A machine learning-based approach for mapping leachate contamination using geoelectrical methods

Ester Piegari[1*], Giorgio De Donno[2], Davide Melegari[2], Valeria Paoletti[1]

[1] Dipartimento di Scienze della Terra, dell'Ambiente e delle Risorse, Università degli Studi di Napoli Federico II, Naples, Italy

[2] Dipartimento di Ingegneria Civile Edile e Ambientale, “Sapienza” Università di Roma, Rome, Italy

* corresponding author: ester.piegari@unina.it

Abstract

The growth of urbanized areas combined with the overall increase of world’s population is leading to an increase of waste disposal sites that put a serious threat to the environment and human health. Leachate is the main source of pollution in landfills and its negative impacts continue for several years even after landfill closure. In recent years, geophysical methods are recognized as effective tools for providing an imaging of the leachate plume. However, they produce subsurface cross-sections in terms of individual physical quantities, leaving room for ambiguities on interpretation of geophysical models and uncertainties in the definition of contaminated zones. In this work, we propose a machine learning based approach for mapping leachate contamination through an effective integration of geoelectrical tomographic data. We apply the proposed approach for the characterization of two urban landfills. For both cases, we use machine learning techniques to perform a multivariate analysis on datasets consisting of electrical resistivity, chargeability and normalized chargeability (chargeability-to-resistivity ratio) data extracted from previously inverted model sections. By executing a K-means clustering analysis, we find the best partitioning of the datasets into different classes and get updated cross-sections that provide a quantitative integration of the tomographic data, allowing an objective identification of the most polluted zones. Our findings, also supported by borehole data for one of the investigation sites, show that the combined use of geophysical imaging and unsupervised machine learning is promising in environmental applications and can yield new perspectives for the characterization of leachate distribution in landfills.

Keywords: Leachate contamination detection**,** Machine Learning, K-means clustering Geophysical imaging, Electrical Resistivity Tomography, Induced Polarization Tomography

1. Introduction

Although there is an increased awareness on the importance of environment protection, urban waste management is one the most important environmental issues. In many countries there is still a little use of recycling or reuse actions, and the majority of municipal solid waste is destined for landfills (WHO, 2015). Waste decomposition generates leachate that is a highly contaminated liquid consisting of a mixture from organic degradation products, liquid waste and rainwater. Leachate infiltration causes serious environmental issues to groundwater and soils and, therefore, identifying and monitoring its flow pathways has major implications on designing a risk mitigation strategy (Mukherjee et al., 2015; Lavagnolo et al., 2019; Vaccari et al., 2019; Morita et al., 2021; Ergene et al., 2022). To this aim, geophysical methods often represent the only cost-effective, rapid and non-invasive choice for mapping large areas, such as those encountered in urban landfills, down to tens of meters (e.g., Di Maio et al., 2018).

In last decades, many studies have demonstrated that geoelectrical methods can be effective in identifying landfill leachate (e.g., Soupios et al., 2007; De Donno and Cardarelli, 2017; Raji and Adeoye, 2017; Soupios et al., 2017; Power et al., 2018, Zaini et al., 2022). As leachate is a fluid with high concentrations of ions, it can be successfully imaged by low values of electrical resistivity ρ and high values of chargeability M, sensed respectively by electrical resistivity tomography (ERT) and induced polarization (IP) surveys (Everett, 2013). The contrast between the electrical properties of leachate and those of the surrounding media generally facilitates its identification. However, geophysical inverted models leave ambiguities in defining contaminated zones, particularly in presence of clayey soils, which are often placed at the bottom of the landfill as a low-permeability barrier in combination with a synthetic liner (geomembrane). Therefore, the zones characterized by the highest values of chargeability frequently do not match with the most conductive ones, and the question of how to combine information from different geophysical methods is still open. Many studies consider the so-called normalized chargeability, M_n (ratio of M and ρ), as it is directly linked to surface polarization (Slater and Lesmes, 2002). High values of M_n are generally related to leachate contamination, even though M_n can be also largely and significantly affected by clay content (Slater and Lesmes, 2002). Additionally, both resistivity and chargeability are related to the saturation levels, and to pick up the threshold of M_n for fully saturated zones (the most dangerous ones) remains a subjective choice. In fact, focusing only on the highest values of M_n may lead to false-positive results in the identification of leachate, as they may not arise from the concomitance of low ρ and high M values. For instance, large values of normalized chargeability may be caused by only large M related to the presence of pockets of clays with high-chargeability or by extremely low resistivity values of leachate-saturated zones. Thus, a residual uncertainty persists in properly identifying leachate accumulation zones.

With this study, we propose the application of machine learning techniques to combine ρ, M and M_n data with the aim to get one comprehensive section integrating all information from the electrical models.

Machine Learning (ML) is a branch of artificial intelligence based on the idea that systems can learn from data, and this learning comes from analyses of data through statistical tools to make predictions or finding patterns in data (Zhang et al., 2022). In recent years, applications of clustering algorithms to geosciences are continuously growing due to their potential of classifying large datasets into groups of data that share similar features (Lyra et al., 2014; Lindsey et al., 2018; Bernardetti and Bruno, 2019; Karpatne et al., 2019; Straus, 2019; Abdideh and Ameri, 2020, Kamer et al., 2020; Cesca. 2020; Piegari et al., 2022). There are many different types of clustering algorithms, which can be broadly categorised into three types: i) partitioning algorithms, which divide the dataset into a number of groups (clusters) and require the number of clusters as an input data; ii) hierarchical algorithms, which provide a tree-based representation of data points (dendrogram) showing the hierarchical relationship between clusters; iii) density-based algorithms, which groups data points on the basis of their spatial density. In this study, we use the K-means algorithm, which is one of the most common partitioning algorithms. In addition to being a fast, robust and simple iterative algorithm, it has the advantage of producing tighter (spherical) clusters than hierarchical and density-based clustering (Shukla and Naganna, 2014; Zhang et al., 2022).

In the following sections, we describe the proposed methodology and show the results of geophysical surveys performed into two urban waste landfills. Then, we report and discuss the results of the K-means clustering analysis for the two case studies.

2. Methods

We illustrate our methodological approach in Fig. 1. Traditionally, to characterize leachate contamination by means of geophysical methods, ERT and IP surveys are carried out with the aim of retrieving two cross sections showing the distribution of ρ and M, after data inversion. In this work, we propose to go beyond traditional approaches by: (i) combining all inverted values in a proper joint space and (ii) applying a cluster analysis to the unique large dataset. The retrieved cluster indices are used to get integrated cross sections that allow direct quantitative identification of leachate accumulation zones. In the following subsections, we give details about the inversion procedure and the clustering algorithms.

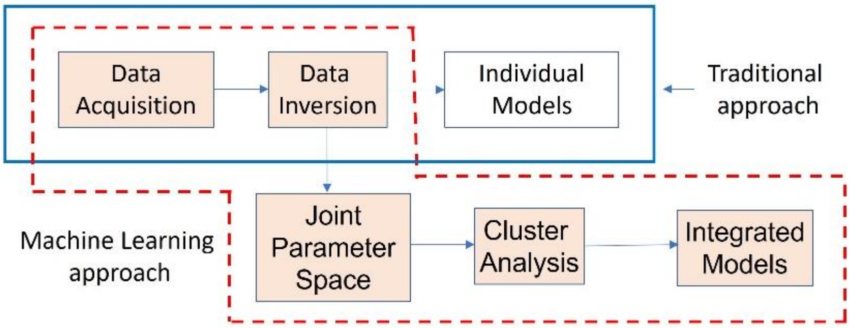

Figure 1. *Sketch of the traditional and ML based approaches. The red dotted line encloses the proposed procedure.*

2.1 Geophysical Imaging

2.1.1 Forward modelling

The resistive response of a 2.5D subsoil (where the conductivity varies only in the *x*-*z* plane), is described within a domain D by the Fourier-transformed Poisson's equation under the hypothesis of an external point source located at (x_s, z_s) (Dey and Morrison, 1979):

$$-\nabla \cdot [\sigma(x,z)\nabla\phi(x,z,\lambda)] + \lambda^2\sigma(x,z)\phi(x,z,\lambda) = I\delta(x_S)\delta(z_S) \quad \forall(x,z) \in D\,, \tag{1}$$

where σ is the conductivity (ρ=1/σ is the resistivity), ϕ the transformed electric potential, λ the transformed variable and I the injected current.

Eq. (1), subjected to Dirichlet's and Neumann's boundary conditions on surface and lateral and bottom boundaries respectively, is solved numerically. A widespread used technique is the Galërkin formulation of the Finite Element Method (De Donno and Cardarelli, 2017), which is employed in this work.

