Environmental Earth Sciences (2026) 85:99
https://doi.org/10.1007/s12665-025-12784-2

ORIGINAL ARTICLE

Repurposing underutilized monitoring data from contaminated sites for sustainable groundwater characterization

Laura Landi[1] · Marco Rotiroti[2] · Chiara Zanotti[2] · Alessandro Amorosi[1] · Enrico Dinelli[1] · Maria Filippini[1]

Received: 1 August 2025 / Accepted: 20 December 2025 / Published online: 5 February 2026

Abstract

Knowledge of water quality is crucial for sustainable management of the natural resource, yet such information is often limited by data scarcity and high monitoring costs. This study proposes a novel approach, leveraging underutilized hydrogeochemical data from (potentially) contaminated sites to characterize natural groundwater composition at the mesoscale. Although originally collected for remediation purposes, these data, if properly processed, can yield valuable insights into natural groundwater conditions, offering a cost-effective sustainable alternative to extensive monitoring programs. The proposed workflow is applied in the Ferrara province (Po Plain, Italy), a region affected by both anthropogenic and natural groundwater quality issues. Aggregated data processed through reproducible steps underwent multivariate analysis, confirming the method's ability to depict natural geochemical heterogeneity. Results were validated against the official regional monitoring network, demonstrating improved spatial and temporal resolution. This approach provides a scalable solution for enhancing groundwater quality assessment and supporting sustainable groundwater management, without requiring significant new data collection.

Keywords Standardized workflow · Monitoring data · Groundwater characterization · Multivariate analysis · Geogenic contamination

Introduction

Groundwater is an essential freshwater source, supplying water to billions of people worldwide (Famiglietti, 2014). Its quality depends on a combination of physical, chemical, and biological processes (Cauich-Kau et al., 2025), and it is a crucial aspect for public health and environmental safety (Zahra et al., 2025). Understanding the hydrogeochemical status of groundwater, whether natural or impacted by anthropogenic activities, is fundamental for effective resource management. In particular, it is essential for distinguishing geogenic signals from human-induced contamination (Reimann & Garrett, 2005), establishing natural background levels that underpin regulatory standards (Edmunds & Shand, 2008; Muller et al., 2006), and assessing aquifer vulnerability and its capacity for natural attenuation (Foster et al., 2022). Moreover, this understanding also supports policy and decision-making by informing evaluations of groundwater suitability for various uses (WHO, 2017) and aids in predicting hydrogeochemical responses to environmental changes such as land use shifts and climate variability (Appelo and Postma, 2005; Edmunds, 2009).

The natural composition of groundwater can be assessed at different spatial scales and through different approaches (Busico et al., 2024). At large scales, assessments typically rely on data from regional monitoring networks or other datasets with relatively low spatial and temporal resolution (Molinari et al., 2019; Schiavo et al., 2024; Yan et al., 2025). These datasets are commonly analyzed using statistical methods (e.g., regression analysis, probability plots, boxplots analysis, component separation; Gałuszka, 2007; Muller et al., 2006; Reimann et al., 2005; Yan et al., 2024; Parrone et al., 2019), geostatistical techniques (e.g., kriging, object oriented spatial statistics; Dalla Libera et al., 2017;

✉ Laura Landi
laura.landi11@unibo.it

[1] Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum University of Bologna, Via Zamboni 67, Bologna 40126, Italy

[2] Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza 1, Milan 20126, Italy

Di Curzio et al., 2019; Li and Heap, 2014; Menafoglio et al., 2021; Mirzaei & Sakizadeh, 2016), and increasingly, machine learning approaches (e.g., random forest, artificial neural network; Abu et al., 2025; X. Li et al., 2025; Podgorski and Berg, 2020; Yadav et al., 2024; Yang et al., 2024; Wadoux et al., 2020). These methods produce large-scale overviews of groundwater composition, helping to identify broad geochemical patterns and background levels. At the local scale, investigations generally rely on ad hoc field surveys, which are more targeted but also costly, time-consuming, and often limited in applicability beyond their specific study sites (Washington et al., 2024). Ghiglieri et al. (2012), Giang et al. (2014), Rotiroti et al. (2015) provide examples of this approach.

To our knowledge, an integrated approach that is both relevant for understanding local-scale heterogeneities while still applicable beyond individual sites, is lacking. This gap largely arises from difficulties in data availability, high monitoring costs, and resource constraints (Bulut et al., 2020; Dixit et al., 2024; Mukherjee et al., 2024; Zhang et al., 2024). The reliability of chemical composition assessments depends on the extent and quality of monitoring data (Burt et al., 2014; Wagner, 1992). Regional monitoring networks are often too sparse to grab more detailed heterogeneity and support meaningful analysis at smaller scale, while setting up new dense networks or conducting extensive surveys is financially and logistically difficult.

However, an alternative source of groundwater monitoring data exists. In anthropized areas, numerous widely distributed sites have been investigated for suspected contamination through environmental characterization efforts, leading to the collection of extensive monitoring data. Where contamination was confirmed, the sites underwent further monitoring during the remediation processes. Hereafter, all the sites investigated for suspected contamination will be referred to as "contaminated sites", while acknowledging that contamination was not always confirmed.

The main objective of investigating contaminated sites is to identify contamination and implement appropriate mitigation strategies. To achieve this, groundwater is systematically monitored and characterized throughout the entire procedure (Benavides Höglund et al., 2024), which may span from several months to decades. Consequently, a vast amount of groundwater monitoring data is collected and stored. On the other hand, these data are fragmented across multiple sites, managed by different local owners and obtained by different laboratories with different analytical capabilities and detail, limiting their integration into a broader network. Moreover, the groundwater geochemistry at the contaminated sites is obviously impacted by human activities.

Although remediation procedures primarily focus on contamination, they also provide opportunities to identify and extract information on natural groundwater conditions (Güler, 2009).

This study explores the potential of treating such extensive yet dispersed monitoring observations as a unified dataset representing the natural groundwater composition. This approach aims to support groundwater characterization, geochemical analyses, and natural background level estimations at a "meso-scale" (i.e., between the large regional scale and the local site-specific scale), embracing local heterogeneity while offering insights applicable to broader-scale applications. Such scale integration is particularly relevant for informed groundwater management and strategic decision-making.

Our approach is applied in the Ferrara province of the Po Plain (northeastern Italy), a highly anthropized area with hundreds of contaminated sites, also known for the critical presence of geogenic arsenic contamination (Amorosi and Sammartino, 2024; Filippini et al., 2021). The coexistence of anthropogenic influences and strong natural geochemical processes occurring in groundwater make it an ideal case study for evaluating the proposed methodology. Based on the Ferrara experience, we propose a reproducible workflow to repurpose data from contaminated sites for assessing natural groundwater composition. The workflow's efficiency is validated through multivariate analyses aimed at identifying the main natural geochemical processes occurring in groundwater. The consistency of our findings is further explored by comparisons with the area's stratigraphic architecture and groundwater geochemistry from a regional monitoring network for quality assessment.

Methods

Study area

The study area is the administrative province of Ferrara, located in the lower Po Plain (northeastern Italy). It covers approximately 2,600 km² and is bordered by the Po River to the north, the Adriatic Sea to the east, and the Reno River to the south (Fig. 1). In terms of land use, agricultural and urbanized areas are predominant, covering 72% and 16% of the area, respectively. Due to extensive anthropization, groundwater is monitored in numerous contaminated sites, particularly in urbanized areas.

The hydrogeological setting consists of a multi-layered aquifer system formed by stacked depositional sequences of Quaternary alluvial deposits, spanning hundreds of meters below the topographic surface (Garzanti et al.,

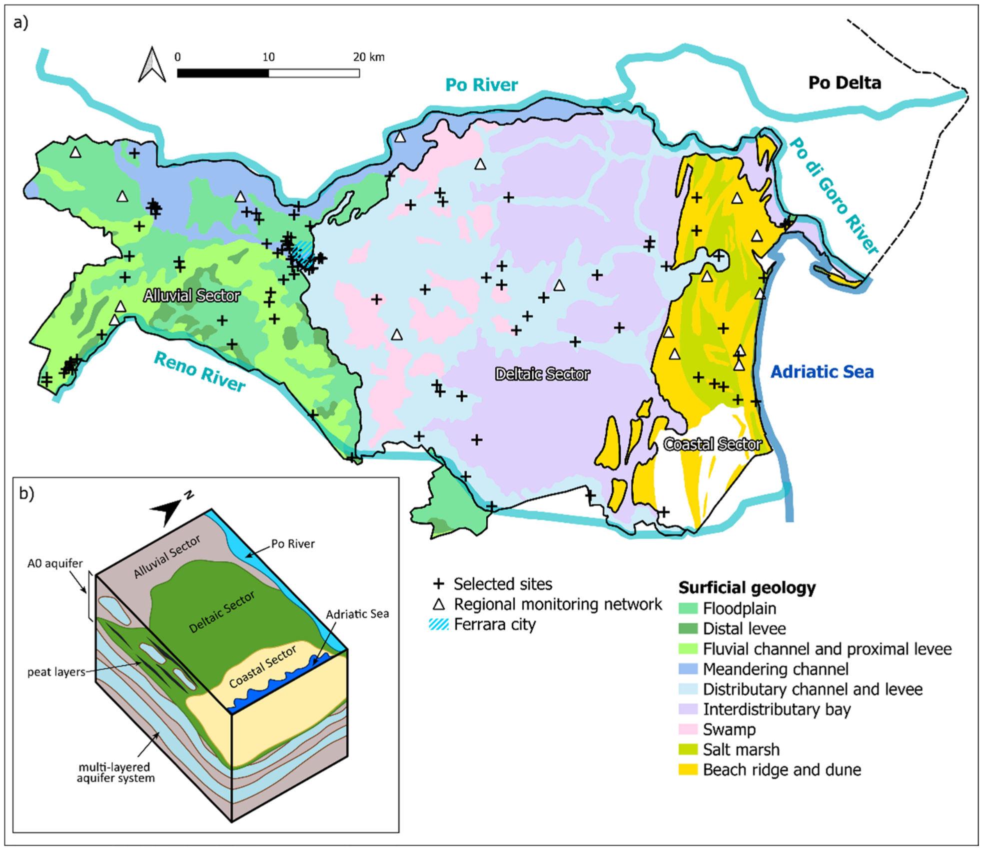

Fig. 1 (**a**) Surficial geology and depositional sectors of the study area. distribution of the 108 contaminated sites selected for data collection (black crosses) is compared with regional monitoring network for groundwater quality assessment of the shallow aquifer (white triangles). (**b**) Simplified hydrogeological conceptual model of the study area

