Estimating Contaminated Soil Volumes Using a Generative Neural Network: A Hydrocarbon Case in France

Herbert Rakotonirina, Paul Honeine, Olivier Atteia, Antonin van Exem

Estimating Contaminated Soil Volumes Using a Generative Neural Network: A Hydrocarbon Case in France

Herbert RAKOTONIRINA[1,2], Paul HONEINE[1],
Olivier ATTEIA[3], Antonin VAN EXEM[2]

[1] Univ Rouen Normandie, INSA Rouen Normandie, Université Le Havre Normandie, Normandie Univ, LITIS UR 4108, F-76000 Rouen, France

[2] Tellux, 11 Rue du Moulin À Poudre, 76150 Maromme, France

[3] EPOC (UMR 5805), CNRS, Univ. Bordeaux & Bordeaux INP, Bordeaux, France

Published in Mathematical Geosciences, May 2025
https://doi.org/10.1007/s11004-025-10193-6

Abstract

The estimation of the volumes of contaminated soil to be treated is a crucial step in soil remediation. Numerous techniques exist for estimating the distribution of pollutants in soils, such as inverse distance weighting, kriging, Gaussian sequential simulation, and sequential indicator simulation. Unfortunately, these methods require significant computational resources to achieve precise estimations. Moreover, both kriging and Gaussian simulation require the transformation of non-normal distributions, often seen in hydrocarbon contamination, to produce accurate results. In this paper, we propose a generative neural network to generate 3D maps of contaminant distributions without prior training, and to estimate the contaminated volumes. This differentiates this work from other Deep Learning approaches that necessitate training data. The proposed method relies on a convolutional neural network for image reconstruction and inpainting. Rather than solely depending on the concentration of chemicals determined in the laboratory, we utilize hyperspectral imaging data from soil cores to achieve a more precise depiction of soil contaminants. We assess the proposed method using a synthetic 3D dataset and a real case of hydrocarbon pollution on a polluted site in France. The method demonstrates competitive performance with efficiently managed computation time, achieved through the use of GPU accelerator. This study offers a new, practical way to improve soil pollution management using fast, and data-driven techniques.

Keywords: Geostatistics, Volume estimation, Hydrocarbon pollution, Deep learning, Environmental data, Kriging, Geostatistical conditional simulation

1 Introduction

Estimating the volume of contaminated soil is essential for managing polluted sites, as it enables effective planning of decontamination efforts, cost evaluation, resource allocation, and ensures environmental and public health safety. Geostatistical methods are frequently used for this estimation due to their ability to manage spatial data variations and provide uncertainty estimates. Tao et al. (2022) showed in their study that many geostatistical methods were used to delineate pollution zones in 3D, such as ordinary kriging/indicator kriging (Liu et al., 2017; Tao et al., 2014; Ren et al., 2016) or geostatistical simulation (Jiang et al., 2016). They also pointed out major underlying challenges, such as adhering to the stationarity assumption, which implies that the variable is invariant under translation, and the isotropy assumption, which requires the variable to be invariant under rotation. Although classical interpolation methods offer a volume estimate, they do not provide the associated uncertainties. Kriging, a major geostatistical method that predicts unknown values based on spatial correlation, can provide an uncertainty for each individual estimate, but it does not provide a global uncertainty for the total estimated volume (Xie et al., 2011; Agyeman et al., 2022). In recent years, there has been an increasing interest in 3D interpolation from borehole data. This progress is primarily attributed to the development and application of data-driven approaches –more specifically machine learning methods– to this task, as demonstrated by studies such as (Abbaszadeh Shahri et al., 2021; Wang and Gan, 2023; Lialestani et al., 2022), and more recently (Lialestani et al., 2024; Tychola et al., 2024). However, these methods do not automatically account for the uncertainties associated with the estimations. Although a recent study Abbaszadeh Shahri et al. (2024) introduces uncertainty estimation, it is limited to point estimates and does not extend to the estimated volume.

To address these limitations on the uncertainty of estimated volumes, Demougeot-Renard and De Fouquet (2004) proposed a geostatistical methodology based on sequential Gaussian simulations (SGS) introduced by Journel (1974). This approach tackles the problem of the non-integration of spatial correlations in simulation techniques. This simulation is conditional because the values simulated at the data locations match the observed values; it is geostatistical as it incorporates the spatial correlation function, represented by the variogram, into the simulation process. This allows the evaluation of the volume of soil to be treated and the associated uncertainty based on observed data. The steps of this method can be summarized as follows: (*i*) Generation of point conditional simulations of pollutant concentrations on a fine grid, using site survey data as conditional data. (*ii*) Calculation of block conditional simulations of pollutant concentrations on a coarse grid, replicating the grid used for the remediation of soils to be cleaned. A simulated block value is the mean of the sum of the simulated point values included in the remediation grid. (*iii*) Calculation of block conditional simulations of pollutant concentrations, model sampling and analysis errors. (*iv*) Calculation of exceeding probabilities for pollutant concentrations per block. (*v*) Evaluation of the remediation volume and associated

uncertainty based on block simulations and probabilities.

The superiority of the SGS method in estimating volumes of contaminated soil is confirmed by a recent comparative study conducted by Metahni et al. (2019), for the estimation of volumes of soil contaminated by As, Cr, Cu, Pentachlorophenol, and dioxins/furans. That study concluded that geostatistical interpolation methods, such as kriging, are effective in capturing the spatial structures of pollution distribution in soils, unlike deterministic methods such as inverse distance weighting. The SGS method involves performing spatial interpolation that estimates unknown values by averaging the values of nearby points, giving more weight to closer points based on the inverse of their distance. However, the tendency of kriging to produce smoothed results reduces its effectiveness in estimating exceedance probabilities for specific thresholds. Consequently, SGS is preferred for estimating volumes of contaminated soils and quantifying the associated uncertainties.

The estimation of uncertainties related to contaminated soil volumes is crucial, as it helps control the risk of error, especially in contexts where the number of observed values is limited. Moreover, the models usually proposed often struggle to accurately represent reality. This in-depth understanding of uncertainties aids in making more informed and safer decisions regarding the management of contaminated sites. Guridi et al. (2023) suggested combining Deep Learning and geostatistical simulation to significantly improve the accuracy of estimations in the context of complex environmental pollution contexts.

However, geostatistical simulation-based approaches, such as SGS, are constrained by certain methodological requirements, notably the need to transform data into a Gaussian distribution before proceeding with the simulation. This step is crucial for SGS, due to the Gaussian assumption of the investigated model. However, this is not always suitable for pollution data, which often exhibits an asymmetric distribution. Indeed, these data are characterized by an over-representation of values close to zero, and conversely, by extremely high concentration values. This particularity is highlighted by Aghadadashi et al. (2019) in their study, illustrating the challenges associated with applying traditional geostatistical methods to environmental pollution data. In most applications, simulations are performed layer by layer in 2D, due to the challenges associated with modeling the variogram in 3D. Additionally, the high computational cost of 3D modeling often leads researchers to favor the 2D approach. However, this process limits the integration of vertical and horizontal information in the modeling, which can affect the accuracy and relevance of the obtained results.