Once the solution of eq. (1) is achieved, potential is back-transformed and the apparent resistivity can be predicted as:

$$\rho_a{}^{pre} = C\frac{\Delta V}{I}\,, \tag{2}$$

where C is the geometric factor and ΔV the potential difference, both obtained depending on the specific quadrupole sequence.

The capacitive response of a medium can be assessed in the time-domain through the chargeability *M*, which is proportional to the induced polarization (IP) phenomena occurring in subsoil due to the switch-on or switch-off of an external direct current (DC) source (Siegel, 1959).

In this case, the IP forward solution is given sequentially with the resistivity modelling by calculating the potential V_M resulting from solution of eq. (1) but with the conductivity replaced by $\alpha = \sigma(1 - M)$.

The predicted apparent chargeability M_a is then found as (Oldenbug and Li, 1994):

$$M_a{}^{pre} = \frac{V_M - V}{V_M}. \tag{3}$$

2.1.2 Data inversion

Time-domain measurements are performed employing a DC electrical source, measuring the apparent resistivity ρ_a and apparent integral chargeability M_a, resulting from the potential decay after current switch-off (Binley and Slater, 2020):

$$\rho_a{}^{obs} = K\frac{V_p}{I}; \tag{4a}$$

$$M_a{}^{obs} = \frac{\int_{ti}^{tf} V_i dt}{V_p \Delta t}, \tag{4b}$$

where V_p is the measured voltage during application of the DC current I and V_i the residual voltage after switch-off the electrical current, integrated over a time window Δt defined between times t_i and t_f. Usually, the time window is divided into a few shorter logarithmically spaced gates (often 20), and the integral in eq. 4b is computed by sum of values achieved for each gate.

Apparent values are inverted for resistivity and chargeability using a Gauss-Newton iterative formulation (Loke and Barker 1996), where the chargeability dataset is inverted following the linear approximation proposed by Oldenburg and Li (1994). We set inequality constraints on chargeability ($M \geq 0.1$ mV/V) to avoid negative values in the inverted models. The goodness of fit is evaluated for each line in terms of mean absolute percentage error for resistivity, while for chargeability models is more convenient to use the mean absolute error, expressed in mV/V (De Donno and Cardarelli, 2017).

The contribution of the surface conduction mechanisms, quantified by the normalized chargeability M_n (Slater and Lesmes 2002), can be calculated at the end of the inversion process by:

$$M_n = \frac{M}{\rho} \,. \tag{5}$$

2.2 K-means clustering

K-means is a partitioning algorithm that divides the dataset into K predefined non-overlapping groups (clusters) of similar data. The similarity measure is based on a distance-based metric that is typically the Euclidean distance. The algorithm starts assigning a set of K means μ_1, μ_2,…, μ_K, also called centroids, computes the distance between each point and the K centroids, and assigns each point to the cluster i with the closest centroid μ_i. Actually, it assigns data points to a cluster in order to minimize the sum of the squared distance (SSE) between the data points and the cluster's centroids (Bhattacharya et al., 2021):

$$SSE = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} ||d_n - \mu_k||_2^2 \,, \tag{6}$$

where $||..||_2$ is the Euclidean L2 norm, d_n denote the data points, N is the total number of points in the dataset and r_{nk} is a binary variable equal to is 1 if d_n is assigned to cluster k and 0 otherwise. Once each point has been assigned to a centroid, the procedure is iterated, i.e., the centroid of each cluster is recomputed as the mean of all data points belonging to the same cluster, and the category of each point is adjusted again until a maximum number of iterations is reached, or the adjustment range is less than a given threshold.

K-means is a very powerful algorithm that, in addition to its simplicity, has the advantage of having convergence guaranteed, as at each iteration step *SSE* always decreases (Bhattacharya et al., 2021). However, K-means finds local minimum of *SSE* and different initial positions of centroids determine different cluster solutions. To overcome this problem, the algorithm is run many times placing the centroids in different random starting points, recording the variance at each step and selecting the configuration corresponding to the minimum variance. Another drawback of the algorithm is that the clustering depends on the number of *K* that needs to be specified in advance. The optimal value of *K* is found by elbow method, if there is not any geologic *a priori* information to constrain the number of clusters (Bhattacharya et al., 2021). This method consists in varying the number of clusters *K*, then computing the percentage of the explained variance *EV* for each trial:

$$EV_j = \frac{\sum_{i=1}^{j}(SSE_i - SSE_{i+1})}{\Delta SSE_{max}}\ j = 1,2,\dots,K-1 \qquad (7)$$

where ΔSSE_{max} corresponds to the *SSE* deviation between 1 and the highest *K* among those analysed.

The idea is that when increasing *K*, at some point the addition of another cluster does not significantly improve the modelling of the data, and, thus, the best value of *K* is chosen as the elbow of the *EV* curve (Thorndike, 1953). Our cluster analysis was performed by using software packages available in the Statistics and Machine Learning Toolbox of MATLAB.

3. Field data

3.1 Case 1 (Southern Italy)

3.1.1 Site location and geophysical measurements

The survey area of Case 1 is located in the Campania region (southern Italy). Basically, no information about the landfill design is available. This site is within a geological context characterized by a dense alternation of layers of greyish silty clays and clay and of lithoid arenaceous levels. This alternation is characteristic of the clayey-silty deposits largely outcropping both in the landfill area itself and in the neighbouring areas (Urban Plan of Montecorvino Pugliano, 2011).

Our geophysical survey consisted of resistivity and induced polarization measurements along four profiles, each one 142.5 m long, oriented approximately SW-NE (Fig. 2), spaced about 50 m apart (Profile1 to Profile4). The data along each profile were collected simultaneously by using a Syscal Pro 96 Switch resistivimeter (IRIS Instruments) in multi-electrode configuration with a unit electrode spacing of 1.5 m. We used the pole-dipole array and, for each profile, we positioned the (remote) current electrode 150 m away from electrode # 96, outside the landfill area. Data were filtered for outliers, negative DC and/or IP voltage values or decay curves with increasing voltage.

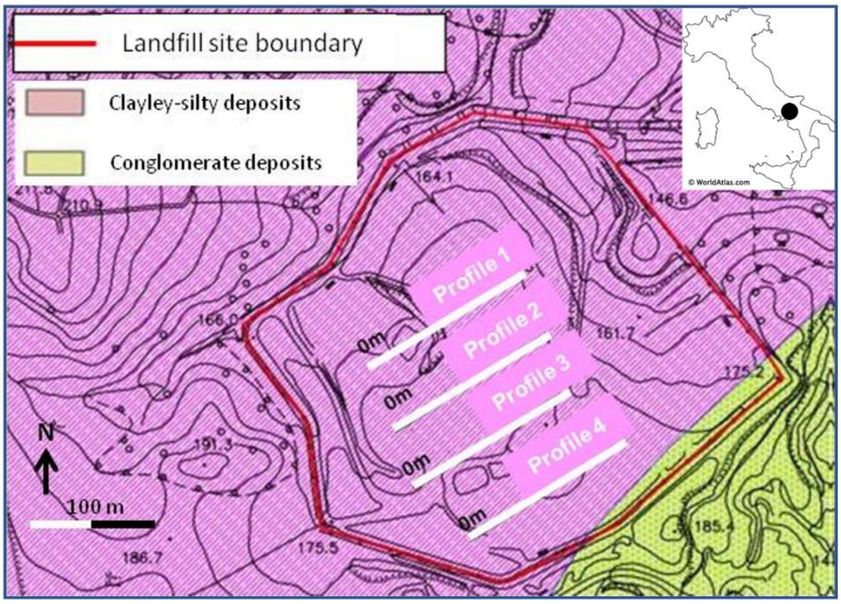

Figure 2. *Map of the survey area for the Case study 1 (landfill in Southern Italy). White lines show the resistivity and induced polarization measurement profiles.*

3.1.2 Inversion results

The inverted models are shown in Fig. 3 (resistivity) and Fig. 4 (chargeability).