2011). These deposits exhibit a rhythmic alternation of coarse-grained and fine-grained sediment bodies (Fig. 1b), reflecting glacial-interglacial cycles of Middle-Late Pleistocene age (Campo et al., 2020). This study focuses on the uppermost 15 m below the ground surface, where the so-called shallow “A0” aquifer is hosted within heterogeneous Holocene deposits (Molinari et al., 2007). The A0 aquifer is made of poorly interconnected sandy ribbons and lenses (aquifer) embedded within a silty or clayey matrix (aquitard). The heterogeneity of the Holocene succession is reflected in the surface geology (Fig. 1a), which is composed of alluvial sediments of Apennine provenance to the west (Amorosi and Sammartino, 2024), transitioning to the east into a complex coastal plain which shows contrasting (Po River versus Apennine) source-rock compositions (Amorosi et al., 2014). Southwest of the modern Po delta (Fig. 1a), an abandoned delta lobe hosts a series of marsh and swamp deposits, which form a thick wedge-shaped sediment body (Stefani and Vincenzi, 2005; Fig. 1b) that thins out to the west (Bruno et al., 2022). The easternmost sector of the study area is made up of coastal deposits (Fig. 1a) that consist almost entirely of beach-barrier sand, elongated parallel to the present shoreline (Amorosi and Sammartino, 2024). As a result of this shallow stratigraphic architecture, the A0 aquifer-aquitard system is typically enriched in peat and organic matter (Amorosi et al., 2017; Bruno et al., 2019), especially in the deltaic sector (Fig. 1b), where swamp deposits may crop out at the surface.

The A0 aquifer was selected for this study as it is the primary focus of groundwater remediation procedures, being the most susceptible to water quality issues arising from surface anthropogenic activities. Moreover, the widespread presence of peat deposits locally triggers chemical reactions associated with the degradation of organic matter, leading to changes in the redox state of the aquifer. This, in turn, promotes the mobilization of metals and semimetals (e.g., Fe, Mn and As) in groundwater (Berg et al., 2008), whose natural occurrence overlaps with possible anthropogenic sources of metal pollution.

Data collection

In Italy, contaminated sites are managed locally by private owners, according to the polluter pays principle, under the supervision of the provincial sections of the Regional Environmental Protection Agency (ARPA), as stipulated in Legislative Decree *D. lgs n. 152/2006*. Typically, the sites vary significantly in terms of contamination type and extent, and not all of them have the potential to provide relevant information for the proposed study. Therefore, preliminary criteria were established to select suitable sites prior to site-specific data harvesting. Sites were not considered if they met any of the following conditions:

- self-certified sites: these involve simplified procedures for minor incidents where groundwater is not fully characterized, but only a few parameters are measured once, to assess potential contamination;
- sites specifically related to oil spills, road accidents, and damages to underground tanks or electrical boxes: these cases follow specific procedures in which only a limited set of contaminants is typically monitored;
- extremely complex industrial aggregates where distinguishing between anthropogenic contamination and natural background signals is significantly challenging due to the high density of industrial activities (e.g., the Ferrara Petrochemical Plant).

After this initial selection, data were carefully harvested from the remaining sites. The information associated to each sample was systematically stored and aggregated into a unified, homogeneous data matrix ("sample matrix" hereafter), including 18 chemical species and physico-chemical parameters: Arsenic (As), Iron (Fe), Manganese (Mn), Sulfate (SO_4), Ammonium (NH_4), Nitrate (NO_3), Nitrite (NO_2), pH, ORP, Electrical Conductivity (EC), Chemical Oxygen Demand (COD), Dissolved Oxygen (DO), Oxidizability (Ox), Chlorine (Cl), Calcium (Ca), Magnesium (Mg), Potassium (K) and Sodium (Na).

Additional collected details include site-specific index contaminants, general site information (site history, contamination type, groundwater flow directions), piezometer information (location, depth, position relative to the contamination source, stratigraphic log) and sample information (sampling date and procedure), which are further detailed in the SM.

The proposed workflow

An aggregated sample matrix from contaminated sites suffers two main limitations: (1) lack of homogeneity due to e.g., variations in sampling procedures, changes in analytical methods over time, and potential human errors resulting from the involvement of numerous operators (these errors may include mistakes in sampling procedures and inaccuracies in reporting measurements); (2) occurrence of anthropogenic contamination, that must be removed to obtain a dataset representative of the natural groundwater composition.

Therefore, we developed a data treatment workflow which consists of two main steps: a first step for optimizing the homogeneity/quality of data (described in the following Sect. "Data quality optimization") and a second step for identifying and removing the site-specific anthropogenic influence on groundwater composition (Sect. "Removal of anthropogenic impact"). To enhance the reproducibility of the workflow, we minimized subjectivity by implementing a tiered approach that combines statistical analysis with expert knowledge (Fig. 2).

At the end of each of the two main steps, we determined a single representative value for each analyzed variable at each piezometer. Different approaches were applied based on variable type: for concentrations, we selected the third quartile (Q3) as the representative value. This choice provides an informative measure of the right tail, where potential anthropogenic contamination is expected, while maintaining robustness against outliers. For physico-chemical parameters (EC, ORP, and pH), we used the mean as the representative value to prevent biases toward positive ORP values, basic pH, and brackish conditions that could arise from using Q3. We acknowledge, however, that for parameters such as ORP, uncertainty from multiple sources may persist (Lindberg and Runnells, 1984). Due to the nature of the aggregated data, we could not trace the ORP measurements back to specific analytical instruments. However, we can reasonably assume that the Ag/AgCl electrode is generally the standard for field measurements in remediation procedures (US EPA, 2013). Therefore, when we refer to ORP hereafter, it implies a reference to an Ag/AgCl electrode. This aspect should be taken into account whenever the results show any apparent inconsistencies.

At the end of each of the two main steps we obtained two data matrices: an "anthropized" and a "natural" dataset representing, respectively, the actual composition of shallow groundwater incorporating both natural and anthropogenic signals, and the groundwater natural composition.

Data quality optimization

Piezometer selection based on screened aquifer. Since this study focuses exclusively on the shallow A0 aquifer, a depth threshold was defined as representative of the aquifer thickness based on previous hydrostratigraphic knowledge. Using available data such as piezometer depth, screen depth, and stratigraphic logs, all piezometers presumably tapping into deeper aquifers were identified and excluded from the dataset.

Management of censored data. In environmental datasets, chemical species concentrations commonly fall outside the instrument's measurement ranges, e.g., the limit of detection (LOD), and are recorded as "censored data", particularly for trace elements. Following the recommendations of the U.S. Environmental Protection Agency, we applied the LOD/2 substitution method for values below the detection limit, which is recommended when the percentage