In this paper, we propose to build on our previous work on spatial interpolation with a generative neural network for 2D maps generation (Rakotonirina et al., 2024). This work is revisited here by proposing a 3D generative convolutional neural network (CNN) that generates multiple 3D maps conditioned by the observed values. And then, these maps are used to estimate the volume of contaminated soil and the associated uncertainty with this estimation. This Deep Learning method frees us from the constraints traditionally associated with geostatistical methods, such as the assumption of Gaussian distribution

data and the constraints linked to the use of variographical analysis like the spatial stationarity (consistent statistical properties across the study area) and the isotropy (directional independence of these properties).

To demonstrate the relevance of the proposed method, its ability to perform 3D spatial interpolation is first assessed using a synthetic dataset. The performance of the proposed method is compared to that of ordinary kriging, using 1% of observed values relative to the interpolated values. Subsequently, a concrete application of volume calculation is presented through a case study of total petroleum hydrocarbon (TPH) contamination in France. This case study consists of a 3D dataset derived from six boreholes, reflecting the limited data availability often encountered in real-world analyses. The dataset is a 3D dataset of soil core samples analyzed using hyperspectral imaging. This analysis, conducted by the startup Tellux, provides the concentrations at every point in the soil cores; this is done by inferring the hyperspectral model trained on ground-truth laboratory chemical analysis (Dhaini et al., 2021; Exem et al., 2023; Feray et al., 2023). To evaluate the performance of the proposed method, its spatial interpolation capability is first analyzed through a comparison with ordinary kriging (OK) and the average maps generated by sequential Gaussian simulation (SGS). Subsequently, the volume estimates obtained from SGS and the proposed method are compared to assess their ability to produce estimates with an uncertainty that aligns with actual values. A validation protocol is implemented by excluding portions of two boreholes, which are then used for evaluation.

The main contributions of this paper can be summarized as follows:

- **Deep Learning method volume estimation with uncertainty:** Our recent work on spatial interpolation in 2D in (Rakotonirina et al., 2024) is revisited here, by proposing a generation conditioned by the observed values in 3D and by adopting the volume uncertainty calculation method introduced by Demougeot-Renard and De Fouquet (2004).

- **3D spatial interpolation with limited data:** The model's ability is demonstrated on performing 3D spatial interpolation with a limited number of boreholes, by recovering both statistical and spatial information by comparing it with OK methods and the mean of a Gaussian simulation.

- **Deep Learning maps generation for variographic analysis:** The maps generated by the proposed method are used for variographic analysis. Our method improves performance compared to variographic analysis with observed data and has also shown that the generated maps present anisotropy.

The rest of this paper is structured as follows. Section 2 introduces the proposed method, that is divided into two subsections: 3D spatial interpolation and volume estimation. The synthetic dataset and the case study are presented in Section 3, and the experimental results are discussed in Section 4. Finally, Section 5 concludes the paper.

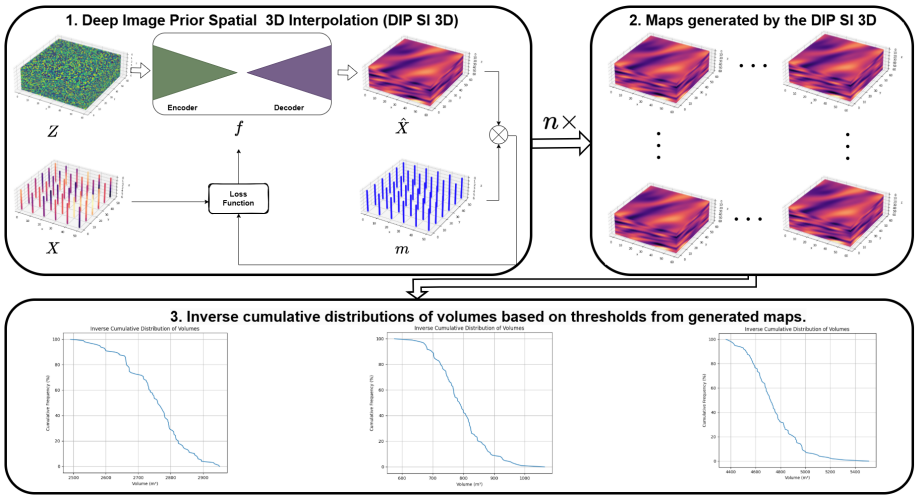

Figure 1: The process of the proposed volume calculation method: (1) Spatial interpolation by minimizing the difference between the interpolated map and observed values; (2) Repeating the interpolation n times with varying random input maps to generate n plausible distributions; (3) Statistical analysis on the resulting maps to estimate the probable volume distribution.

2 Proposed method

This section presents a generative neural network to estimate the volume of contaminated soil and the associated uncertainty. This is addressed in two parts: the first part involves traditional three-dimensional spatial interpolation DIP-SI-3D (Deep Image Prior Spatial Interpolation), and the second part focuses on generating probable maps of pollution distribution in the soil and estimating the volumes from these maps.

2.1 Deep Learning 3D spatial interpolation

In this section, the workflow of 3D spatial interpolation with DIP-SI is described in detail.

2.1.1 DIP-SI-3D

The proposed 3D spatial interpolation method is an adaptation of our 2D method (Rakotonirina et al., 2024), which relies on the generative framework of deep image prior. The operating principle remains similar between 3D and 2D interpolations. As illustrated in Figure 1, the input of the neural network f is a random 3D map Z, which is of the same dimension as the interpolated map $\hat{X}$. The map of observed values X and the binary map m are also in 3 dimensions

to account for both the vertical and horizontal information of the variable of interest. In the binary mask m, the locations corresponding to the observed values in X are equal to 1, while the locations for the values to be interpolated are equal to 0. This implies that the function f is approximated by minimizing the difference between the observed values of the map X and the sampled values of the interpolated map $\hat{X}$. This sampling results from the term-by-term multiplication of the interpolated map with the binary mask, allowing only the locations of the observed data to be considered in the estimation of the error, as shown in step 1 of Figure 1. The loss function $\mathcal{L}$, which adjusts the function f, is the squared error between the map of observed values X and the interpolated map $\hat{X}$ multiplied by the binary mask m, $\mathcal{L}$ can be expressed as follows:

$$\mathcal{L} = \|(\hat{X} - X) \odot m\|^2. \tag{1}$$

Here, the operator $\odot$ designates element-wise product (or Hadamard product), and the norm is a matrix norm.

The 3D method proposed in this paper differs from the 2D method in many aspects. The matrices $X, \hat{X}$ and m are no longer 2D matrices, but 3D cubes, namely tensors. Therefore, this requires a 3D architecture for the neural network f in both its input and its output, as described next. Moreover, special care must be paid to reduce the computational complexity.

2.1.2 3D architecture of the neural network f

In the following section, the specificity of the 3D architecture is presented by providing a detailed description of its two components: the encoder and the decoder.