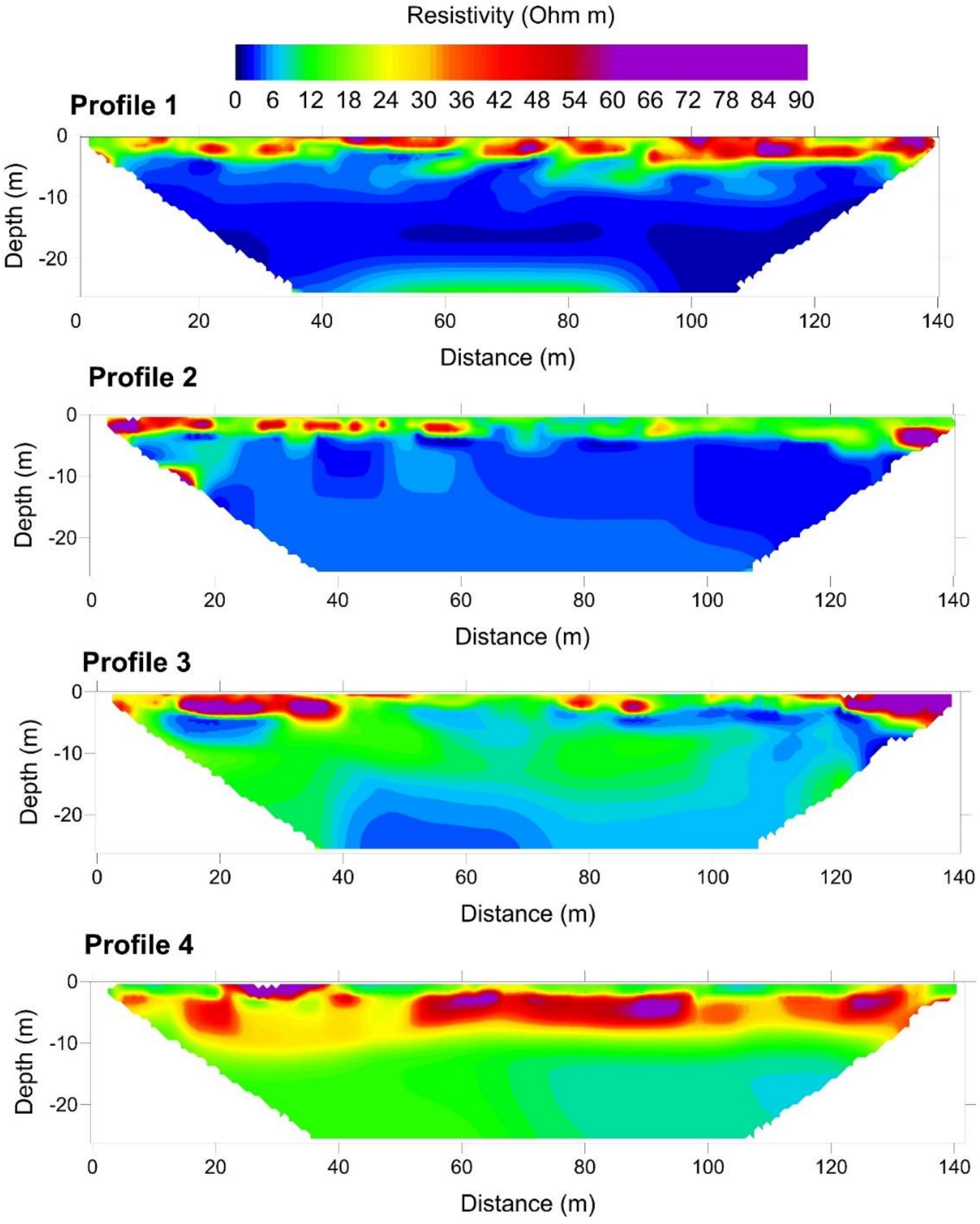

Figure 3. *Resistivity models for the Case study 1.*

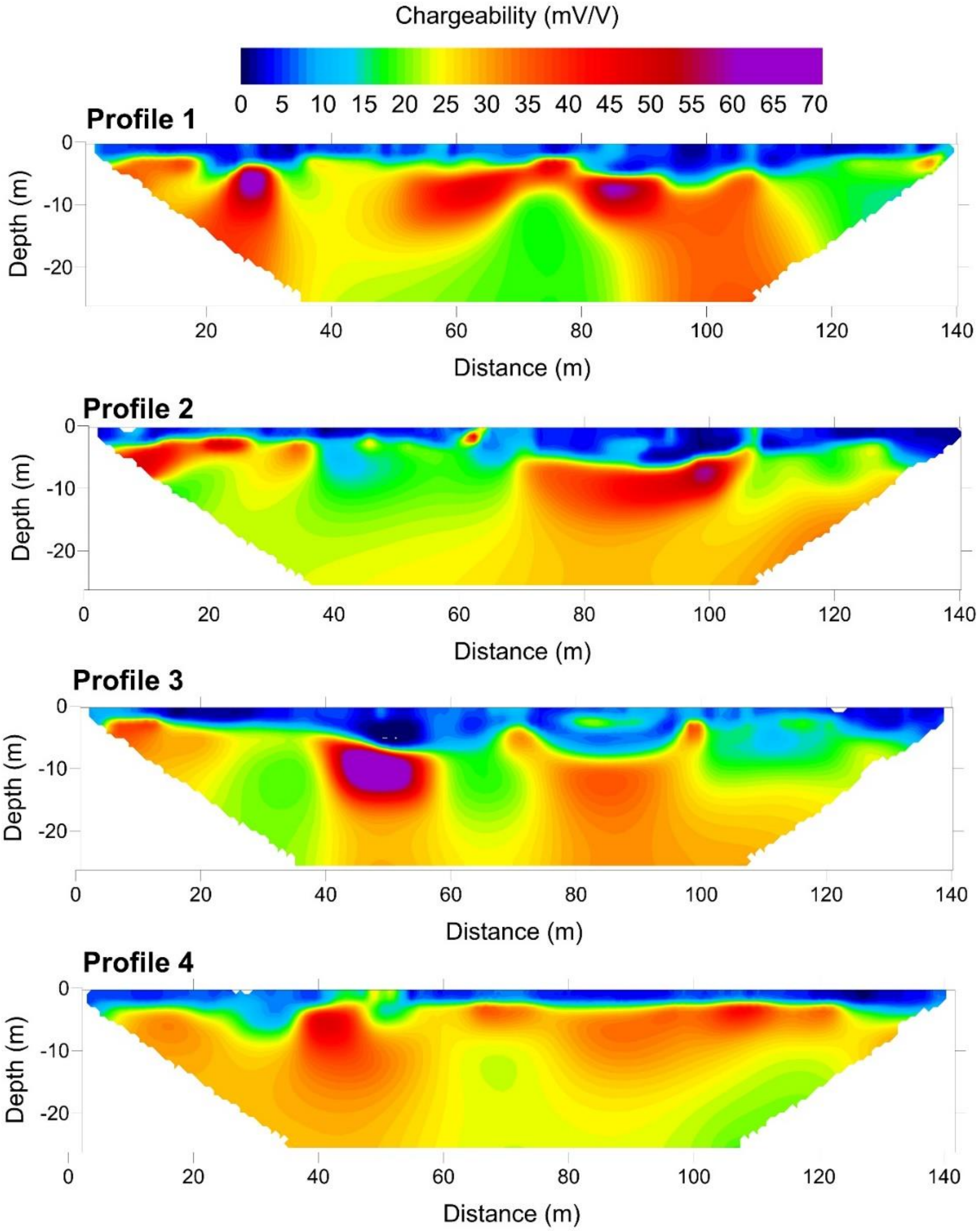

Figure 4. *Chargeability models for the Case study 1.*

For these four profiles we recognize a three-layer model from top to bottom:

- landfill covering and unsaturated waste, having a maximum thickness of 5 m, resistivity higher than 30 Ωm and chargeability close to zero, with local variation of thickness and resistivity (especially for Profile2 and Profile3) due to the high heterogeneity of the covering soil;

- saturated waste (leachate), with resistivity lower than 12 Ωm and chargeability higher than 20 mV/V. We notice strong lateral changes in resistivity and chargeability values throughout the sections due to waste-changes in composition and degree of saturation. The strongest geoelectrical anomalies (ρ < 3 Ωm and M > 35 mV/V) are observed in Profile1 that is the topographically lowermost profile, where an accumulation of leachate is more likely. For the same reason Profile4 is the profile with the smaller variations of resistivity and chargeability;
- a more resistive bottom layer, shown only in Profile1, whose resistivity (about 15 Ωm) is too low to be related to the presence of a bottom liner (geomembrane). The resistivity and chargeability values of this bottom layer of Profile1 (values higher than 10 Ωm and chargeability of about 20 mV/V) suggest that it may be constituted by clayey-silty deposits. For Profile1 we can thus estimate a maximum landfill thickness of about 20 m. The clayey-silty layer at the bottom of the landfill is very likely deeper for the Profile2, Profile3 and Profile4 and therefore cannot be detected in our profiles.

Therefore (Figs. 3-4), we believe that leachate percolates downstream (from Profile4 to Profile1) and accumulates at in the lowermost area, where Profile1 is located. Resistivity models show a gradually decreasing value of resistivity from Profile1 to Profile4, whereas chargeability models highlight for all profiles a high heterogeneity in leachate/waste distribution.