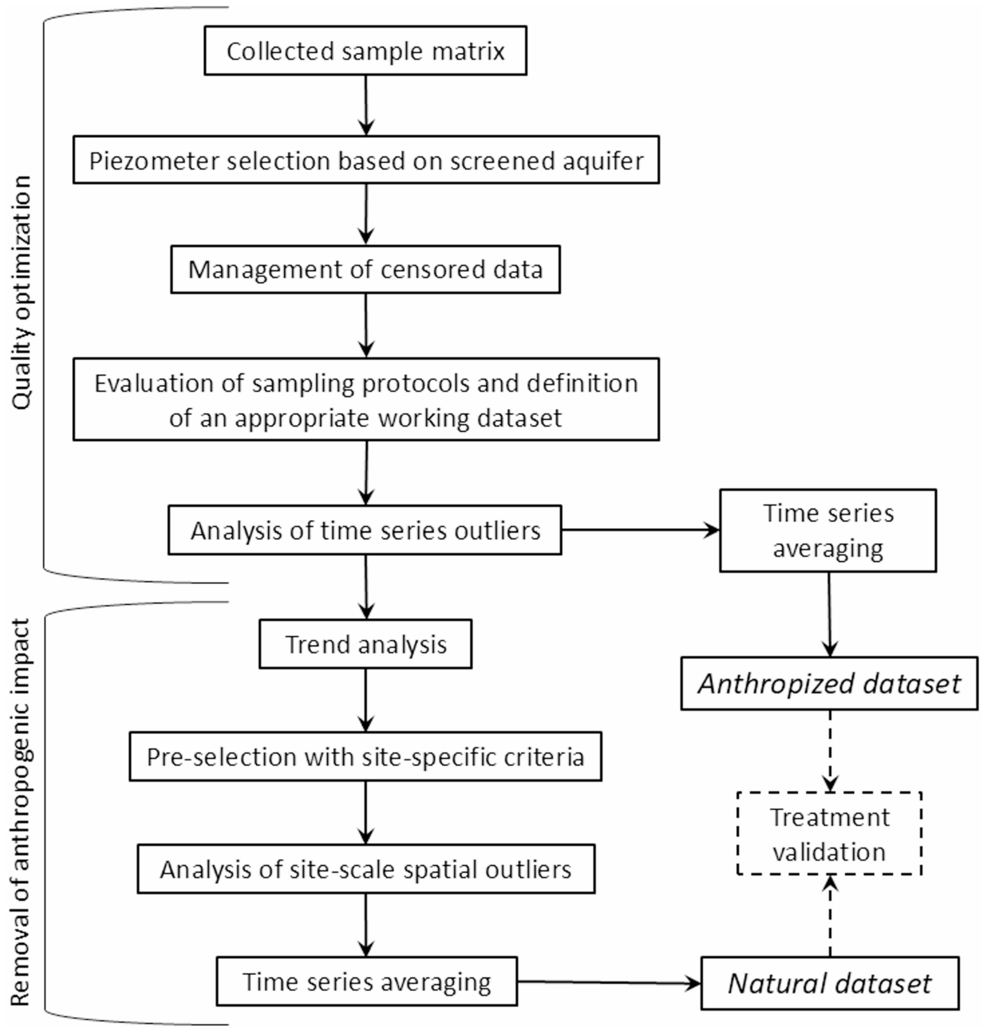

Fig. 2 Flowchart of the proposed standardized workflow for treating aggregated data from contaminated sites in order to obtain a dataset representative of the groundwater natural composition

of censored data is less than 15% (US EPA, 2006). Given the heterogeneity of data sources, we encountered a range of detection limits for the same variable. To address this, we selected, for each variable, the minimum LOD with a significant presence (≥5% of all the censored data; Zanotti et al., 2022). Each censored value was then replaced with half of this selected LOD (presented in the SM) to ensure consistency across the dataset.

Evaluation of sampling protocols. One of the main challenges in aggregating data from contaminated sites is the variability of sampling protocols over space and time. In particular, field filtering has a significant impact on data reliability (Knaack et al., 2021). It is known that for certain species, especially metals, unfiltered samples tend to exhibit higher concentrations and greater variability due to the presence of suspended colloidal particles (Guo et al., 2009). Standard 0.45 µm filter was used for field filtration of the collected samples, in accordance with national and regional guidelines (ARPA, 2013; ISS, 2008). Comparisons of duplicate filtered and unfiltered samples were performed for As, Fe, and Mn through scatterplots to evaluate the consistency between the different field filtering approaches. The portion of data for which the sampling procedure was unknown was also subject to statistical analysis: the different populations were compared through the Mann-Whitney test (Mann and Whitney, 1947; Wilcoxon, 1945) aiming to investigate the potential assignment to one of the two classes (filtered/unfiltered).

Analysis of time series outliers. Outlier detection in time series was performed using the boxplot method (McGill et al., 1978) to identify significant anomalies, which could indicate human errors or rare environmental conditions. When at least three outliers were detected within a single sample across different variables, the entire sample was flagged as "anomalous," provided the time series contained at least eight recorded values. Anomalous samples were considered to be due to potential sampling errors or exceptional environmental conditions and were removed from the dataset. For the remaining outliers, we evaluated their impact on the representative statistics of each time series (i.e., mean and

Q3, as described in Sect. "The proposed workflow"). Outliers that caused a shift in these statistics by more than 15% (hereafter referred to as "extreme outliers") were assumed to be the result of human sampling or recording errors, or at least uncommon environmental conditions. Extreme outliers were removed from the time series of the respective variable. For those variables measured as concentrations (As, Fe, Mn, SO_4, NH_4, NO_3, NO_2, COD, DO, Ox, Cl, Ca, Mg, K, Na) we focused exclusively on upper outliers. This decision was based on the expectation that anomalies in trace elements would primarily manifest as "spikes" in higher concentrations, rather than lower values. Moreover, the presence of censored data and their substitution could affect the detection of lower tail outliers, making them less reliable for analysis. For physico-chemical parameters (pH, ORP, and EC), which are not measured as concentrations and do not include censored data, we investigated both upper and lower outliers.

Removal of anthropogenic impact

Trend analysis. Each time series with a sample size greater than 4 (the minimum sample size required by the used *Kendall* and *rkt* packages for R software; Marchetto, 2024; McLeod, 2022) was tested to detect significant temporal trends that might indicate anthropogenic influences (Mendizabal et al., 2012). Two methods were applied depending on data numerosity: (a) Seasonal Kendall (SK) test (Hirsch et al., 1982) was applied to time series with at least two seasons providing each a minimum of 4 values, whereas (b) Mann-Kendall (MK) test (Kendall, 1948; Mann, 1945) was applied to all of the other time series with numerosity greater than 4. The MK test is the most commonly used test for detecting trends, but the SK test is more robust against seasonal variations and outliers, making it preferable when applicable. Piezometers that showed significant trends (p-value <0.05) for at least one variable were considered potentially contaminated and were subsequently removed from the dataset.

Pre-selection with site-specific criteria. Pre-selection (PS) is a widely recommended method in national and European guidelines for water quality assessment (Bulut et al., 2020; Muller et al., 2006), aimed at removing anthropogenic contamination from a dataset. However, commonly recommended PS indicators (e.g., nitrate, sulfate, chloride or ammonium) are not frequently measured in the remediation procedures. Moreover, these indicators cannot be used in conditions of salinized and anaerobic groundwater (Hinsby et al., 2008). Since both conditions were encountered in our study area, we decided to rely on site-specific indicators, corresponding with index contaminants representative of the type of anthropogenic impact at the site. The pre-selection threshold was set at 75% of the regulatory limit, as recommended by national guidelines (ISPRA, 2018). Piezometers exceeding the threshold at least once in their time series were considered potentially contaminated and removed from the dataset.

Analysis of site-scale spatial outliers. Spatial outliers were investigated within each site using the boxplot method, considering the single representative value for each piezometer defined at the end of the first workflow step of quality optimization (see Sect. "Data quality optimization"). We hypothesize that spatially localized anomalies at the site scale are likely linked to anthropogenic influence. These spatial outliers were further validated using site-specific knowledge about piezometers installed upgradient of the contamination source to monitor background groundwater conditions. An outlier was considered to be a natural anomaly if: (a) it was an upgradient piezometer, (b) its value was either lower (in the case of an upper outlier) or higher (in the case of a lower outlier) than the values measured in upgradient piezometers at the same site. This assumption is based on the idea that upgradient piezometers represent the composition of uncontaminated groundwater. Spatial outliers that could not be classified as natural anomalies were considered anthropogenic and removed from the dataset. As explained in Sect. "Data quality optimization", only upper outliers were considered for variables measured as concentrations, while both upper and lower outliers were examined for physicochemical parameters (pH, ORP, and EC).

Dataset testing through multivariate analyses

Multivariate analyses were performed on both the anthropized and the natural datasets. Given the heterogeneity of our data, we had to address the presence of missing data to proceed with the analyses. To minimize this issue, only the seven variables with less than 50% missing data were selected: As, Fe, Mn, ORP, pH, EC, and DO. Regarding these variables, removing cases with missing values through the common practice of "complete case analysis" would have led to an unacceptable reduction in the dataset (Baker et al., 2014). On the other hand, model-based approaches would have required defining a model for the complete data, along with assumptions about the missing data mechanism (Cheng and Huang, 2021; Lakshminarayan et al., 1999). Therefore, for the remaining missing data, we opted to apply an imputation procedure using a spatial nearest-neighbor approach: for each missing value, we identified the nearest site where the variable was available and replaced the missing value with the median of that site's observations. This method leverages spatial proximity to estimate missing values, ensuring that imputation is based on geographically relevant data rather than arbitrary or statistically derived estimates. By using the median rather than a single nearest-neighbor value, we reduce the influence of potential outliers and improve the robustness of the imputation process.

A factor analysis (FA) was then used to highlight correlations between variables, aiming to identify the main geochemical processes occurring in groundwater. The FA was based on the calculation of the correlation matrix using the Spearman method (Spearman, 1904). Compared to the Pearson method, the Spearman correlation coefficient does not require the assumption of linear relationships between variables, and it is less sensitive to outliers (Bocianowski et al., 2024). The FA was performed on standardized data (Judd, 1980), then an *oblimin* rotation was applied (Carroll, 1953). According to the Kaiser criterium (Kaiser, 1960), only factors with eigenvalues > 1 were considered significant. The significance and reliability of the FA was evaluated through the statistical indicators of Root Mean Square Error of Approximation Index (RMSEA) and Tucker Lewis Index (TLI) (Hu and Bentler, 1999; Marsh et al., 2004).

A hierarchical cluster analysis (CA) was eventually applied to the obtained factor scores, allowing the definition of hydrogeochemical facies where similar geochemical processes occur. The selected methodology for CA was the Ward method (Ward, 1963), using the Euclidean distance. To minimize the influence of missing data imputation on the multivariate analysis, after the CA we discarded all piezometers where the most representative variable(s) of the corresponding cluster had been previously imputed. As a result, each considered piezometer relies on the real values of the key variables defining its cluster. The resulting clusters were eventually plotted to visualize the spatial distribution of hydrogeochemical facies in the study area.