Encoder: The architecture starts with a sequence of 3D convolutions, using convolution kernels of size 3x3x3 with reflective padding to maintain the spatial dimensions of the data. These layers gradually increase the number of channels, from 32 to 256, enabling hierarchical feature extraction from the raw data. The ReLU activation function is employed at each layer to introduce non-linearity, essential for capturing complex patterns in the data and where the relationships between vertical and horizontal information are non-linear. The dimension of the data is reduced using the average pooling operation, which helps to concentrate the information while reducing the computational complexity. After dimension reduction, the data is flattened and passed through a linear layer to produce the latent representation. In all the experiments carried out in 3D, the latent vector has a size of $s = 10$, which allows to control the computation time.

Decoder: The decoding process begins with transforming the latent vector into a 3D tensor, which is then progressively enlarged through upsampling to return to the original dimensions. The decoding uses skip connections, where feature maps of the same dimensions between the Encoder and Decoder are concatenated before each subsequent convolution, facilitating the recovery of detailed spatial features lost during encoding. The specificity of the 3D decoder also lies in the upsampling method used. Unlike the DIP-SI approach in 2D, which uses bicubic interpolation, this technique is not applicable in 3D.

Table 1: Detailed architecture of f for 3D spatial interpolation, where t corresponds to the size of the flattened vector before and after the encoding vector of size s ($s = 10$ in experiments).

Layer	**Type**	**Parameters**
Conv1	Conv3D	Input: 1, Output: 32, Kernel: (3,3,3)
Conv2	Conv3D	Input: 32, Output: 64, Kernel: (3,3,3)
Pool1	AvgPool3D	Kernel: (2,2,2)
Conv3	Conv3D	Input: 64, Output: 128, Kernel: (3,3,3)
Conv4	Conv3D	Input: 128, Output: 128, Kernel: (3,3,3)
Pool2	AvgPool3D	Kernel: (2,2,2)
Conv5	Conv3D	Input: 128, Output: 256, Kernel: (3,3,3)
Conv6	Conv3D	Input: 256, Output: 256, Kernel: (3,3,3)
Pool3	AvgPool3D	Kernel: (2,2,2)
Linear1	Linear	Input: t, Output: s
Linear2	Linear	Input: s, Output: t
Conv7	Conv3D	Input: 512, Output: 256, Kernel: (3,3,3)
Conv8	Conv3D	Input: 512, Output: 256, Kernel: (3,3,3)
Upsample1	Interpolate	Scale factor: 2, Mode: nearest
Conv9	Conv3D	Input: 384, Output: 128, Kernel: (3,3,3)
Upsample2	Interpolate	Scale factor: 2, Mode: nearest
Conv10	Conv3D	Input: 256, Output: 128, Kernel: (3,3,3)
Conv11	Conv3D	Input: 128, Output: 64, Kernel: (3,3,3)
Conv12	Conv3D	Input: 64, Output: 32, Kernel: (3,3,3)
Conv13	Conv3D	Input: 32, Output: 1, Kernel: (3,3,3)

While tricubic interpolation could be considered, it would be extremely resource-intensive. Therefore, we have chosen to use the nearest neighbor method, which provides relevant results. Finally, the network output is produced by the last convolutional layer, thus completing the reconstruction process of the architecture of f.

Table 1 details the hyperparameters of each convolution layer.The parameters are optimized taking into account the available computing capacity and the convergence of the model. As shown in our previous work Rakotonirina et al. (2024), the size of the encoding vector s and the upsampling function both affect the performance of the model.

2.2 Deep Learning volume estimation

After introducing the DIP-SI architecture for 3D estimation, this section focuses on volume estimation. Specifically, the ability of the method presented in Section 2.1 to generate multiple interpolated maps $\hat{X}$ conditioned on the observed values in X is demonstrated.

2.2.1 Conditional generation with DIP-SI-3D (DIP-GEN-3D)

DIP-GEN-3D involves utilizing the DIP-SI-3D spatial interpolation method to create multiple probable maps of the variable of interest. The results of $\hat{X}$ interpolation with DIP-SI-3D vary for two full learning processes. The variability in the results is primarily due to different instances of the generating function f, which performs convolutions on Z using weights ω. These weights are iteratively updated using the Adam optimizer until the loss function defined in Equation (1) reaches its minimum. This loss function is designed to be flexible, permitting a wide range of admissible solutions. As a result, multiple configurations of f can achieve optimized outcomes. This flexibility is essential for modeling systems where uncertainty or variability in the data plays a significant role.

This variation allows, not only to estimate the volume of contaminated soil, but also to quantify the uncertainty associated with this estimation. The measured uncertainty reflects the variability of the results obtained through different initializations and configurations of the input map Z and the weights ω. This approach provides a more robust and reliable estimate, considering the range of possible outcomes rather than a single optimistic or pessimistic result. In summary, both the use of randomness in the generation of Z and the initialization of convolution weights ω enrich the analysis by allowing a more comprehensive exploration of possible scenarios and their implications.

2.2.2 Volume calculation

During the generation step, n different maps are produced. These maps are crucial for the analysis as they represent potential spatial variations of the studied phenomenon, taking into account uncertainties and natural variations. To process and analyze these maps, we have chosen to adopt the methodology described by Demougeot-Renard and De Fouquet (2004). However, instead of using SGS for the generation, the DIP-SI-3D is used as described in the following.

A set of n maps representing the spatial distributions of pollution is generated, denoted as $\{\hat{X}_1, \hat{X}_2, ..., \hat{X}_n\}$. Each of these maps is used to estimate the volume of contaminated soils based on a contamination threshold specified by environmental authorities. This threshold defines the minimum and maximum contamination limits for classifying the soil, namely S_{min} and S_{max}. The contaminated volume V for each map in the set $\{\hat{X}_1, \hat{X}_2, ..., \hat{X}_n\}$ is computed by counting the number of cells (i, j, l) that satisfy the condition relative to the threshold. Each cell contributes to the overall volume estimate based on its spatial resolution, given by res_i, res_j, and res_l, which are the dimensions of each cell along the abscissa axis i, the ordinate axis j, and depth axes, respectively. This is expressed by the following formula for each map $\hat{X}_k$:

$$V = \sum_{i,j,l} P(\hat{X}_k(i,j,l); S_{min}; S_{max}) \times V_{grid} \qquad \forall k \in \{1, 2, \dots n\}, \tag{2}$$

The function $P(\hat{X}_k(i,j,l); S_{min}; S_{max})$ is defined as an indicator function that takes the value 1 if the value of the contaminated soil at point (i, j, l) meets the

specified threshold conditions, that is $S_{min} \leq \hat{X}(i,j,l) < S_{max}$, and 0 otherwise. V_{grid} represents the volume of an individual grid cell, computed as the product of the cell dimensions along the three axes res_i, res_j, res_l.

Following these calculations, a set of estimated volumes for each map is obtained, denoted $\{V_{\hat{X}_1}, V_{\hat{X}_2}, ..., V_{\hat{X}_n}\}$. We leverage this set to analyze the cumulative distribution of the estimated volumes, which provides a clear view of the statistical distribution of the data. To ensure a robust estimation of the contaminated soil volume, in line with regulatory requirements, the percentiles are used as key statistical indicators. Specifically, the percentiles of the inverse cumulative distribution are computed. These values delineate the range within which we can be reasonably confident that the contaminated soil volume lies. The choice of percentiles as references for the confidence interval is strategic. It allows us to cover a broad range of potential scenarios, thereby minimizing the risk of underestimating or overestimating the contamination volumes. This is crucial for planning remediation measures, managing environmental risks, and ensuring transparent communication with regulatory authorities and the public.