3.2 Case 2 (Central Italy)

3.2.1 Site location and geophysical measurements

The investigated area of Case 2 is a landfill built in the ‘80s for being used as municipal waste disposal of a medium-size city located in Central Italy. For this landfill, a few more information about the original design is available. The landfill site (Fig. 5), located on a steep slope (slope percent around 40%), was provided with a bottom liner (geo-membrane) overlying the in situ marly-arenaceous flysch. An embankment was built at the bottom (elevation around 445 m a.s.l.) to prevent slope instability phenomena. Geoelectrical investigation was planned for reconstructing the landfill depth and evaluating the leachate accumulation since it could reach a level that may trigger a slope failure. The supposed depth of the buried waste is greater upstream (southern part, elevation: 510-520 m a.s.l.), while reducing downstream (northern zone, elevation: 460-450 m a.s.l.).

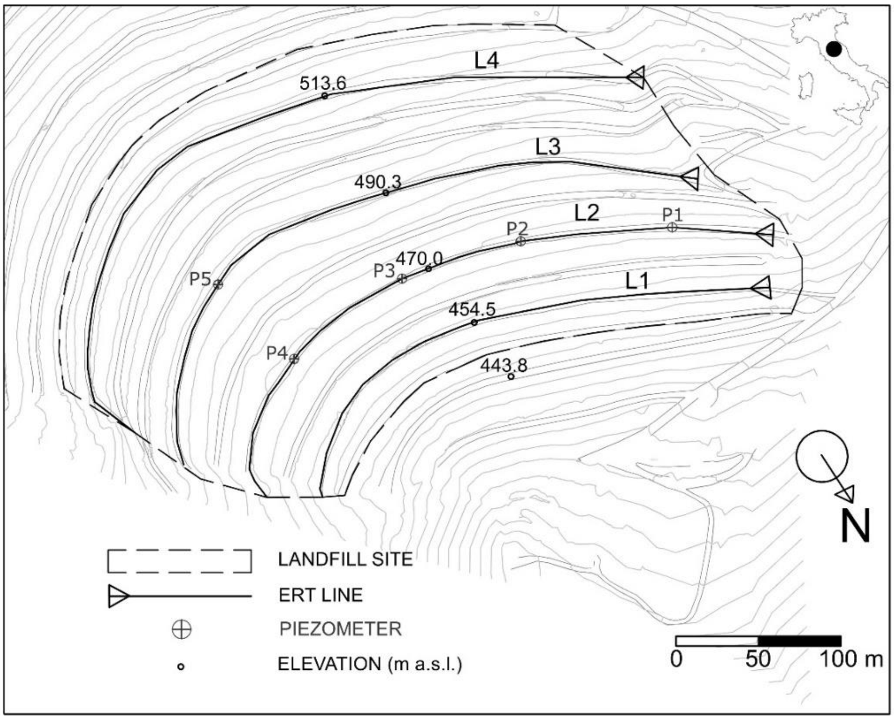

Figure 5. *Map of the survey area for the Case study 2 (landfill in Central Italy).*

The geophysical campaign encompasses four electrical profiles spaced approximately 40-50 m apart (L1 to L4), using the road tracks built for site management (Fig. 5). Five wells are located along L2 and L3 and piezometric levels were logged during the campaign to validate the geophysical models. Time-domain ERT and IP data were acquired by 48 electrodes spaced 5 m apart (Fig. 6), using the IRIS Instruments Syscal Pro resistivimeter. We employed the dipole-dipole array for data acquisition, using a maximum dipole length *a* = 5 and a maximum dipole separation factor *n* = 6 (945 data points for each baseline), as it combines significant depth of investigation and good lateral resolution needed to image the leachate variations. In fact, laying cables outside the landfill for set-up a pole-dipole configuration was unfeasible in such a steep and complex environment. For covering the long lines, we used the roll-along technique by overlapping 36 electrodes for each baseline. We set the current injection time to 2 s (2 stacks), a time delay of 40 ms and a logarithmic sampling of the IP decay curve using 20 gates (first gate centered at 40 ms, last gate at 1.7 s). Similarly to what done for Case 1, data were filtered for outliers, negative DC and/or IP voltage values or decay curves with increasing voltage.

3.2.2 Inversion results

The inverted models are shown in Fig. 6 (resistivity) and Fig. 7 (chargeability). The ERT and IP sections were computed in 2.5D mode by linearization of the curvilinear profiles, to compare the results with Case study 1. The error committed, calculated as the percentage deviation between 2D and 3D geometric factors, is between 0.5% (L1) and 1.5% (L4), due to the low curvature of profiles. We superposed the piezometric levels logged in the available wells to the electrical models (white filled areas in Figs. 6 and 7), in order to validate the geophysical results.

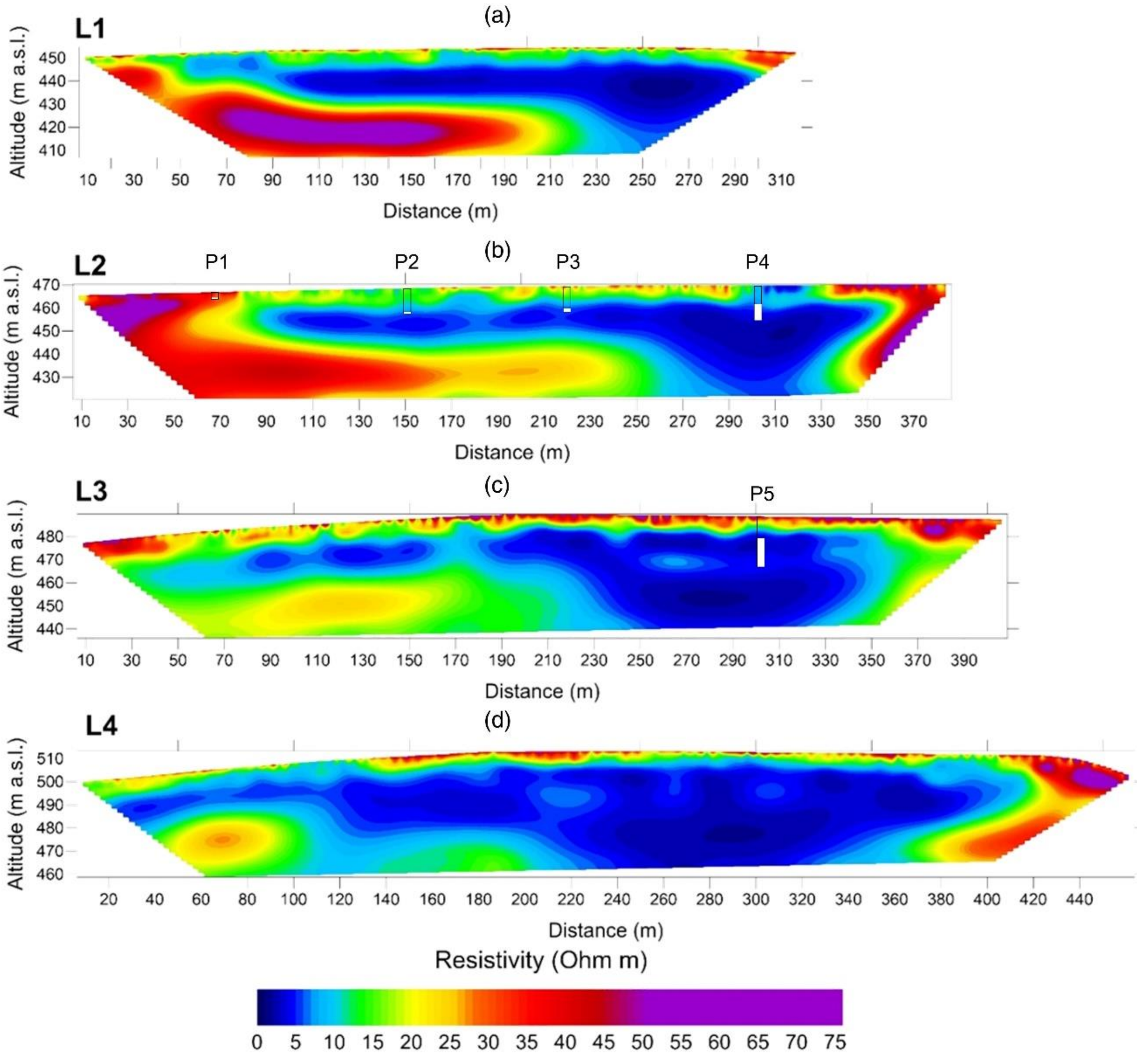

Figure 6. *Resistivity models for the Case study 2: (a) L1, (b) L2, (c) L3, (d) L4. Absolute error is 2.2%, 4.2%, 7.4% and 4.5% respectively. The piezometric levels in wells (white filled areas) are superposed to the models.*