Given the significant influence of the area's stratigraphic architecture on groundwater geochemistry, particularly due to the presence of organic matter-enriched deposits, we compared the clustered piezometers with the stratigraphic facies association (sensu Bruno et al., 2019) intersected by their screens. Specifically, we distinguished between deltaic swamp deposits, primarily composed of soft grey clay, organic clay and peat layers, and alluvial and coastal deposits, which mainly consist of yellow/brown sand, silt, or compact clay with no evident organic matter. We defined three classes based on the intersection between piezometer screen and facies association: (1) a "swamp" class, including piezometers predominantly (> 25% of screen length) screening swamp deposits; (2) a "non-organic" class, comprising piezometers entirely screening alluvial and subordinate coastal deposits; and (3) a "transition" class, consisting of piezometers primarily screening alluvial/coastal deposits but positioned near the transition to swamp deposits, with the contact surface located just below the screen's end or intersecting swamp deposits for less than 25% of the screen length.

Results and discussion

Data quality optimization and the "anthropized dataset"

As of 2023, records of more than 350 contaminated sites were stored in the ARPA offices of Ferrara, containing data from both completed and ongoing remediation procedures. The initial site selection identified 108 sites potentially suitable for this study (detailed in the SM) providing 1,695 piezometers, 12,741 samples, and 72,772 measured values. The aggregated dataset from the 108 sites spans 43 years of monitoring, from February 1981 to December 2023. Due to the diverse nature of the data sources, the dataset is highly heterogeneous. Several factors contribute to this heterogeneity, including variations in the measured variables for individual samples (i.e., not all 18 chemical parameters are available for every sample) and inconsistent temporal coverage (detailed in the SM).

Based on previous stratigraphic and hydrostratigraphic studies (Bruno et al., 2017; Colombani and Mastrocicco, 2016; Gaiolini et al., 2024; Molinari et al., 2007), a threshold depth of 15 m bgs was established to differentiate between the piezometers tapping the surficial, Holocene A0 aquifer (piezometer depth or maximum screen depth < 15 m) and those reaching deeper, Pleistocene aquifers. Out of the 1,695 piezometers, 1,480 were classified as representative of the A0 aquifer based on this threshold. The remaining 13% of piezometers reaching deeper aquifers was excluded from the dataset.

For three of the 18 chemical variables included in the sample matrix, the proportion of censored data exceeded the recommended 15% threshold for applying the LOD/2 substitution method: As, NO_3 and NO_2 had censored data rates of 27%, 50% and 61%, respectively. However, the LOD/2 approach is still considered reasonable for datasets where censored data comprise up to 60% of observations, provided that the dataset is large (> 20 samples), exhibits a highly skewed distribution (geometric standard deviation, GSD > 3), and contains multiple LOD values (Hewett and Ganser, 2007; Hornung and Reed, 1990). All these conditions were met in this study. Moreover, it is important to note that the subsequent averaging of time series into a single representative value per piezometer helped mitigate the impact of censored data, reducing the proportion of censored values for As, NO_3, and NO_2 to 14%, 40%, and 56%, respectively.

Comparisons of duplicate filtered and unfiltered samples for As, Fe, and Mn, presented in the SM, confirmed the influence of solid particulate in measurement of dissolved metals, with unfiltered samples showing greater variability and in general greater concentration. This aligns with previous studies addressing the same issue (Filippini et al., 2021; Kumanova et al., 2015; Tang et al., 2024). Given the risk of introducing bias and significantly increasing uncertainty, we established that unfiltered samples were not reliable enough to be included in the analysis. The 19% of the samples available at this stage of the workflow were labeled as “unknown” due to missing information on the sampling procedure. A Mann-Whitney test performed at the site scale, to assess whether the unknown samples were more similar to the filtered or unfiltered population, yielded inconclusive results, showing inconsistencies when compared to the actual data distribution. This is likely due to the small sample size (<8) in many cases and the potential mixing of sampling protocols among the unknown samples. For these reasons, we decided to rely exclusively on data from confirmed filtered samples, which constitute 41% of the dataset available at this stage of the workflow.

The analysis of outliers in time series identified a total of 1,166 outlier values, with 46 samples exhibiting at least three outliers across different variables. These samples were labeled as anomalous and removed from the dataset. Among the remaining outliers, 398 (34% of all outliers and less than 1% of the total data) were classified as “extreme”, as they affected the representative statistics of their piezometer by more than 15%, and were subsequently removed from the dataset.

After completing all quality optimization steps (Fig. 2), the sample matrix included data from 86 sites, 993 piezometers, and 22,261 individual values. The step-by-step data reduction process described earlier in this Section is also summarized in the SM. After extracting a single representative value for each piezometer (according to the criteria outlined in Sect. "The proposed workflow"), a refined data matrix comprising 993 piezometers was obtained, constituting the anthropized dataset, which is expected to be representative of the current groundwater conditions, incorporating both natural and anthropogenic signals.

Removal of anthropogenic impact and the “natural dataset”

The analysis of temporal trends was performed on 1,965 out of 5,608 time series that met the MK test requirements (i.e., sample size greater than 4). Of these, 212 series also met the criteria for the SK test. The MK and SK tests identified 138 and 25 significant trends, respectively, totaling 163 trends linked to 111 piezometers. These piezometers were considered to be affected by anthropogenic contamination and were removed from the dataset.

Pre-selection based on site-specific index contaminants identified 293 piezometers where at least one analytical value exceeded the 75% of the regulatory limit. These piezometers were considered influenced by anthropogenic activities and were excluded from the dataset. The most common contaminants were Total Hydrocarbons (THC), Total Organohalogenated compounds (TOH) and Vinyl Chloride Monomer (VCM), exceeding respectively in 27%, 4% and 3% of the initial 993 piezometers from the anthropized dataset.

The analysis of spatial outliers at the site scale detected a total of 146 local anomalies, 101 of which could not be attributed to natural variation after the comparisons with upgradient piezometers. These 101 anomalous piezometers were classified as anthropogenically influenced and removed from the dataset.

After the entire cleaning process (Fig. 2), which resulted in the removal of 505 piezometers, 5 of them (0.99% of total removed) were local anomalies not validated by the position of upgradient piezometers (due to unavailable information), trend analysis (due to insufficient sample size), or PS (due to missing measurements for index contaminants). The remaining 99% of removed piezometers went through at least one of these steps.

The 488 piezometers remaining after filtering out those influenced by anthropogenic contamination constitute the natural dataset, which is expected to be representative of the natural groundwater composition.

The treatment workflow modified the statistical distribution of measured variables (Fig. 3, detailed in SM), reducing the number of outliers and shifting the median values to ranges consistent with previous studies conducted under similar environmental conditions (Cinti et al., 2023; Greggio et al., 2020; Orecchia et al., 2022; Rapti-Caputo and Martinelli, 2009).

Notably, despite the natural dataset retaining only 49% of the piezometers constituting the anthropized dataset, the overall spatial distribution remains largely unaffected (Fig. S4), indicating that the anthropogenic signal was removed in a spatially homogeneous manner across the sites.

Dataset testing

Factor analysis

The FA performed on the anthropized and natural datasets identified, in both cases, 3 significant factors (i.e., eigenvalues > 1), respectively called FA1-3 and FN1-3 (Table 1), explaining a total cumulative variance of 43%.