3 Study case and experiments

To demonstrate the capacity of the method DIP-SI-3D, it is first evaluated on synthetic anisotropic data, and then we present a concrete application of volume calculation in a case study of total petroleum hydrocarbon (TPH) contamination in France.

3.1 Synthetic dataset

The first dataset is a synthetic dataset generated using the Python library GSTools (Müller et al., 2022). This dataset will be specifically designed to meet the conditions required for optimal kriging performance, with a distribution that does not require transformation. The creation of the dataset can be summarized in three key steps: (i) A 3D grid of dimensions $60 \times 60 \times 60$ is established, defining the spatial coordinates where the random field will be evaluated. (ii) A three-dimensional Gaussian model is instantiated with a variance of 30 and length scales of 16, 8, and 4 for the axes of abscissa, ordinate, and depth, respectively. The angles $(0.8, 0.4, 0.2)$ are used to specify the orientation of the field in the three-dimensional space. The length scale corresponds to the range of the variogram in the direction of the chosen angle. This is a generation method based on the study by Heße et al. (2014). (*iii*) A spatially structured random field is generated using the Gaussian model. This field has a mean of 50 and is initialized with a fixed random seed (`seed = 1`) to ensure the reproducibility of the results.

The sampling strategy to select the observed values involves choosing 6 points spaced 10 units apart starting from 0 along the x and y axes, while all values of z are selected. This approach aims to demonstrate the effectiveness of our method in handling data from soil core sampling. Figure 2 shows that

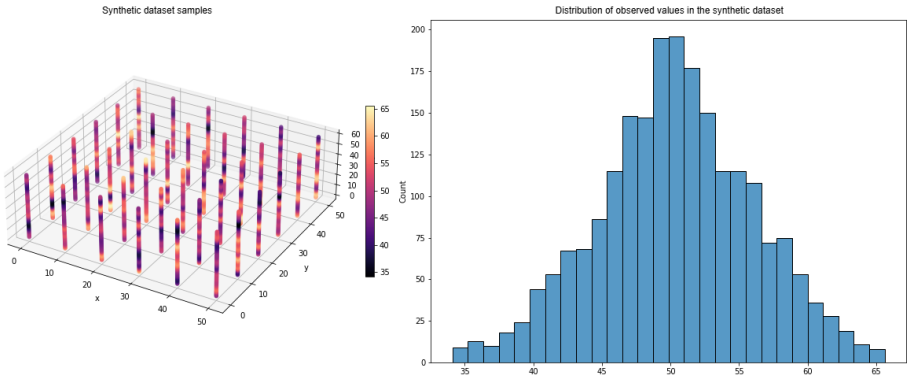

Figure 2: Representation of the generated synthetic data map and its distribution.

the distribution of the selected values closely resembles a Gaussian distribution, which is expected since the dataset was generated using a Gaussian model.

3.2 Case Study: TPH contaminated site

The studied dataset comes from a site contaminated by TPH located in France. The goal of this study is to conduct a soil pollution assessment in order to reuse the land. Estimating the volume of contaminated soil on the site allows for better anticipation of the waste storage management. To achieve this, six boreholes (B1-B6), each with a depth of 5 meters, were drilled. The interpolation grid covers an area of $2612m^2$, with a resolution of 2 m for the horizontal axes and 0.1 m for the vertical depth axis. This dataset is particularly specific due to the limited number of boreholes, resulting in sparse data along the horizontal axes. With a resolution of 2 m on the horizontal axes, each grid cell covers $4m^2$, resulting in $2612/4 = 653$ points to interpolate. Therefore, we have a ratio of 0.91% of points relative to the 6 boreholes. The performance of geostatistical methods may be affected by this limitation, especially due to the constraints of variographic analysis, which heavily depends on the variation of values based on the distances between points.

The pollution contamination in TPH comes from the analysis of hyperspectral images performed by Tellux. Tellux offers machine learning algorithms that correlate indices derived from hyperspectral imaging, as demonstrated by Achard and Elin (2019) and Kühn et al. (2004), with TPH concentrations obtained from laboratory chemical analyses. More details on the investigated methods are given in (Dhaini et al., 2021; Exem et al., 2023; Feray et al., 2023). This provides us with TPH concentrations at all points along the depth axis of the boreholes.

Figure 3 illustrates the locations of the six boreholes as well as the total hydrocarbon concentrations measured along each well. The objective is to use

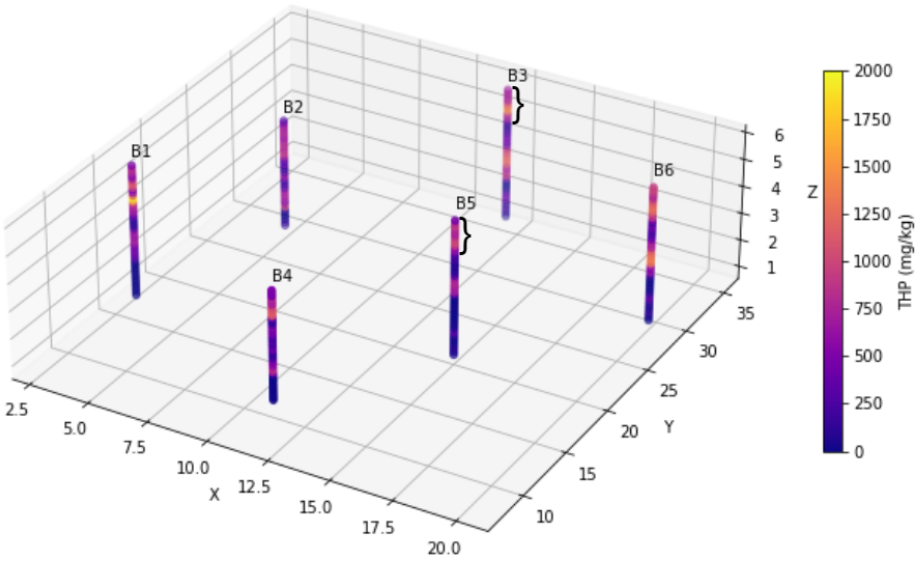

Figure 3: Representation of total hydrocarbon concentrations in the space for the 6 boreholes with the selection of evaluation data for total hydrocarbons indicated by black markers for B3 and B5.

these concentrations to generate multiple possible maps of the pollution distribution. As mentioned in the previous section, due to the lack of more precise elevation data, we used the digital elevation model (DEM) provided by the French Institute for Geographic Information (IGN). The DEM accuracy varies according to the zones and stakes. It reaches 20 centimeters of Root Mean Square Error (RMSE) in flood-prone areas or coastal zones to meet the requirements of the European Flood Directive. Pichon et al. (2016) estimated the average error of the IGN DEM to be around 0.5 m. The wells are thus adjusted according to the altitude of the borehole point.