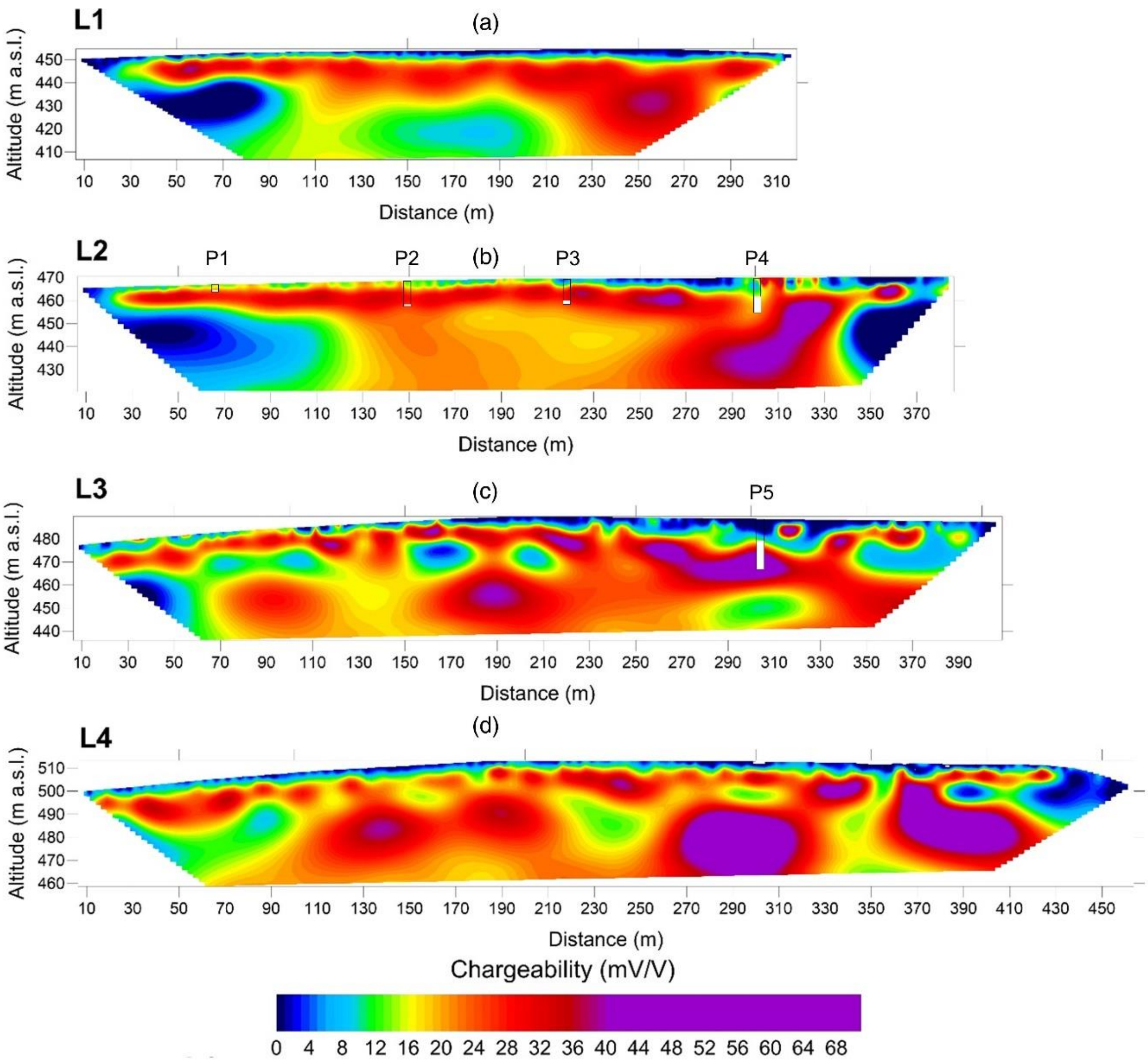

Figure 7. *Chargeability models for the Case study 2: (a) L1, (b) L2, (c) L3, (d) L4. Absolute error is 1.1 mV/V, 2.1 mV/V, 3.7 mV/V, and 4.6 mV/V, respectively. The piezometric levels in wells (white filled areas) are superposed to the models.*

Overall, we reconstruct for the four profiles a three-layer model from top to bottom:

- landfill covering and unsaturated waste, having an average thickness of about 8-10 m, resistivity higher than 15-20 Ωm and chargeability close to zero, with local variation of thickness and resistivity due to the high heterogeneity of the covering soil;
- saturated waste (leachate), with resistivity lower than 10 Ωm and chargeability higher than 10 mV/V. Strong changes in resistivity and chargeability values are clearly seen

throughout sections depending on composition and degree of saturation of the waste mass. The strongest geoelectrical anomalies ($\rho < 3$ Ωm and $M > 35$ mV/V) are located in the deepest zones of the landfill between $x = 240 – 350$ m from the beginning of the lines;

- bottom liner (geomembrane) overlying bedrock (marly-arenaceous flysch), with resistivity values higher than 10 Ωm and chargeability lower than 10 mV/V. The resistivity values are lower than expected for such a dielectric material (geomembrane), because of the lack of sensitivity at greater depths (De Donno and Cardarelli 2017). A series of sloped terraces is clearly visible, which ends in the deepest part of the landfill, where we detect the strongest geophysical response. Consequently, the maximum depth of the landfill can be estimated around 60-70 m upstream (L4) and 30-40 m downstream (L1).

Therefore, leachate likely accumulates at the bottom of the landfill and percolates downstream (from L4 to L1) through a preferential pathway, whose lateral extent progressively reduces downstream from about 80 m (L4, between 240 to 320 m) to about 30 m (L1, between 240-270 m). Chargeability models (Fig. 7) enhance the significant heterogeneity in leachate accumulation, mainly for L3 and L4 lines (where accumulation zones are larger), because of changes in degree of saturation or in waste composition. The piezometric levels in P4 and P5 wells confirms the geophysical reconstruction, while significant discrepancies can be noticed between chargeability models and leachate levels for P1, P2 and P3. This residual ambiguity leads the way for a more quantitative integration of geophysical data, which encompasses the joint use of both inverted models as an input for the clustering analysis.

4. Results

4.1 Case 1 (Southern Italy)

The results of the K-means clustering analysis in the 3D parameter space defined by the inverted values of ρ, M and M_n are shown in Figure 8. Since the inverted values span many orders of magnitude, we applied a log10 transformation and then a normalization in the range [0,1]. The clustering algorithm was iterated by varying K from 1 to 50 and for each run; we considered up to 500 different initial configurations of centroids choosing the configuration that minimizes the distortion (i.e., the sum of point-to-centroid distances). We retrieved the best choice for the number of clusters K_{best}=11 (Fig. 8a) using the elbow method with an explained variance threshold equal to 95%.

The clustering analysis highlights 11 different regions in the parameter space, whose centroids have the coordinates reported in Table 1 and are shown in Fig. 8b by blue markers. We associated the cluster indices to a colour scale (green-yellow-red) by computing the Euclidean distance to the point of normalized coordinates (0,1,1). This point corresponds simultaneously to the lowest values of ρ and the highest values of M and M_n, and therefore can be associated to the highest degree of contamination by leachate.

Being ERT and IP data acquired along the same profiles, triplets of values (ρ, M, M_n), in addition to being characterized by a cluster index, are also associated to the same depths. Thus, we use the retrieved cluster indices to get the integrated depth sections shown in Fig. 9. It is worth noting that such cross sections do not represent a geoelectrical model in a traditional sense, but combine information from all the investigated geophysical quantities.

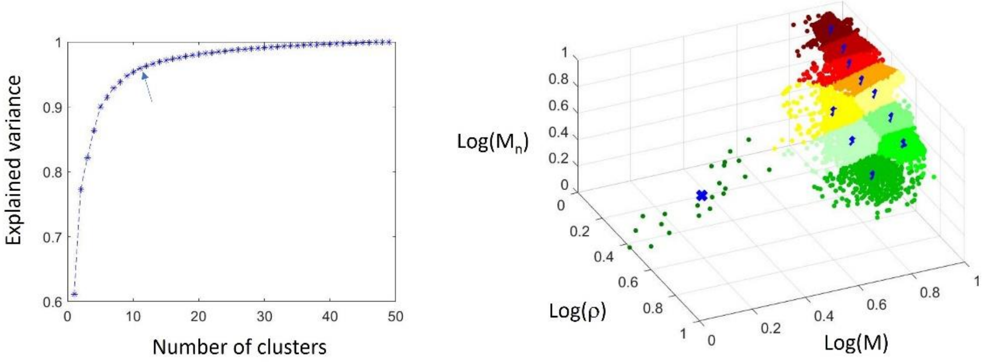

Figure 8. *Clustering analysis for the Case study 1: (a) Explained variance as a function of number of clusters. The best number of clusters is indicated by the arrow. (b) Scatterplot of geoelectrical data, where clusters are marked by different colours. In each cluster, the location of the centroid is shown by a blue cross.*