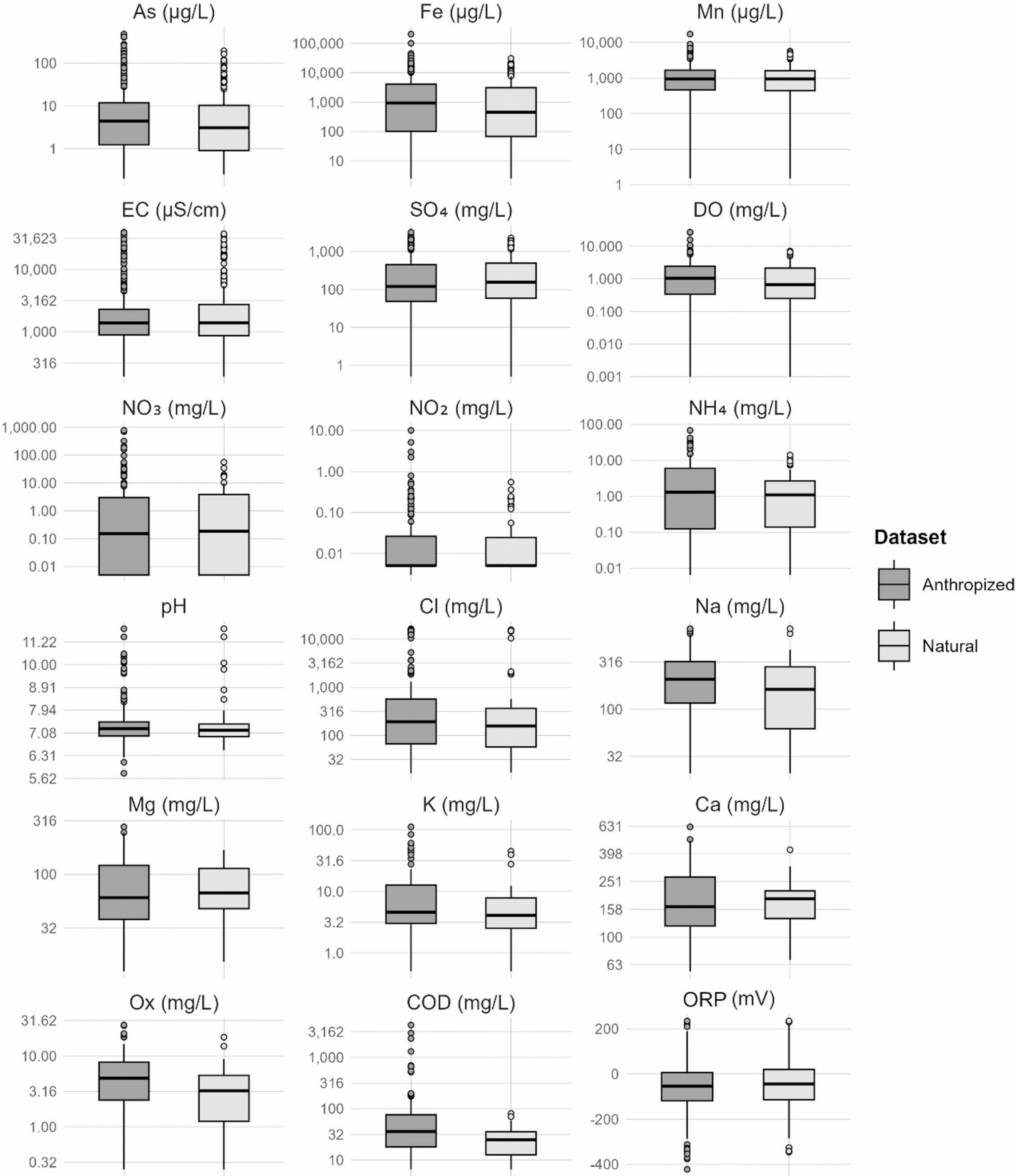

Fig. 3 Statistical summaries of the anthropized and natural dataset

For the anthropized dataset, FA1 was primarily represented by the positively correlated variables Fe and As (i.e., loading values >|0.3|), suggesting the release of metals in groundwater under reducing conditions. FA2 was mainly associated with pH, likely indicating some specific anomalies affecting groundwater acidity. FA3 was strongly represented by the negatively correlated variables EC and ORP, likely indicating brackish conditions related to the influence

Table 1 Loading values resulting from the factor analysis on the anthropized and natural dataset (values highlighted in bold >|0.3|)

Dataset	Anthropized			Natural		
Factors	FA1	FA2	FA3	FN1	FN2	FN3
Variable						
Fe	**1**	−0.01	0	0.05	**0.81**	−0.03
Mn	0.27	−0.01	0.12	−0.23	**0.35**	0.17
As	**0.31**	0.14	−0.14	**0.99**	0.02	0.01
EC	−0.02	−0.02	**−0.57**	0.01	−0.06	**−0.49**
ORP	−0.01	0	**0.68**	0.02	−0.03	**0.88**
pH	0	**1**	0	0.05	−0.21	0.06
DO	0.02	0.16	0.17	−0.17	0.26	0.25

of seawater. The statistics of the analysis, with an RMSEA index of 0.086 and a TLI of 0.645, suggest low reliability (Hu and Bentler, 1999). This is likely due to the overlap of natural and anthropogenic signals, which complicates the identification of variable correlations.

For the natural dataset, FN1 was primarily represented by As, indicating groundwater with strongly reducing conditions. FN2 was predominantly represented by Fe and Mn, suggesting moderately reducing conditions. FN3 was represented by the negatively correlated variables EC and ORP, again indicating brackish conditions. In this case, the statistics of the analysis indicated high reliability, with an RMSEA index of 0.04 and a TLI of 0.938.

Cluster analysis

The CA was performed on the natural dataset, which gave the most conclusive results in terms of FA. Through the combined CA on the obtained factor scores, we identified 6 clusters (C1-C6, details in the SM) each indicating a hydrogeochemical facies related to different redox zones (Table 2; Fig. 4). A median Mn concentration consistently exceeding 320 µg/L across all clusters suggests that multiple factors prevent the groundwater from reaching strictly oxidative conditions (Close et al., 2016). Instead, the clusters C1 to C5 seem to describe a range from slightly to strongly reducing conditions, while cluster C6 represents brackish groundwater near the coast (median EC of 20,865 µS/cm is one order of magnitude greater than that of the other clusters), aligning with previous studies focused on the influence of seawater intrusion in the Po Plain aquifers (Antonellini et al., 2008; Colombani and Mastrocicco, 2016; Orecchia et al., 2022). More in detail, at less reducing conditions of cluster C1 (median Mn, Fe and As of 1,489 µg/L, 2.5 µg/L and 2.6 µg/L, respectively), Mn serves as the primary electron acceptor and is reduced into the aqueous phase, while Fe and As remain stable in the solid phase (Hamer et al., 2020). As redox processes transition to proper Mn/Fe-reduction, Fe becomes unstable and is released into groundwater (Neidhardt et al., 2014), leading to increasing

Table 2 Median values of the original variables for each cluster obtained from the cluster analysis applied to the factor analysis scores of the natural dataset. Empty cells denote the absence of available data

Variable	Cluster					
	C1	C2	C3	C4	C5	C6
Fe (µg/L)	2.5	227.0	607.0	13350.0	6400.0	103.9
Mn (µg/L)	1489.0	844.8	867.0	1960.0	371.0	269.4
As (µg/L)	2.6	2.5	2.8	16.3	75.1	0.3
ORP (mV)	223.5	−2.9	−131.2	−93.9	−123.0	−133.1
EC (µS/cm)	829.0	1227.0	1701.0	2632.6	2286.0	20865.0
pH	7.1	7.3	7.1	6.9	7.1	7.1
SO_4 (mg/L)	-	129.4	145.0	114.0	27.6	1693.6
NO_3 (mg/L)	0.01	1.99	0.21	0.01	-	-
NO_2 (mg/L)	0.005	0.005	0.005	0.008	0.020	0.005
NH4 (mg/L)	0.1	1.6	0.2	1.1	-	-
DO (mg/L)	0.1	1.7	0.2	2.8	0.3	0.3
Ox (mg/L)	-	2.4	3.3	4.4	-	-
COD (mg/L)	38.7	17.8	8.0	25.3	-	-
Cl (mg/L)	-	110.3	67.5	305.9	-	-
Mg (mg/L)	101.0	65.2	25.1	114.0	-	-
Ca (mg/L)	-	179.0	70.0	232.5	-	-
K (mg/L)	-	4.2	36.4	4.0	-	-
Na (mg/L)	-	150.0	630.0	224.0	-	-