To validate the model in the absence of evaluation volume data, we study its interpolation capability and compare it with ordinary kriging (OK) and the average of a conditional SGS. To this end, some of the data are excluded from boreholes B3 and B5 as illustrated in Figure 3, which could not be obtained by vertical interpolation of the cores, in order to use them to validate the model. These points, excluded during the training phase, can also be used to evaluate the estimation of the number of contaminated cells in the grid.

The distribution of the dataset is typical of hydrocarbon contamination cases as shown in Figure 4, where a pronounced asymmetrical distribution towards low concentration values is often observed. This characteristic is consistent with observations reported in the literature on similar pollution cases, as shown by studies conducted by Liu et al. (2013) and Liu et al. (2010). These studies highlight that hydrocarbon concentrations in the environment tend to have pre-

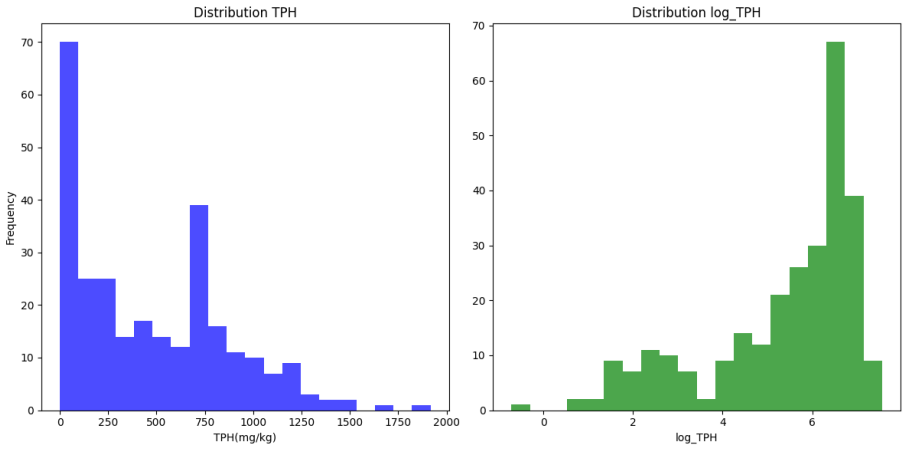

Figure 4: Representation of the distributions of the TPH variable (left figure) and its logarithmic transformation (right figure).

dominantly low values with a few occurrences of extremely high values, creating this highly skewed distribution.

To adequately compare the results obtained by different methods, a log-normal transformation is performed on the TPH values. This transformation involves applying the natural logarithm to the TPH values before performing spatial interpolation using ordinary kriging. After the interpolation, an inverse transformation by exponentiation is applied to return to the original TPH values as shown in the study by Kishné et al. (2003). The concentration values are defined for all the wells. The presence of only six boreholes limits the ability to model three variograms in three directions. Therefore, we opted for a single omnidirectional variogram to ensure consistency.

For performing SGS, the first step is to transform the variables into a Gaussian distribution. However, as shown in Figure 4, even after transformation, we are far from achieving a Gaussian distribution. This observation explains why the results from SGS are not consistent with those obtained using ordinary kriging, even when using the same variogram. Several types of transformations were explored to make the data Gaussian, such as Gaussian anamorphosis performed with the RGeostats tool (MINES ParisTech / ARMINES, 2023) and the Box-Cox transformation. However, the results obtained with the validation data from boreholes $B3$ and $B5$ are inconclusive.

The limited number of borehole points also complicates the modeling of the variogram. Therefore, we propose using data from simulations with DIP-GEN-3D to model the variograms in the three directions needed for SGS. Specifically, the variograms were modeled from the average of the realizations obtained with the DIP-GEN-3D method and then used them in SGS. Figure 5 illustrates the three variogram models thus obtained.

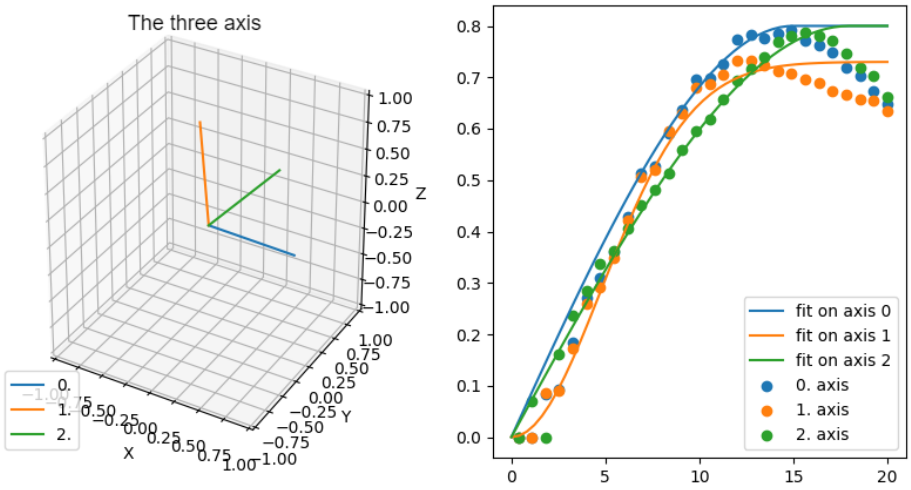

Figure 5: Representation of variograms obtained from the mean maps of the proposed method: the figure on the right displays the directions of the three axis, while the figure on the left presents the variograms for each axis.

4 Experimental results

In the following, the results obtained by the proposed method are compared with geostatistical methods.

4.1 Synthetic dataset

Based on the observed data described in Figure 2, the spatial interpolation is performed using ordinary kriging and the proposed method. For kriging, we used the same directions as those employed in the dataset creation to conduct the variographic analysis. It should be noted that directional information for variographic analysis is not directly available in practical applications but can be obtained through expertise and prior knowledge of the study site. The 3D kriging is performed using the Python library PyKrige, which is a dedicated toolkit for kriging. Developed by Murphy (2014), it provides a wide range of features related to geostatistical methods.

Three metrics are used to compare the results: The mean absolute error (MAE), defined as

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{X}_i - X_i \right|, \tag{3}$$

Table 2: Performance measurements between OK and the proposed DIP-SI-3D method.

Methods	**RMSE**	**MAE**	R^2
OK	3.11	2.28	0.69
DIP-GEN-3D	**2.53**	**1.76**	**0.80**

the root mean square error (RMSE), defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{X}_i - X_i\right)^2}, \tag{4}$$

and the coefficient of determination (R^2), defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{X}_i - X_i)^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}. \tag{5}$$

The results obtained from ordinary kriging are presented in Figure 6 and Table 2. This illustration consists of three columns: the first column shows the ground truth, which is the synthetic dataset generated for this experiment. The second column illustrates the results obtained using the proposed DIP-SI-3D method. The third column displays the results from ordinary kriging. Comparative metrics between the two methods, ordinary kriging and DIP-SI-3D, indicate that the latter outperforms ordinary kriging.

By comparing the maps in the first row, one can observe that the DIP-SI-3D method more accurately reproduces the ground truth map, while ordinary kriging tends to smooth the results. This difference is particularly evident in areas where real values are known and replace interpolated values, creating a visible discontinuity on the map. The three distributions presented in the second row are similar, which is consistent with the fact that the data follows a Gaussian distribution. This similarity also reflects the typical results obtained with ordinary kriging.

