Négrel, P., Oorts, K., Matschullat, J., De Caritat, P., Albanese, S., Anderson, M.E., Baritz, R., Batista, M.J., Bel-Ian, A., Cicchella, D., De Vivo, B., De Vos, W., Dinelli, E., Đuriš, M., Sadeghi, M., 2018. GEMAS: establishing geochemical background and threshold for 53 chemical elements in european agricultural soil. Appl. Geochem. 88, 302–318. https://doi.org/10.1016/j.apgeochem.2017.01.021.
ResourceWatch (2019). Global Power Plant Database v1.2.0. http://resourcewatch.org/.
Rigby, R.A., Stasinopoulos, D., 2005. Generalized additive models for location, scale and shape. J. R. Stat. Soc. Ser. C. Appl. Stat. 54 (3), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x.
Rigby, R.A., Stasinopoulos, M.D., Heller, G.Z., De Bastiani, F., 2019. Distributions for modeling location, scale, and shape. In Chapman and Hall/CRC eBooks. https://doi.org/10.1201/9780429298547.
Rodríguez-Lado, L., Hengl, T., Reuter, H.I., 2008. Heavy metals in european soils: a geostatistical analysis of the FOREGS geochemical database. Geoderma 148 (2), 189–199. https://doi.org/10.1016/j.geoderma.2008.09.020.
Sakala, J., Vacha, R., Cechmankova, J., 2011. Evaluation of arsenic occurrence in agricultural soils of the bohemian Forest region. Retrieved from Silva Gabreta 17 (2–3), 55–67. https://www.npsumava.cz/wp-content/uploads/2019/06/sg17_2_55_67.pdf.
Salminen, R., Batista, M. J., Bidovec, M., Demetriades, A., De Vivo, B., De Vos, W., Duris, M., Gilucis, A, Gregorauskiene, V., Halamic, J., Heitzmann, P., Lima, A., Jordan, G., Klaver, G., Klein, P., Lis, J., Locutura, J., Marsina, K., Mazreku, A., O'Connor, P. J., Olsson, S.Å., Ottesen, R. T., Petersell, V., Plant, J. A., Reeder, S., Salpeteur, I., Sandström, H., Siewers, U., Steenfelt, A., Tarvainen, T. Geochemical Atlas of Europe. Part 1 - Background Information, Methodology, and Maps. (2005) Retrieved October 31, 2023, from http://weppi.gtk.fi/publ/foregsatlas/article.php?id=5.
Saxe, J. K., Bowers, T. S., & Reid, K. R. (1964). Arsenic. In Elsevier eBooks (pp. 279–292). https://doi.org/10.1016/b978-012507751-4/50035-5.
Shahabi-Ghahfarokhi, S., Åström, M.E., Josefsson, S., Apler, A., Ketzer, M., 2021. Background concentrations and extent of cu, as co, and U contamination in Baltic Sea sediments. J. Sea Res. 176, 102100 https://doi.org/10.1016/j.seares.2021.102100.
Sheppard, S., 1992. Summary of phytotoxic levels of soil arsenic. Water Air Soil Pollut. 64 (3–4), 539–550. https://doi.org/10.1007/bf00483364.
Singh, S.K., Taylor, R.W., Pradhan, B., Shirzadi, A., Pham, B.T., 2022. Predicting sustainable arsenic mitigation using machine learning techniques. Ecotoxicol. Environ. Saf. 232, 113271 https://doi.org/10.1016/j.ecoenv.2022.113271.
Smith, A.H., Lopipero, P., Bates, M., Steinmaus, C., 2002. Arsenic epidemiology and drinking water standards. Science 296 (5576), 2145–2146. https://doi.org/10.1126/science.1072896.
Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V., De Bastiani, F., 2017. Flexible regression and smoothing. In Chapman and Hall/CRC eBooks. https://doi.org/10.1201/b21973.
Stasinopoulos, M.D., Rigby, R.A., De Bastiani, F., 2018. GAMLSS: a distributional regression approach. Stat. Model. 18 (3–4), 248–273. https://doi.org/10.1177/1471082x18759144.
Tarvainen, T., Albanese, S., Birke, M., Poňavič, M., Reimann, C., 2013. Arsenic in agricultural and grazing land soils of Europe. Appl. Geochem. 28, 2–10. https://doi.org/10.1016/j.apgeochem.2012.10.005.
Taylor, S.R., McLennan, S.M., 1995. The geochemical evolution of the continental crust. Rev. Geophys. 33 (2), 241–265. https://doi.org/10.1029/95rg00262.
Thornton, I., & Farago, M. E. (1997). The geochemistry of arsenic. In Springer eBooks (pp. 1–16). https://doi.org/10.1007/978-94-011-5864-0_1.
Tibshirani, J., Athey, S., Sverdrup, E., Wager, S., 2023. grf: generalized random forests. R Package Version 2 (3). https://CRAN.R-project.org/package=grf.
Tóth, G., Hermann, T., Szatmári, G., Pásztor, L., 2016. Maps of heavy metals in the soils of the European Union and proposed priority areas for detailed assessment. Sci. Total Environ. 565, 1054–1062. https://doi.org/10.1016/j.scitotenv.2016.05.115.
[USEPA], 2023a. United States Environmental Protection Agency. Retrieved October 31, 2023, from IRIS. https://iris.epa.gov/AdvancedSearch/.
[USEPA], May 2023 (2023b). Regional screening level (RSL). Retrieved October 31, 2023, from Summary Table. https://semspub.epa.gov/work/HQ/404057.pdf.
[USGS] United States Geological Survey. (2022). Landsat Collection 2 Level-2 Science Products. Landsat-7 image courtesy of the U.S. Geological Survey.
Van Eynde, E., Fendrich, A.N., Ballabio, C., Panagos, P., 2023. Spatial assessment of topsoil zinc concentrations in Europe. Sci. Total Environ. 892, 164512 https://doi.org/10.1016/j.scitotenv.2023.164512.
Veneman, P.L.M., Murray, J., Baker, J.H., 1983. Spatial distribution of pesticide residues in a former apple orchard. J. Environ. Qual. 12 (1), 101–104. https://doi.org/10.2134/jeq1983.00472425001200010017x.
Venteris, E.R., Basta, N.T., Bigham, J.M., Rea, R.G., 2014. Modeling spatial patterns in soil arsenic to estimate natural baseline concentrations. J. Environ. Qual. 43 (3), 936–946. https://doi.org/10.2134/jeq2013.11.0459.
Voegelin, A., Weber, F., Kretzschmar, R., 2007. Distribution and speciation of arsenic around roots in a contaminated riparian floodplain soil: micro-XRF element mapping and EXAFS spectroscopy. Geochim. Cosmochim. Acta 71 (23), 5804–5820. https://doi.org/10.1016/j.gca.2007.05.030.
Williams, J., Kim, H.W., Crespi, C.M., 2020. Modeling observations with a detection limit using a truncated normal distribution with censoring. BMC Med. Res. Method. 20 (1) https://doi.org/10.1186/s12874-020-01032-9.
Wu, J., Teng, Y., Chen, H., Li, J., 2016. Machine-learning models for on-site estimation of background concentrations of arsenic in soils using soil formation factors. J. Soil. Sediment. 16 (6), 1787–1797. https://doi.org/10.1007/s11368-016-1374-9.