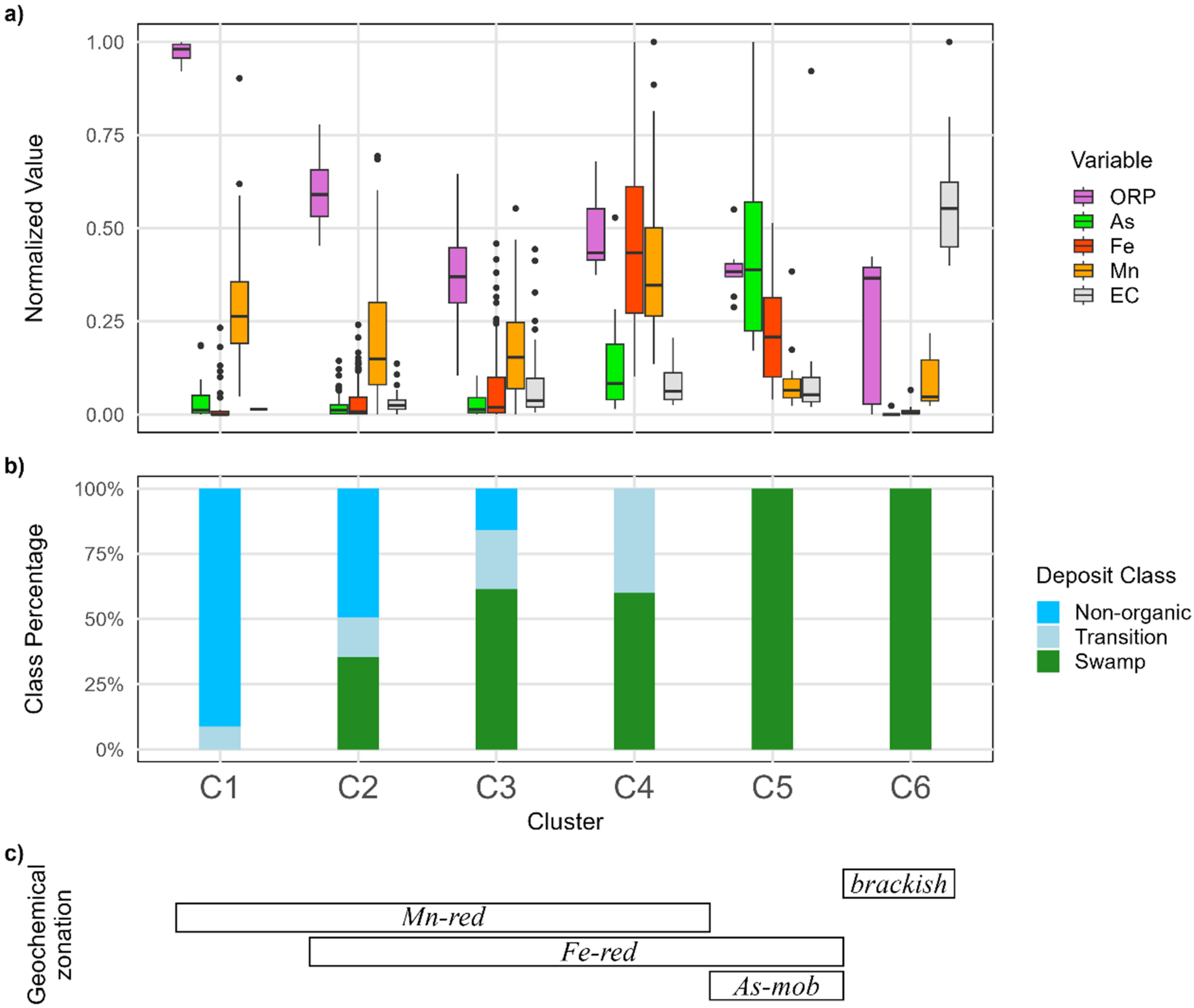

Fig. 4 Hydrogeochemical zonation. (**a**) Boxplots of normalized values for key variables across the clusters. (**b**) Percentage of piezometers correlated with stratigraphic facies associations for each cluster. (**c**) Interpreted hydrogeochemical zones (Mn reduction, Fe reduction, As mobilization and brackish groundwater)

concentrations. Initially, Mn concentrations still exceed Fe concentrations in clusters C2 (median Mn, Fe and As of 844.8 µg/L, 227 µg/L and 2.5 µg/L, respectively) and C3 (median Mn, Fe and As of 867 µg/L, 607 µg/L and 2.8 µg/L, respectively), indicating Mn-reduction as the predominant process (Ying et al., 2017). As the conditions become more reducing, the initially predominant Mn-reduction process is overlapped by stronger Fe-reduction in cluster C4 (median Fe, Mn and As of 13,350 µg/L, 1,960 µg/L and 16.3 µg/L respectively), where Fe and Mn reach peak concentrations. Eventually, more reducing conditions observed in cluster C5 lead to the mobilization of As in groundwater (McArthur et al., 2004), reaching a peak median concentration of 75 µg/L.

Additionally, we observed that cluster C5 is also characterized by lower SO_4 concentration (Table 2). Although SO_4 was not included as a direct input variable in the multivariate analysis due to insufficient measurements, we consider it a key indicator of additional natural processes. Specifically, this SO_4 depletion may suggest the occurrence of sulfate reduction under advanced reducing conditions, potentially influencing arsenic release in groundwater (see SM for further details). The interactions between SO_4 and As during reducing processes can be complex: sulfate reduction processes may promote As attenuation or mobilization in groundwater depending on many factors and environmental conditions (Gao et al., 2021; Luo et al., 2025; Wang et al., 2017). According to Kao et al. (2011), in aquifers influenced by seawater (such as our study area), strong As mobilization can occur alongside sulfate reduction due to the co-occurrence of multiple geochemical processes associated with the sulfur cycling.

Our hydrogeochemical model (Fig. 4) aligns with previous studies conducted in similar environments (Carraro et al., 2013, 2015; Dalla Libera et al., 2020; A. Molinari et al., 2012; Rotiroti et al., 2014, 2021; Sracek et al., 2018), suggesting that the observed processes reflect the natural geochemical evolution of groundwater in the area.

Correlation with stratigraphic facies associations

The comparison with stratigraphic facies associations was performed for 54% of clustered piezometers with available stratigraphic log information (172 piezometers). A clear correlation was observed between facies associations and redox zonation, with major occurrence of swamp deposits over non-organic deposits with increasing reducing conditions from cluster C1 to C5 (Fig. 4). This is consistent with recent studies on sedimentary and stratigraphic controls influencing aquifer geochemistry in the Po Plain (Amorosi and Sammartino, 2024; Bosi et al., 2025; Carraro et al., 2015) and in similar settings worldwide (Buschmann et al., 2007; Desbarats et al., 2014; Kazmierczak et al., 2022; Weinman et al., 2008).

In detail, the C1 cluster, representing the slightest reducing conditions, is predominantly (91%) associated with non-organic deposits, with a smaller proportion (9%) linked to transition deposits. The C2 cluster shows a 50% correlation with non-organic deposits, 15% with transition deposits, and 35% with swamp deposits. The C3 cluster is correlated with 16% of non-organic, 23% of transition, and 61% of swamp deposits. The C4 cluster shows a 40% correlation with transition and 60% with swamp deposits. The C5 cluster, which represents the strongest reducing conditions in the area, is entirely (100%) correlated with swamp deposits.

Moreover, the spatial distribution of redox clusters C1-C5 is consistent with the shallow stratigraphic architecture of the study area: in the southwestern (SW) sector, where the shallow A0 aquifer is mostly embedded in alluvial deposits of Apennine origin relatively poor in organic matter, only piezometers pertaining to clusters C1-C3 occur (Fig. 5). The wedge-shaped body of swamp deposits extending across the study area is found here relatively deeper and thinner (Bruno et al., 2022), making it less likely to be intersected by the shallow piezometers. In the northeastern (NE) sector, where the A0 aquifer is mostly encased in deposits of the modern

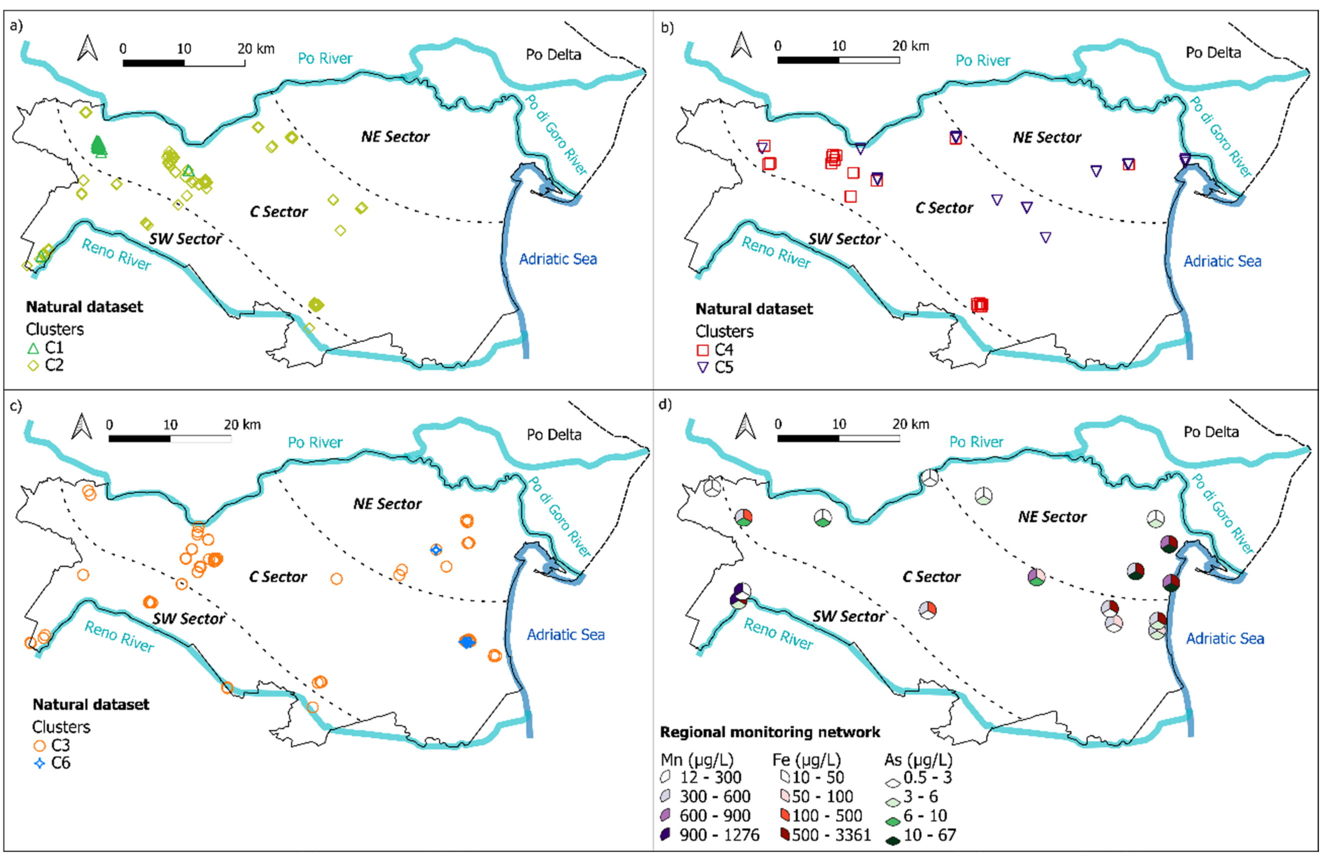

Fig. 5 Spatial distribution of clusters across the southwestern (SW), central (C) and northeastern (NE) sectors (**a**, **b**, **c**). Concentrations of Mn, Fe and As from the regional monitoring network are reported for comparison (**d**)

delta plain, and swamp sediments rich in organic matter are encountered at shallow depths (Bruno et al., 2017), only piezometers pertaining to clusters C3-C5 occur (Fig. 5). In the central (C) sector, groundwater exhibits all identified redox clusters (C1-C5), with an irregular distribution reflecting the highly heterogeneous geological setting of this area (Fig. 5). The A0 aquifer in this region is primarily encased within the abandoned Po delta lobe deposits, overlapped by alluvial deposits, or outcropping coastal deposits (Bruno et al., 2017). Depending on filter depth, piezometers may capture groundwater from upper alluvial sediments or from swamp deposits enriched in organic matter.