Table 1. Centroids coordinates of the clusters shown in Fig. 8b.

cluster color	ρ **(Ωm)**	M **(mV/V)**	M_n **(mS/m)**
dark green	7.81	0.01	0.001
	46.54	4.13	0.09
	38.43	21.71	0.57
↓	19.42	4.62	0.24
	20.76	26.28	1.27
light yellow	9.14	4.92	0.54
yellow	11.23	26.74	2.38
	7.52	23.70	3.15
↓	4.82	23.21	4.82
	3.43	27.04	7.88
dark red	2.10	27.32	13.0

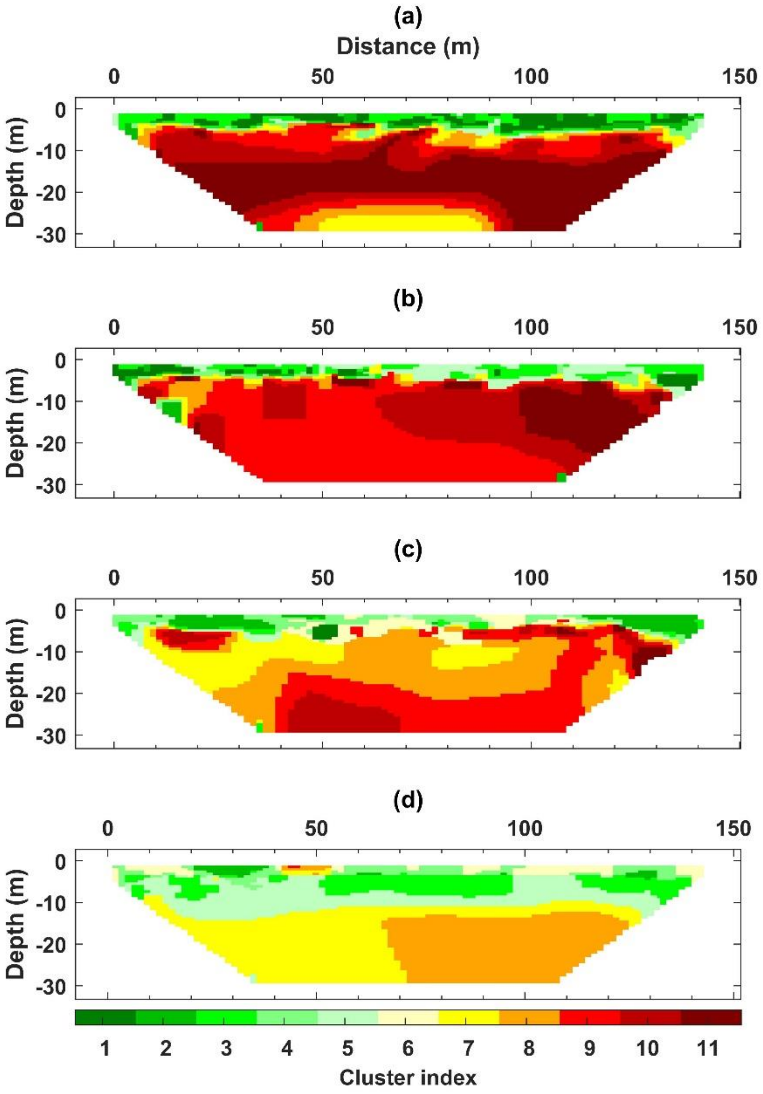

Figure 9. *Final cross-sections based on the proposed ML-based approach for the four profiles of the case study 1.*

4.2 Case 2 (Central Italy)

The results of the K-means clustering analysis performed on the second dataset (Figure 10) were obtained following the same clustering procedure as for Case 1. In this case, the 95% variance threshold determines the best configuration achieved with 10 different clusters (Fig. 10a). Looking at the distribution of data points in the parameter space (Fig. 10b), we note that the values of *M* span over a wider range with respect to the Case study 1 (Fig. 8b). The darkest green cluster in Fig. 10b that groups data with a shape of straight line is related to the inequality constraints set on the chargeability model during inversion to enforce positiveness of chargeability (minimum value of 0.1 mV/V in this case).

The final output of the clustering algorithm are the cross-sections shown in Fig. 11, where we observe a very good agreement between well and predictions of leachate contamination from cluster analysis.

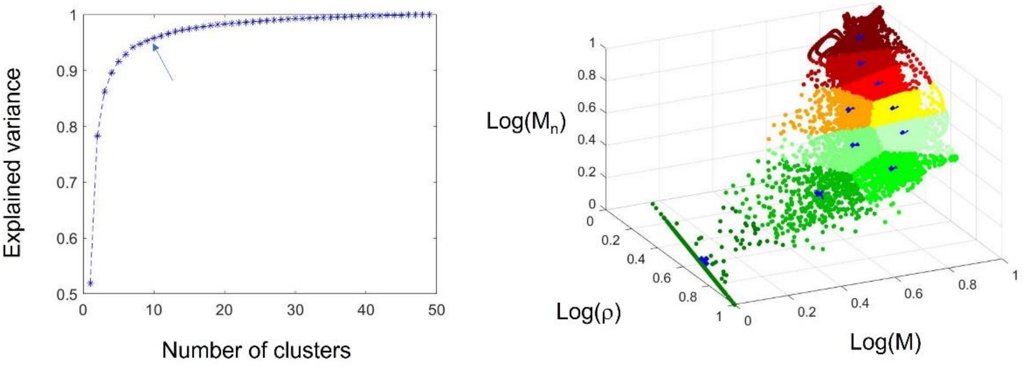

Figure 10. *Clustering analysis for the case study 2: (a) Explained variance as a function of number of clusters. The best number of clusters is indicated by the arrow. (b) Scatterplot of geoelectrical data, where clusters are marked by different colours. In each cluster, the location of the centroid is shown by a blue cross.*

Table 2. Centroids coordinates of the clusters shown in Fig. 10b.

cluster color	ρ (Ωm)	M (mV/V)	M_n (mS/m)
dark green	31.77	0.11	0.003
	30.30	1.89	0.06
↓	40.68	9.27	0.23
	18.39	6.64	1.36
	23.93	18.38	0.77
yellow	14.02	21.74	1.55
	9.25	10.08	1.09
↓	7.77	23.56	3.03
	4.59	23.19	5.05
dark red	2.83	32.37	11.44

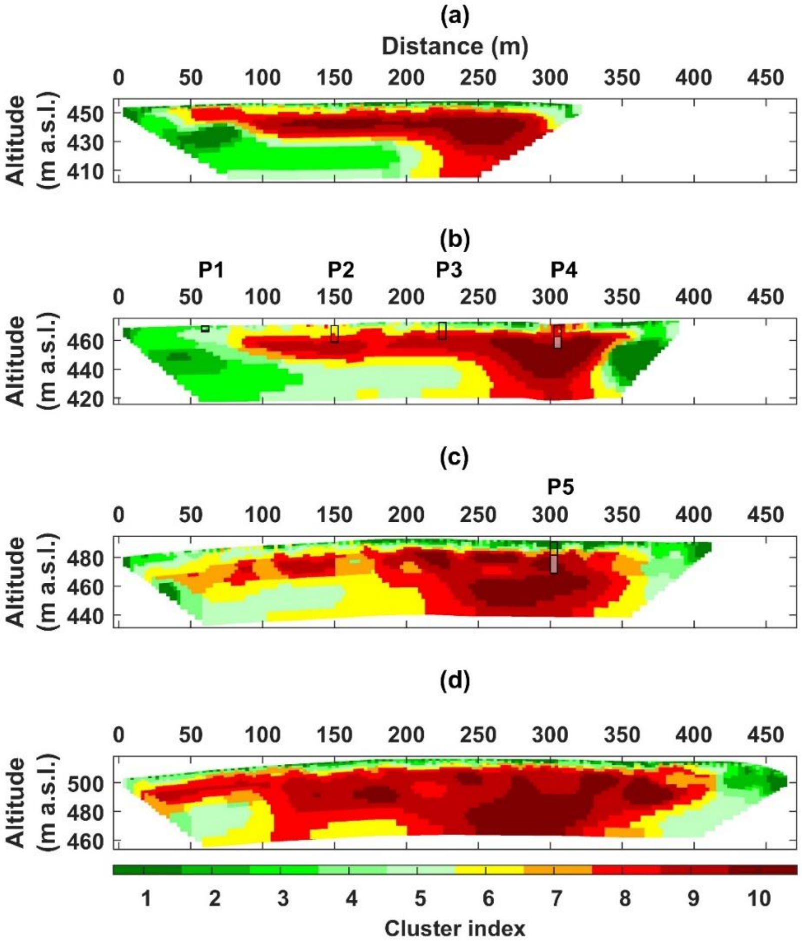

Figure 11. *Final cross-sections based on the proposed ML-based approach for the four profiles of the Case study 2. The piezometric levels are superposed to the models.*

5. Discussion

To discuss and show the effectiveness of the proposed ML-based approach, in Figs. 12 and 13 we compare the M_n models, achieved by the ratio of chargeability and resistivity of each pixel in Figs. 4, 5 and 6, 7 (for the two study-cases respectively), with the sections retrieved by the cluster analyses. In many cases the areas characterized by high values of M_n fall within the most hazardous zones identified by our cluster analyses (red clusters). Nevertheless, in many other zones there are significant differences between M_n models and the models obtained by our ML-approach. It is worth noting the lack of lateral continuity of the chargeable zones (i.e. Figs. 12a, 13a), compared to the respective ML images (Figs. 12e, 13e). This effect observed in the M_n sections increases the uncertainty on the model interpretation based only on inversion results, preventing a clear detection of the leachate accumulation zones. The effectiveness of K-means clustering in identifying different groups of data is particularly clear for the L2 and L3 lines of Case study 2 (Fig. 13), where a quantitative comparison with data from wells is available. The piezometric levels match well with the most hazardous areas (dark red clusters) identified by our clustering analysis, whereas the levels are less correlated with the highest M_n values. The very low level (0.2 m) logged in P1 is not highlighted by both procedures.