4.2 Case Study: TPH contaminated site

The initial analysis of the results involves evaluating the spatial interpolation by comparing the results obtained with OK, the mean of SGS, and the mean of simulations obtained using the proposed method. These comparisons assess the model's ability to capture spatial and statistical information of the variable of interest. Table 3 shows that the proposed method outperforms the other two geostatistical approaches on all three metrics evaluated: mean absolute error (MAE), root mean square error (RMSE), and R^2 coefficient score. It is also important to note the difference between the results of OK and the mean of SGS. Theoretically, the results of these two methods should converge. However,

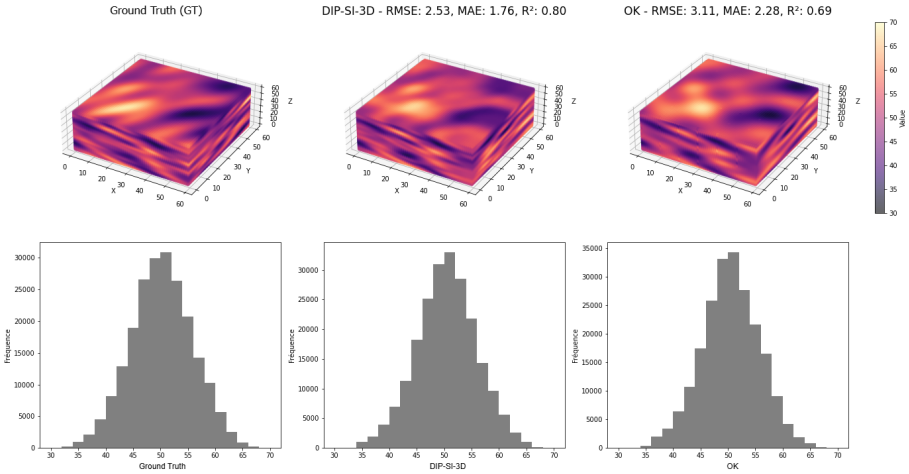

Figure 6: Illustration of ground truth (first column) and the 3D interpolation results with the proposed DIP-SI-3D method (second column) and OK (last column).

in the study, a variogram derived from the mean of the maps generated by the proposed method was used for SGS, whereas for OK, the variogram was based solely on the observed data. This divergence highlights the capability of the model to effectively integrate vertical and horizontal information from very limited data. Therefore, using the proposed generation method to augment the amount of data could be a viable solution to overcome variogram modeling issues with restricted data.

After evaluating the performance of various spatial interpolation methods, their ability to estimate volumes was assessed. Volume estimates are based on thresholds defined by the authorities in France for hydrocarbons. In this case, the maximum measured concentration is 2000 mg/kg, leading to two classes: the class between 0 and 500 mg/kg and the class between 500 and 2000 mg/kg. Given the absence of real data on the volumes for the two classes, the number of cells in the test data from the B3 and B5 boreholes is used. Specifically, the actual values available for these data are classified and compared with the estimates produced by SGS and the proposed method. Figure 7 and Table 4 show that the inverse cumulative distribution of the estimated numbers of contaminated cells is consistent with the actual number for the proposed method, while the SGS tends to overestimate the concentrations. For this case, the P10 and P90 percentiles are used, providing us with an 80% confidence interval. However, this choice is not universally applicable and depends on the statistical distribution of the estimated volumes and the required confidence level for the case studied. Although the data remain limited, they provide insight into our method's ability to estimate contamination volumes.

After evaluating the model's results using the test data, we present the

Table 3: Performance measures for different methods: OK, SGS (mean), the proposed method

Methods	**RMSE**	**MAE**	R^2
OK	321.44	261.92	0.35
SGS (mean)	320.10	258.79	0.34
This paper	**233.53**	**163.40**	**0.58**

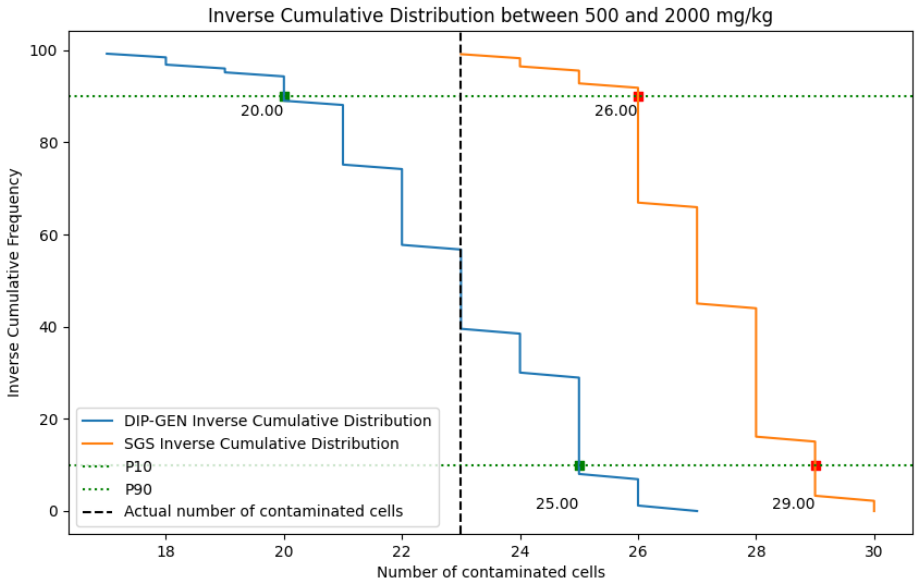

Figure 7: Representation of the distribution of estimated volumes for the two thresholds for the validation data from proposed method (DIP-GEN) and SGS

volume estimations for each threshold across the entire grid using Equation (2). Figure 8 shows the inverse cumulative distribution for 2 classes of contaminated soil volume. In addition to volume estimation, the simulations allow us to obtain the mean map $\hat{X}_{\text{avg}}$, the uncertainty map u, and the probability p of exceeding a certain threshold, respectively defined by

$$\hat{X}_{\text{avg}}(i,j,l) = \frac{1}{n}\sum_{k=1}^{n} \hat{X}_k(i,j,l), \tag{6}$$

$$u(i,j,l) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\hat{X}_k(i,j,l) - \hat{X}_{\text{avg}}(i,j,l)\right)^2}, \tag{7}$$

Table 4: Estimated number of contaminated cells from test data for the two classes of pollution (mg/kg), and compared to ground truth values

	Classes of contamination (mg/kg)	
	[0-500]	**[500-2000]**
Groundtruth (V_{real})	10 cells	23 cells
SGS	4 to 7 cells	26 to 29 cells
This paper	**8 to 13 cells**	**20 to 25 cells**