Finally, the C6 cluster, indicative of brackish groundwater conditions, is exclusively associated with swamp deposits. However, specific redox conditions cannot be determined in this case, as the few piezometers in C6 linked to stratigraphic information lack metal concentration data. The brackish groundwater cluster likely represents mixed redox conditions. Nonetheless, its location in the eastern part of the study area increases the likelihood of intersecting deltaic deposits.

Comparison of the aggregated dataset with the regional monitoring network

In the study area, 18 monitoring wells of the regional groundwater quality monitoring network managed by ARPA tap into the A0 aquifer (Fig. 1). In accordance with the Water Framework Directive (*WFD*, *Dir. 2000/60/EC*) transposed into national legislation (Legislative Decree *D. lgs n. 152/2006)*, the regional monitoring program has been collecting groundwater samples biannually since 2010. Data from the regional network, covering the period between 2010 and 2023, consist of 361 samples. The regional network follows European and national directives for assessing groundwater quality in areas unaffected by local anthropogenic influences and can therefore be considered representative of the natural groundwater composition.

We compared this "validated" dataset with our findings, in order to assess the reliability of our insights into the groundwater chemical status of the study area. The small sample size (18 piezometers) prevents the regional dataset from undergoing advanced statistical analysis (de Winter et al., 2009; Hogarty et al., 2005; MacCallum et al., 1999, 2001). Therefore, we focused on observing the spatial distribution of key chemical indicators for redox status (i.e., Fe, Mn and As; Fig. 5). In the southwestern sector, we observed the highest Mn concentration (up to 1,276 µg/L), moderate Fe concentrations (25 µg/L to 1,660 µg/L) and low concentrations of As (<6 µg/L). Conversely, in the northeastern sector, the highest Fe concentrations (up to 3,300 µg/L) and As concentrations (up to 66 µg/L) coincide with moderate Mn concentrations (70 µg/L to 720 µg/L).

This comparison indicates that the regional monitoring network aligns with the geochemical zonation identified in this study, confirming the reliability of our workflow in generating a dataset with meaningful insights into the natural groundwater composition of the shallow aquifer, despite originating from aggregated data from contaminated sites. It is important to note that our geochemical model offers a more detailed depiction of groundwater geochemical heterogeneity at the meso-scale. In fact, we observed a significant smaller-scale variability, highlighted by the close proximity of piezometers associated with different redox zones in several areas. While the regional network provides a valuable overview of the general groundwater composition, its limited spatial resolution prevents it from capturing this smaller-scale natural heterogeneity. Additionally, the larger size of our aggregated dataset enabled more advanced statistical analyses compared to the regional network, enhancing the robustness and reliability of our findings.

Conclusions

This study demonstrates the effectiveness of the proposed treatment workflow for aggregating and repurposing available monitoring datasets as a reliable basis for investigating the natural composition of shallow groundwater. The main conclusions can be summarized as follows:

Methodological effectiveness. The proposed workflow successfully extracted information on the natural composition of shallow groundwater from heterogeneous monitoring datasets originally affected by human activities. Its validity is confirmed through consistency with both the stratigraphic architecture of the area and the regional groundwater quality monitoring network. While the removal of all anomalies is not guaranteed, the presence of a few outliers is outweighed by the significant volume of reliable data retrieved. The workflow is designed to be as objective and reproducible as possible (balancing a broadly applicable framework with the integration of site-specific knowledge), while remaining adaptable to expert knowledge for fine-tuning, such as selective outlier removal tailored to local contexts.

Hydrochemical insights. Multivariate analysis on the processed data provided key insights into the hydrochemical composition of shallow groundwater in the Ferrara province, leading to an improved conceptual model of the A0 aquifer. The study revealed redox-driven natural processes, closely linked to the stratigraphic architecture of the region, that explain elevated concentrations of As, Fe, and Mn, reaching up to 75 µg/L, 13,350 µg/L, and 1,960 µg/L, respectively.

This effect is particularly pronounced in the northeastern sector, near the Po River delta, where the strongest reducing conditions are observed. Moving southwestward, the effect gradually diminishes due to the deepening and fragmentation of the organic matter-enriched swamp deposits.

Practical implications. Although more complex than conventional methods for assessing groundwater composition, this method enables the strategic use of unconventional data that would otherwise remain largely unused, offering a sustainable solution to data scarcity. It supports practical applications such as assessing the natural background levels (NBLs) of specific contaminants, a critical step required by international and national water quality directives. Indeed, one of the major challenges in determining NBLs is distinguishing between natural and anthropogenically influenced groundwater. Our methodology successfully tackles this issue, making it a valuable tool for both contaminant-specific assessments and broader groundwater quality studies, providing a foundation for multi-scale groundwater resource management.

In summary, the proposed workflow provides a scientifically sound and operationally practical approach for characterizing natural groundwater systems leveraging existing monitoring data, addressing the issue of data scarcity and supporting sustainable groundwater management without requiring significant new data collection.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12665-025-12784-2.

Acknowledgements This work was supported by Emilia-Romagna Region (RER) and the Regional Environmental Protection Agency (ARPAE). We would like to acknowledge Igor Villani (RER), Cristina Baroni (RER), Marco Marcaccio (ARPAE), Marco Roverati (ARPAE), Gabriella Dugoni (ARPAE) and other employees, who allowed the investigation by providing their support on data collection and interpretation.

Author contributions Conceptualization: L.L., M.R., C.Z., A.A., E.D., M.F.; Methodology: L.L., M.R., M.F.; Formal analysis: L.L., M.R., C.Z., M.F.; Data curation: L.L.; Writing—original draft preparation: L.L.; Writing—review and editing: M.R., C.Z., A.A., E.D., M.F.; Visualization: L.L.; Supervision: M.F. All authors read and approved the final manuscript.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability Data are available in a publicly accessible repository at: https://doi.org/10.5281/zenodo.15125732

Declarations

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

Competing interests The authors declare no competing interests.

References

Abu M, Egbueri JC, Agbasi JC (2025) Kriging-interpolated mapping and predictive modeling of groundwater F − and NO3 − contamination with chemometric and health risk assessments in Ghana's Birimian Province. Environ Geochem Health. https://doi.org/10.1007/s10653-025-02453-5

Amorosi A, Sammartino I (2024) Predicting natural arsenic enrichment in peat-bearing, alluvial and coastal depositional systems: a generalized model based on sequence stratigraphy. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2024.171571

Amorosi A, Bruno L, Campo B, Morelli A, Rossi V, Scarponi D, Hong W, Bohacs KM, Drexler TM (2017) Global sea-level control on local parasequence architecture from the Holocene record of the Po Plain, Italy. Mar Pet Geol 87:99–111. https://doi.org/10.1016/j.marpetgeo.2017.01.020

Amorosi A, Guermandi M, Marchi N, Sammartino I (2014) Fingerprinting sedimentary and soil units by their natural metal contents: a new approach to assess metal contamination. Sci Total Environ 500–501:361–372. https://doi.org/10.1016/j.scitotenv.2014.08.078

Antonellini M, Mollema P, Giambastiani B, Bishop K, Caruso L, Minchio A, Pellegrini L, Sabia M, Ulazzi E, Gabbianelli G (2008) Salt water intrusion in the coastal aquifer of the southern Po Plain, Italy. Hydrogeol J 16(8):1541–1556. https://doi.org/10.1007/s10040-008-0319-9

Appelo CAJ, Postma D (2005) Geochemistry, groundwater and pollution (2nd ed.). CRC Press

ARPA. (2013). Direzione Tecnica. Approvazione della Circolare interna recante la Linea Guida per l'esecuzione delle attività di controllo in materia di siti contaminati.