One of the main advantages of the proposed procedure is that the number of different zones is directly retrieved from the cluster analysis and the shapes of such well-defined zones are not affected by the choice of the colour scale. In fact, the scale was automatically distributed according to the distance of the centroids from the point (0, 1, 1), which represents the maximum level of contamination, given the electrical properties of leachate (highly conductive and chargeable). We chose to scale colours from red to green as it is used in alert systems to grade the severity of a hazard event. The final output is easy to interpret even for non-experts in the field of geophysics. Typically, leachate is controlled by drilling monitoring wells and our analyses show that the use of geophysical methods combined with ML techniques can provide very accurate information on optimal locations to plan for such wells. The machine learning-based approach can be therefore used to support decision-makers in the waste management sector and reduce the costs of effectiveness of landfill management.

However, this study has some limitations and there is a room for further developments. In fact, ML techniques are generally more effective if they use large datasets, and they can provide different results if datasets are differently scaled. For this reason, before developing the procedure, we preliminary explored different scenarios (Piegari and Paoletti, 2022; De Donno and Piegari, 2022). We found that clustering gives better results if data are scaled and using a 3D parameter space instead of a 2D space with only ρ and M_n. Additionally, the proposed identification of the most contaminated zones is only based on the electrical properties of leachate, since they are recognized as the most diagnostic for this purpose (Soupios et al., 2017). However, the proposed multivariate analysis could be applied also to other geophysical datasets, such as i.e. seismic tomography, which might add information about consolidation and compaction of waste materials.

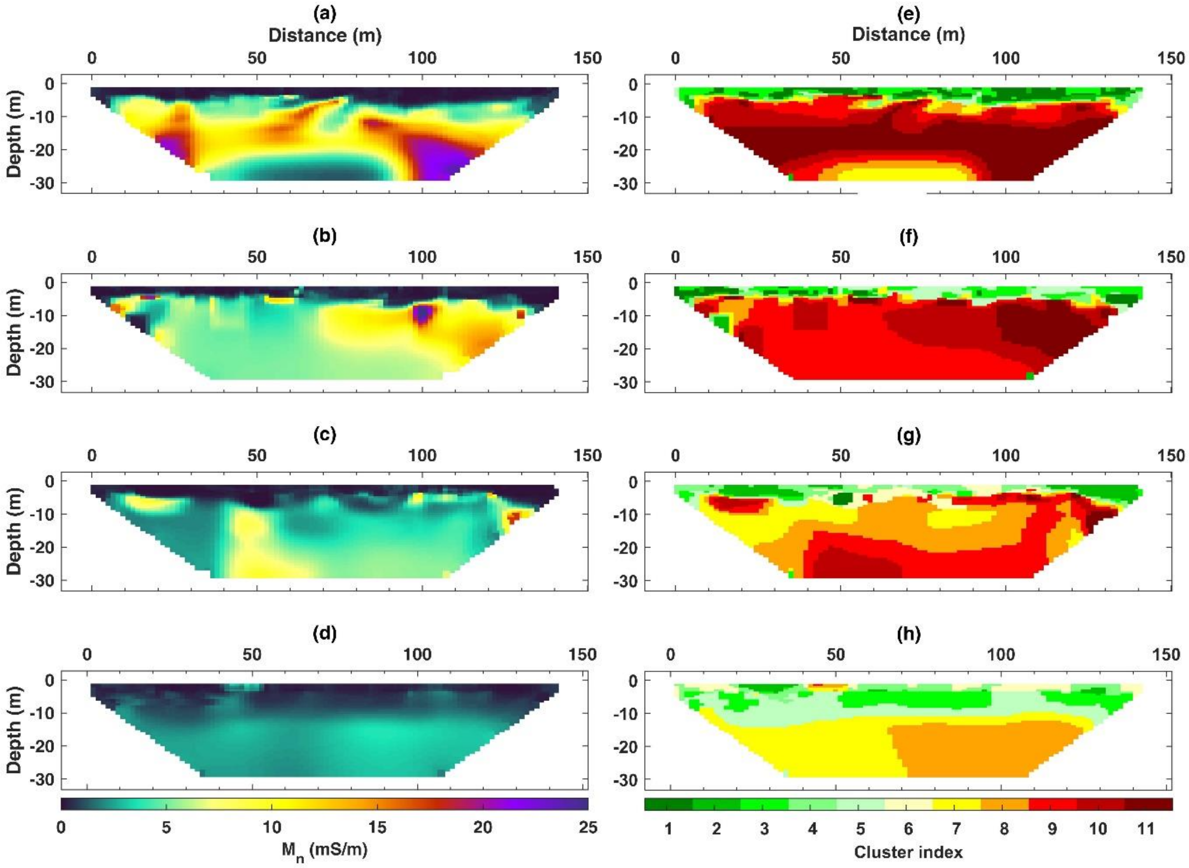

Figure 12. *Comparison between normalized chargeability sections (a-d) and integrated depth sections after clustering analysis (e-h) for the Case study 1.*

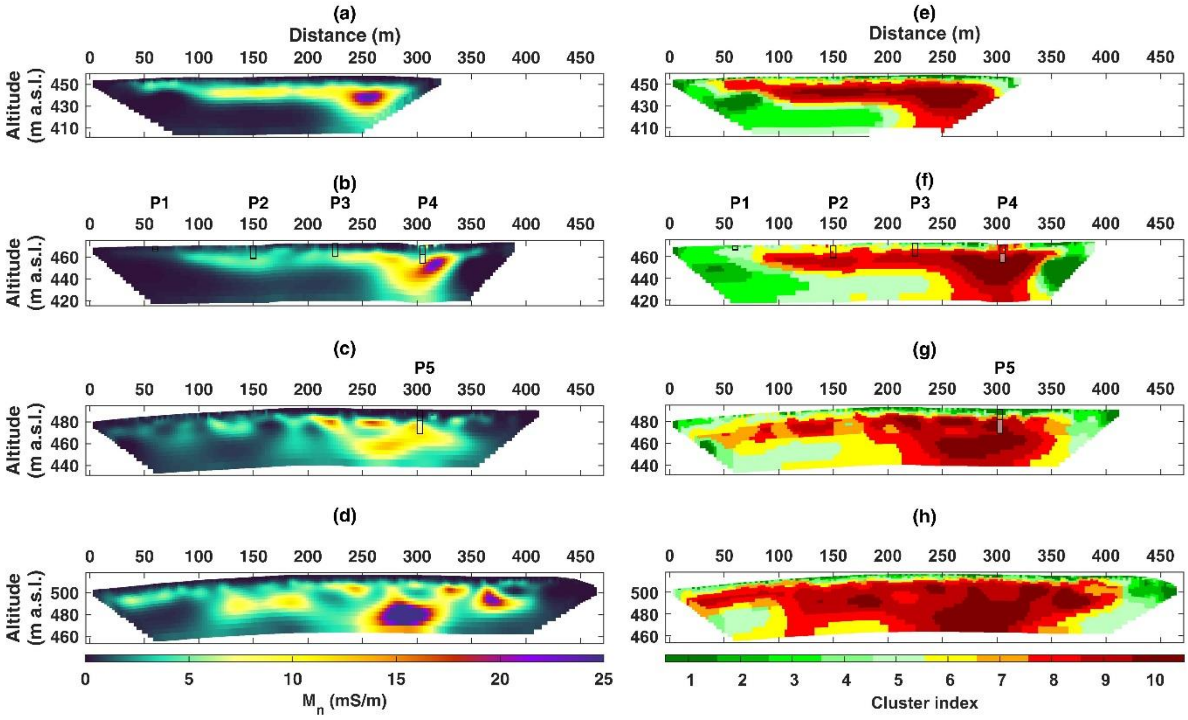

Figure 13. *Comparison between normalized chargeability sections (a-d) and integrated depth sections after clustering analysis (e-h) for the Case study 2.*

6. Concluding Remarks

In this study, we demonstrated that unsupervised machine learning may be successfully used to integrate data from resistivity and IP methods. The proposed ML-based approach is a flexible tool that can be easily adapted to other case studies also for different type of geophysical data. For each of the two investigated landfills, using K-means algorithm we were able to: i) achieve integrated model sections combining information from resistivity, chargeability and normalized chargeability data; ii) identify different regions in the investigated landfills associated with different leachate contamination levels; iii) locate the most hazardous zones. Therefore, our findings offer new perspectives for landfill characterization and have practical implications for landfill management. In fact, the final reconstructions of the investigated landfills help to predict the leachate flow pathways and enables more effective allocation of financial resources for the development of monitoring and remediation systems.