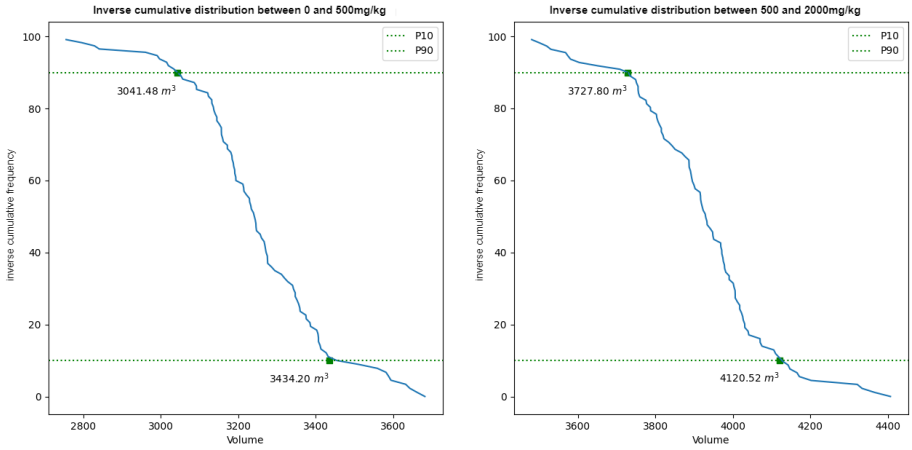

Figure 8: Cumulative inverse distribution of volume estimates for the two classes, 0-500 mg/kg and 500-2000 mg/kg

and, for all $k \in \{1, 2, \dots n\}$,

$$p(i,j,l) = \frac{\sum_{i,j,l} P(\hat{X}_k(i,j,l); S_{min}; S_{max})}{n}. \tag{8}$$

Figure 9 illustrates the model's ability to generate multiple different maps from the observed values. This capability is further confirmed by the uncertainty map. It is noted that the areas where test data were removed exhibit very high uncertainty, which is consistent with the absence of data. However, this also raises an important question: could removing known values during training bias the uncertainty map? The uncertainty map helps identify regions where estimates are uncertain, which is crucial for assessing environmental risks and planning interventions. These maps are essential for optimizing resource allocation during remediation efforts and for clearly communicating risks to stakeholders and the public, thereby facilitating a more informed and responsive management of the situation.

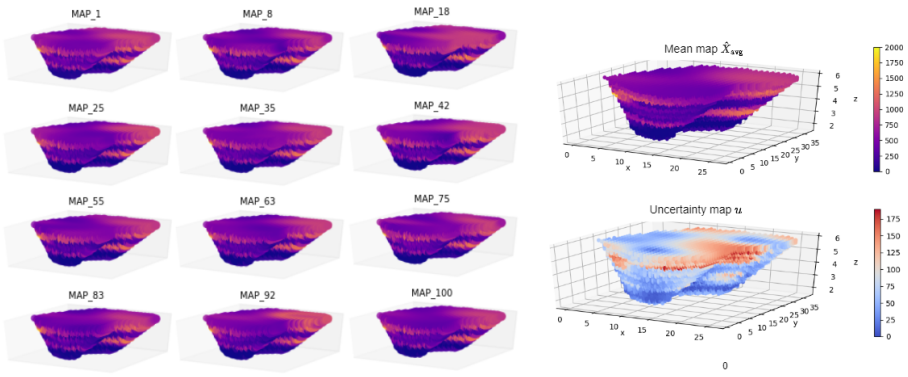

Figure 9: Illustration of the 3D simulation results using the proposed DIP-GEN-3D method. The 12 figures on the left show 12 maps randomly selected from the 100 generated maps. The 2 figures on the right are the map of estimated means (upper figure) and the associated uncertainty map (lower figure).

5 Conclusion

In this paper, a generative neural network was proposed for 3D map generation and contaminated soil volume estimation. The learning process relied on the generation of maps from a random 3D map, adjusted with the observed values. The challenges of 3D generation, compared to 2D, required addressing several issues to develop a computationally feasible method.

The model's ability to perform spatial interpolation in 3D is demonstrated, with better performance than both ordinary kriging and mean of geostatistical simulation. Geostatistical methods exhibit difficulty because of the very limited amount of observed data, which makes variographic modeling difficult. This is accentuated by the fact that the vertical resolution is better than the horizontal resolution. The proposed method is advantageous as it eliminates the need for any prior variographic analysis.

The proposed 3D method was also capable of generating maps conditioned on observed values. This capability was utilized to estimate the volume of contaminated soil. Such estimations are crucial for remediation professionals, as they provide a basis for assessing cleanup costs and managing land post-remediation. The effectiveness of the proposed method was assessed through its application to a case of hydrocarbon contamination in France. The dataset, limited to only six wells and exhibiting an asymmetric distribution, represented the challenges often encountered in polluted sites. Data from two wells, which could not be obtained through vertical interpolation, were excluded to be used as evaluation data. More accurate volume estimates were provided by the proposed method compared to the state-of-the-art method using SGS. Furthermore, it was demonstrated that the results from the proposed method can be used to enhance variogram modeling, thus facilitating the implementation of SGS. This

could represent a practical application of the method in cases where SGS is already endorsed by environmental authorities for volume estimation.

The ability of the proposed method to perform spatial interpolation and conditional map generation has been demonstrated in this paper. However, unlike traditional geostatistical methods, multivariate cases are not handled by the proposed method. This represents a limitation, particularly in applications for contaminated sites and soils where data acquisition costs are very high. In such cases, the incorporation of auxiliary variables could be highly beneficial. This research direction is identified as a major area for future work.

6 Acknowledgment

The authors would like to thank the agency of ecological transition ADEME in France and TELLUX company for the funding of this research work. We would like to express our sincere gratitude to the anonymous reviewers for their valuable feedback and constructive comments, which contributed significantly to enhancing the quality and clarity of this manuscript.

7 Conflict of interest

The authors declare that they have no conflict of interest.

References

Abbaszadeh Shahri A, Chunling S, Larsson S (2024) A hybrid ensemble-based automated deep learning approach to generate 3d geo-models and uncertainty analysis. Engineering with Computers 40(3):1501–1516, DOI: 10.1007/s00366-023-01852-5

Abbaszadeh Shahri A, Kheiri A, Hamzeh A (2021) Subsurface topographic modeling using geospatial and data driven algorithm. ISPRS International Journal of Geo-Information 10(5), DOI: 10.3390/ijgi10050341, URL: https://www.mdpi.com/2220-9964/10/5/341, ISSN 2220-9964

Achard V, Elin C (2019) Automatic mapping of hydrocarbon pollution based on hyperspectral imaging. In IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 5768–5771, DOI: 10.1109/IGARSS.2019.8898455

Aghadadashi V, Molaei S, Mehdinia A, Mohammadi J, Moeinaddini M, Bakhtiari A R (2019) Using gis, geostatistics and fuzzy logic to study spatial structure of sedimentary total pahs and potential eco-risks; an eastern persian gulf case study. Marine pollution bulletin 149:110489, DOI: 10.1016/j.marpolbul.2019.110489

Agyeman P C, Kingsley J, Kebonye N M, Ofori S, Boruuvka L, Vasat R, Kovarek M (2022) Ecological risk source distribution, uncertainty analysis, and application of geographically weighted regression cokriging for prediction of potentially toxic elements in agricultural soils. Process Safety and Environmental Protection 164:729–746, DOI: 10.1016/j.psep.2022.06.051