Acknowledgments

The authors are grateful to Donato Fiore (SOCOTEC S.r.l.) for making the data from case study 1 available. The authors are also grateful to former M.Sc. students Valeria Cariddi (Università degli Studi di Napoli Federico II) e Chiara Pagano ("Sapienza" Università di Roma) for preliminary data processing.

References

Abdideh, M., Ameri, A., 2020. Cluster Analysis of Petrophysical and Geological Parameters for Separating the Electrofacies of a Gas Carbonate Reservoir Sequence. Natural Resources Research 29(3), 1843–1856.

Bhattacharya, S., 2021. A Primer on Machine Learning in Subsurface Geosciences, 1st edn., Vol. 1, pp. 1–172, Springer.

Bernardetti, S., Bruno, P.P.G., 2019. The Hydrothermal System of Solfatara Crater (Campi Flegrei, Italy) Inferred from Machine Learning Algorithms. Frontiers in Earth Science 7, 286.

Binley, A., Slater, L., 2020. Resistivity and induced polarization: Theory and applications to the near-surface earth. Cambridge University Press.

Cesca, S., 2020. Seiscloud, a tool for density-based seismicity clustering and visualization. J Seismol 24, 443–457.

De Donno, G., Cardarelli, E., 2017. Tomographic inversion of time-domain resistivity and chargeability data for the investigation of landfills using a priori information. Waste Management 59, 302–315.

De Donno, G., Piegari, E., 2022. Clustering analysis of ERT/IP data for leachate mapping in urban waste landfills. Near Surface Geoscience 2022, Belgrade, 18-22 September (*accepted).*

Dey, A., Morrison, H.F., 1979. Resistivity modelling for arbitrarily shaped two-dimensional structures. Geophysical Prospecting 27(1), 106-136.

Di Maio, R., Fais, S., Ligas, P., Piegari, E., Raga, R., Cossu, R., 2018. 3D geophysical imaging for site-specific characterization plan of an old landfill. Waste Management 76, 629–642.

Ergene, D., Aksoy, A., Sanin, F.D., 2022. Comprehensive analysis and modelling of landfill leachate. Waste Management, 145, 48-59.

Everett, M., 2013. Near-Surface Applied Geophysics. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139088435

Kamer, Y., Ouillon, G., Sornette, D., 2020. Fault network reconstruction using agglomerative clustering: applications to southern Californian seismicity. Nat. Hazards Earth Syst. Sci. 20, 3611-3625.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Ali Babaie H., Kumar, V., 2019. Machine Learning for the Geosciences: Challenges and Opportunities. IEEE Transactions on knowledge and data engineering 31, 8, 1544.

Lavagnolo, M.C., 2019. Landfilling in developing countries. In: Cossu, R., Stegmann, R. (Eds.), Solid Waste Landfilling: Concepts, Processes, Technologies, Elsevier, pp. 773–796. https://doi.org/10.1016/B978-0-12-407721-8.00036-X.

Lindsey, C.R., Neupaneb, G., Spycher, N., Fairley, J.P., Dobson, P., Wood, T., McLing, T., Conrad, M., 2018. Cluster analysis as a tool for evaluating the exploration potential of Known Geothermal Resource Areas. Geothermics 72, 358–370.

Lyra, G.B., Oliveira-Júnior, J.F., Zeri, M., 2014. Cluster analysis applied to the spatial and temporal variability of monthly rainfall in Alagoas state, Northeast of Brazil. International Journal of Climatology 34(13), 3546–3558, doi.org/10.1002/joc.3926

Loke, M.H., Barker, R.D., 1996. Rapid least-squares inversion of apparent resistivity pseudosections by a quasi-Newton method1. Geophysical prospecting 44(1), 131–152.

Morita, A.K.M., Ibelli-Bianco, C., Anache, J.A.A., Coutinho, J.V., Pelinson, N.S., Nobrega, J., Rosalem, L.M.P., Leite, C.M.C., Niviadonski, L.M., Manastella, C., Wendland, E., 2021. Pollution threat to water and soil quality by dumpsites and non-sanitary landfills in Brazil: A review. Waste Management 131, 163–176.

Mukherjee, S., Mukhopadhyay, S., Hashim, M.A., Gupta B.S., 2015. Contemporary Environmental Issues of Landfill Leachate: Assessment and Remedies. Critical Reviews in Environmental Science and Technology 45:5, 472-590.

Oldenburg, D.W., Li, Y., 1994. Inversion of induced polarization data. Geophysics 59(9), 1327-1341.

Piegari, E., Herrmann, M., Marzocchi, W., 2022. 3-D spatial cluster analysis of seismic sequences through density-based algorithms. Geophysical Journal International 230, 2073-2088, https://doi.org/10.1093/gji/ggac160

Piegari, E., Paoletti, V., 2022. Analysis of geoelectric data through machine learning algorithms for waste leachate detection. Advances in Science, Technology & Innovation, *in press.*

Power, C., Tsourlos, P., Ramasamy, M., Nirvolis, A., Mkandawire, M., 2018. Combined DC resistivity and induced polarization (DC-IP) for mapping the internal composition of a mine waste rock pile in Nova Scotia, Canada. Journal of Applied Geophysics 150, 40–51.

Raji, W.O., Adeoye, T.O., 2017. Geophysical mapping of contaminant leachate around a reclaimed open dumpsite. Journal of King Saud University – Science 29, 348–359.

Shukla, S., Naganna, S., 2014. A Review on K-means DATA Clustering approach. International Journal of Information & Computation Technology. 4(17), 1847–1860. ISSN 0974-2239

Seigel, H.O., 1959. Mathematical formulation and type curves for induced polarization: Geophysics 24, 547–565.

Slater, L.D., Lesmes, D., 2002. IP interpretation in environmental investigations. Geophysics 67(1), 77–88.

Straus, D.M., 2019. Clustering Techniques in Climate Analysis. Climate Science, Oxford https://doi.org/10.1093/acrefore/9780190228620.013.711

Soupios, P., Papadopoulos, N., Papadopoulos, I., Kouli, M., Vallianatos, F., Sarris, A., Manios, T., 2007. Application of integrated methods in mapping waste disposal areas. Environ. Geol. 53(3), 661–675.

Soupios, P., Ntarlagiannis, D., Sengupta, D., Agrahari, S., 2017. Characterization and monitoring of solid waste disposal sites using geophysical methods: current applications and novel trends. Modelling Trends in Solid and Hazardous Waste Management, Edition 2017, Chapter. fifth Ed. Springer, pp. 29. https://doi.org/10.1007/978-981-10-2410-8_5

Thorndike, R.L., 1953. Who belongs in the family? Pyschometrika 18 (4), 267–276.

Urban Plan of Montecorvino Pugliano, 2011. Official Bulletin of the Campania Region, No 1 of 3 January 2011. https://www.comune.montecorvinopugliano.sa.it/?page_id=788

Vaccari, M., Tudor, T., Vinti, G., 2019. Characteristics of leachate from landfills and dumpsites in Asia, Africa and Latin America: an overview. Waste Manag. 95, 416–431.

WHO, 2015. Waste and human health: evidence and needs: WHO meeting report 5–6 November 2015: Bonn, Germany. https://apps.who.int/iris/handle/10665/354227

Zaini, M.S.I., Hasan, M., Zolkepli, M.F., 2022. Urban landfills investigation for leachate assessment using electrical resistivity imaging in Johor, Malaysia, Environmental Challenges 6, 100415.

Zhang, W., Zhang, Y., Gu, X., Wu, C., Han, L., 2022. Application of Soft Computing, Machine Learning, Deep Learning and Optimizations in Geoengineering and Geoscience, 1st edn, Vol. 1, pp. 1-138, Springer.