Demougeot-Renard H, De Fouquet C (2004) Geostatistical approach for assessing soil volumes requiring remediation: Validation using lead-polluted soils underlying a former smelting works. Environmental science & technology 38(19):5120–5126, DOI: 10.1021/es0351084

Dhaini M, Roudaut F J, Garret A, Arzur R, Chereau A, Varenne F, Honeine P, Mignot M, Exem A V (2021) Hyperspectral imaging for the evaluation of lithology and the monitoring of hydrocarbons in environmental samples. In RemTech (International event on Remediation, Coasts, Floods, Climate, Seismic, Regeneration Industry), Ferrara, Italy

Exem A V, Kassem P, Honeine P, Mignot M (2023) High-resolution characterization of total hydrocarbons by infrared hyperspectral imaging in an alluvial soil. In NICOLE Fall Workshop 2023 (Innovative solutions for sustainable redevelopment and land stewardship of contaminated sites and sediments), Malmö, Sweden

Feray C, Jacquemoud S, Honeine P, Exem A V (2023) Hyperspectral characterization of soil matrix effects by coupling physical models and machine learning methods. Poster at the 13th IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS), Athens, Greece

Guridi I, Chassagne R, Pryet A, Atteia O (2023) Uncertainty quantification of contaminated soil volume with deep neural networks and predictive models. Environmental Modeling & Assessment :1–20 DOI: 10.1007/s10666-023-09924-y

Heße F, Prykhodko V, Schlüter S, Attinger S (2014) Generating random fields with a truncated power-law variogram: A comparison of several numerical methods. Environmental Modelling & Software 55:32–48, DOI: 10.1016/j.envsoft.2014.01.013

Jiang S, Wang J, Zhai Y, Yin Z, Teng Y (2016) Determination of the volume of soil requiring remediation in contaminated sites based on conditional simulation. Acta Scientiae Circumstantiae 36(7):2596–2604

Journel A G (1974) Geostatistics for conditional simulation of ore bodies. Economic Geology 69(5):673–687, DOI: 10.2113/gsecongeo.69.5.673

Kishné A S, Bringmark E, Bringmark L, Alriksson A (2003) Comparison of ordinary and lognormal kriging on skewed data of total cadmium in forest soils of sweden. Environmental monitoring and assessment 84:243–263, DOI: 10.1023/A:1023326314184

Kühn F, Oppermann K, Hörig B (2004) Hydrocarbon index–an algorithm for hyperspectral detection of hydrocarbons. International Journal of Remote Sensing 25(12):2467–2473, DOI: 10.1080/01431160310001642287

Lialestani S P M, Parcerisa D, Himi M, Shahri A A (2024) A novel modified bat algorithm to improve the spatial geothermal mapping using discrete geodata in catalonia-spain. Modeling Earth Systems and Environment 10:4415–4428, DOI: 10.1007/s40808-024-01992-7, URL: https://link.springer.com/article/10.1007/s40808-024-01992-7

Lialestani S P M, Shahri A A, Parcerisa D, Himi M (2022) Generating 3d geothermal maps in catalonia, spain using a hybrid adaptive multitask deep learning procedure. Energies 15(13):4602, DOI: 10.3390/en15134602, URL: https://www.mdpi.com/1996-1073/15/13/4602

Liu G, Bi R, Wang S, Li F, Guo G (2013) The use of spatial autocorrelation analysis to identify pahs pollution hotspots at an industrially contaminated site. Environmental monitoring and assessment 185:9549–9558, DOI: 10.1007/s10661-013-3272-6

Liu G, Niu J, Guo W, Zhao L, Zhang C, Wang M, Zhang Z, Guo G (2017) Assessment of terrain factors on the pattern and extent of soil contamination surrounding a chemical industry in chongqing, southwest china. Catena 156:237–243, DOI: 10.1016/j.catena.2017.04.005

Liu Y, Chen L, Zhao J, Wei Y, Pan Z, Meng X Z, Huang Q, Li W (2010) Polycyclic aromatic hydrocarbons in the surface soil of shanghai, china: concentrations, distribution and sources. Organic Geochemistry 41(4):355–362, DOI: 10.1016/j.orggeochem.2009.12.009

Metahni S, Coudert L, Gloaguen E, Guemiza K, Mercier G, Blais J F (2019) Comparison of different interpolation methods and sequential gaussian simulation to estimate volumes of soil contaminated by as, cr, cu, pcp and dioxins/furans. Environmental pollution 252:409–419, DOI: 10.1016/j.envpol.2019.05.122

MINES ParisTech / ARMINES (2023) RGeostats: The Geostatistical R Package. Free download from: `http://cg.ensmp.fr/rgeostats`

Müller S, Schüler L, Zech A, Heße F (2022) Gstools v1. 3: a toolbox for geostatistical modelling in python. Geoscientific Model Development 15(7):3161–3182, DOI: 10.5194/gmd-15-3161-2022

Murphy B S (2014) Pykrige: development of a kriging toolkit for python. In AGU fall meeting abstracts, volume 2014, H51K–0753

Pichon L, Ducanchez A, Fonta H, Tisseyre B (2016) Quality of digital elevation models obtained from unmanned aerial vehicles for precision viticulture. Oeno One 50(3), DOI: 10.20870/oeno-one.2016.50.3.1177

Rakotonirina H, Guridi I, Honeine P, Atteia O, Van Exem A (2024) Spatial interpolation and conditional map generation using deep image prior for environmental applications. Mathematical Geosciences :1–26 DOI: 10.1007/s11004-023-10125-2

Ren L, Lu H, He L, Zhang Y (2016) Characterization of monochlorobenzene contamination in soils using geostatistical interpolation and 3d visualization for agrochemical industrial sites in southeast china. Archives of Environmental Protection

Tao H, Liao X, Cao H, Zhao D, Hou Y (2022) Three-dimensional delineation of soil pollutants at contaminated sites: Progress and prospects. Journal of Geographical Sciences 32(8):1615–1634, DOI: 10.1007/s11442-022-2013-6

Tao H, Liao X, Yan X, et al. (2014) Uncertainty analysis and pollution volumetric calculation of soil bap contents in a contaminated site. Geographical Research 33(10):1857–1865

Tychola K A, Vrochidou E, Papakostas G A (2024) Deep learning based computer vision under the prism of 3d point clouds: a systematic review. The Visual Computer DOI: 10.1007/s00371-023-03237-7, URL: https://link.springer.com/article/10.1007/s00371-023-03237-7

Wang T, Gan V J (2023) Automated joint 3d reconstruction and visual inspection for buildings using computer vision and transfer learning. Automation in Construction 149:104810, DOI: 10.1016/j.autcon.2023.104810, URL: https://www.sciencedirect.com/science/article/pii/S0926580523000705

Xie Y, Chen T b, Lei M, Yang J, Guo Q j, Song B, Zhou X y (2011) Spatial distribution of soil heavy metal pollution estimated by different interpolation methods: Accuracy and uncertainty analysis. Chemosphere 82(3):468–476, DOI: 10.1016/j.chemosphere.2010.09.053