Baker J, White N, Mengersen K (2014) Missing in space: an evaluation of imputation methods for missing data in spatial analysis of risk factors for type II diabetes. Int J Health Geogr. https://doi.org/10.1186/1476-072X-13-47

Benavides Höglund N, Sparrenbom C, Hugman R (2024) Assimilation and value of injection response data for enhanced contaminated site characterization. Hydrogeol J. https://doi.org/10.1007/s10040-024-02854-3

Berg M, Trang PTK, Stengel C, Buschmann J, Viet PH, Van Dan N, Giger W, Stüben D (2008) Hydrological and sedimentary controls leading to arsenic contamination of groundwater in the Hanoi area, Vietnam: the impact of iron-arsenic ratios, peat, river bank deposits, and excessive groundwater abstraction. Chem Geol 249(1–2):91–112. https://doi.org/10.1016/j.chemgeo.2007.12.007

Bocianowski J, Wrońska-Pilarek D, Krysztofiak-Kaniewska A, Matusiak K, Wiatrowska B (2024) Comparison of Pearson's and Spearman's correlation coefficients for selected traits of *Pinus* *sylvestris* L. Biom Lett 61(2):115–135. https://doi.org/10.2478/bile-2024-0008

Bosi E, Colombera L, Mountney NP, Bertoni D, Sarti G, Amorosi A (2025) Sedimentary controls on arsenic distribution in meander-belt deposits of the Po Valley, Italy. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2025.178627

Bruno L, Bohacs KM, Campo B, Drexler TM, Rossi V, Sammartino I, Scarponi D, Hong W, Amorosi A (2017) Early Holocene transgressive palaeogeography in the Po coastal plain (northern Italy). Sedimentology 64(7):1792–1816. https://doi.org/10.1111/sed.12374

Bruno L, Campo B, Di Martino A, Hong W, Amorosi A (2019) Peat layer accumulation and post-burial deformation during the mid-late Holocene in the Po coastal plain (Northern Italy). Basin Res 31(3):621–639. https://doi.org/10.1111/bre.12339

Bruno L, Campo B, Hajdas I, Hong W, Amorosi A (2022) Timing and mechanisms of sediment accumulation and pedogenesis: insights from the Po Plain (northern Italy). Palaeogeogr Palaeoclimatol Palaeoecol. https://doi.org/10.1016/j.palaeo.2022.110881

Bulut OF, Duru B, Çakmak Ö, Günhan Ö, Dilek FB, Yetis U (2020) Determination of groundwater threshold values: a methodological approach. J Clean Prod. https://doi.org/10.1016/j.jclepro.2020.120001

Burt TP, Howden NJK, Worrall F (2014) On the importance of very long-term water quality records. Wiley Interdisciplinary Reviews: Water 1(1):41–48. https://doi.org/10.1002/wat2.1001

Buschmann J, Berg M, Stengel C, Sampson ML (2007) Arsenic and manganese contamination of drinking water resources in Cambodia: coincidence of risk areas with low relief topography. Environ Sci Technol 41(7):2146–2152. https://doi.org/10.1021/es062056k

Busico G, Bordbar M, Rufino F, Sarracino A, Tedesco D (2024) Assessment of NO3−, As, and F−background levels in groundwater bodies: A methodological review and case study utilizing sequential Gaussian simulation (SGS). Groundwater for Sustainable Development, 26. Elsevier B.V. https://doi.org/10.1016/j.gsd.2024.101211

Campo B, Bruno L, Amorosi A (2020) Basin-scale stratigraphic correlation of late Pleistocene-Holocene (MIS 5e-MIS 1) strata across the rapidly subsiding Po Basin (northern Italy). Quat Sci Rev. https://doi.org/10.1016/j.quascirev.2020.106300

Carraro A, Fabbri P, Giaretta A, Peruzzo L, Tateo F, Tellini F (2013) Arsenic anomalies in shallow Venetian Plain (Northeast Italy) groundwater. Environ Earth Sci 70(7):3067–3084. https://doi.org/10.1007/s12665-013-2367-2

Carraro A, Fabbri P, Giaretta A, Peruzzoa L, Tateo F, Tellini F (2015) Effects of redox conditions on the control of arsenic mobility in shallow alluvial aquifers on the Venetian Plain (Italy). Sci Total Environ 532:581–594. https://doi.org/10.1016/j.scitotenv.2015.06.003

Carroll JB (1953). An analytical solution for approximating simple structure in factor analysis. Psychometrika 18.

Cauich-Kau D, Castro-Larragoitia J, Cardona Benavides A, García-Arreola ME, García-Vargas GG (2025) An adapted groundwater quality index including toxicological critical pollutants. Groundwater for Sustainable Development. https://doi.org/10.1016/j.gsd.2024.101401

Cheng CH, Huang SF (2021) A novel clustering-based purity and distance imputation for handling medical data with missing values. Soft Comput 25(17):11781–11801. https://doi.org/10.1007/s00500-021-05947-3

Cinti D, Sciarra A, Cantucci B, Galli G, Pizzino L, Procesi M, Poncia PP (2023). Hydrogeochemical investigation of shallow aquifers before and after the 2012 Emilia seismic sequence (northern Italy). Appl Geochem 151. https://doi.org/10.1016/j.apgeochem.2023.105624

Close ME, Abraham P, Humphries B, Lilburne L, Cuthill T, Wilson S (2016) Predicting groundwater redox status on a regional scale using linear discriminant analysis. J Contam Hydrol 191:19–32. https://doi.org/10.1016/j.jconhyd.2016.04.006

Colombani N, Mastrocicco M (2016) Geochemical evolution and salinization of a coastal aquifer via seepage through peaty lenses. Environ Earth Sci. https://doi.org/10.1007/s12665-016-5642-1

Dalla Libera N, Fabbri P, Mason L, Piccinini L, Pola M (2017) Geostatistics as a tool to improve the natural background level definition: an application in groundwater. Sci Total Environ 598:330–340

Dalla Libera N, Pedretti D, Tateo F, Mason L, Piccinini L, Fabbri P (2020) Conceptual model of arsenic mobility in the shallow alluvial aquifers near Venice (Italy) elucidated through machine learning and geochemical modeling. Water Resour Res 56(9):0–3. https://doi.org/10.1029/2019WR026234

de Winter JCF, Dodou D, Wieringa PA (2009) Exploratory factor analysis with small sample sizes. Multivar Behav Res 44(2):147–181. https://doi.org/10.1080/00273170902794206

Desbarats AJ, Koenig CEM, Pal T, Mukherjee PK, Beckie RD (2014) Groundwater flow dynamics and arsenic source characterization in an aquifer system of West Bengal, India. Water Resour Res 50(6):4974–5002. https://doi.org/10.1002/2013WR014034

Di Curzio D, Rusi S, Signanini P (2019) Advanced redox zonation of the San Pedro Sula alluvial aquifer (Honduras) using data fusion and multivariate geostatistics. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2019.133796

Dixit M, Arora M, Jayaprakash HP, Achutha VR (2024) Achievements and significance of national hydrograph stations network and 55 years of continuous groundwater level monitoring in India. Groundw Sustain Dev. https://doi.org/10.1016/j.gsd.2024.101313

Edmunds WM (2009) Geochemistry's vital contribution to solving water resource problems. Appl Geochem 24(6):1058–1073. https://doi.org/10.1016/j.apgeochem.2008.11.018

Edmunds WM, Shand P (2008) Natural groundwater quality. Blackwell Publishing

Famiglietti JS (2014) The global groundwater crisis. Nat Clim Chang 4(11):945–948. https://doi.org/10.1038/nclimate2425

Filippini M, Zanotti C, Bonomi T, Sacchetti VG, Amorosi A, Dinelli E, Rotiroti M (2021) Deriving natural background levels of arsenic at the meso-scale using site-specific datasets: an unorthodox method. Water 13:452

Foster S, Hirata R, Gomes D, D'Elia M, Paris M (2022) Groundwater quality protection: a guide for water utilities, municipal authorities, and environment agencies. World Bank / WHO

Gaiolini M, Colombani N, Chierici V, Montanari L, Mastrocicco M (2024) Numerical modelling of groundwater level and salinity evolution in a low-lying coastal area under intensive agricultural activity: numerical modelling of groundwater level and salinity evolution in a low-lying coastal area under intensive agricultural activity. Water Resources Management. https://doi.org/10.1007/s11269-024-04044-y

Gałuszka A (2007) A review of geochemical background concepts and an example using data from Poland. Environ Geol 52(5):861–870. https://doi.org/10.1007/s00254-006-0528-2

Gao J, Zheng T, Deng Y, Jiang H (2021) Microbially mediated mobilization of arsenic from aquifer sediments under bacterial sulfate reduction. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2020.144709

Garzanti E, Vezzoli G, Andò S (2011) Paleogeographic and paleodrainage changes during Pleistocene glaciations (Po Plain, Northern Italy). Earth-Sci Rev. https://doi.org/10.1016/j.earscirev.2010.11.004

Ghiglieri G, Pittalis D, Cerri G, Oggiano G (2012) Hydrogeology and hydrogeochemistry of an alkaline volcanic area: the NE Mt. Meru slope (East African Rift-Northern Tanzania). Hydrol Earth Syst Sci. https://doi.org/10.5194/hess-16-529-2012

Giang N. Van, Thanh LN, Hiep VQ, Hida N (2014) Hydrological and hydrogeological characterization of groundwater and river water in the North Hanoi industrial area, Vietnam. Environ Earth Sci. https://doi.org/10.1007/s12665-014-3086-z

Greggio N, Giambastiani BMS, Mollema P, Laghi M, Capo D, Gabbianelli G, Antonellini M, Dinelli E (2020) Assessment of the main geochemical processes affecting surface water and groundwater in a low-lying coastal area: implications for water management. Water. https://doi.org/10.3390/W12061720

Güler C (2009) Site characterization and monitoring of natural attenuation indicator parameters in a fuel contaminated coastal aquifer: Karaduvar (Mersin, SE Turkey). Environ Earth Sci 59(3):631–643. https://doi.org/10.1007/s12665-009-0060-2

Guo H, Zhang BO, Yang S, Li Y, Stüben D, Norra S, Wang J (2009) Role of colloidal particles for hydrogeochemistry in as-affected aquifers of the Hetao Basin, Inner Mongolia. Geochem J 43.

Hamer K, Gudenschwager I, Pichler T (2020) Manganese (Mn) concentrations and the mn-fe relationship in shallow groundwater: implications for groundwater monitoring. Soil Syst 4(3):1–19. https://doi.org/10.3390/soilsystems4030049

Hewett P, Ganser GH (2007) A comparison of several methods for analyzing censored data. Ann Occup Hyg 51(7):611–632. https://doi.org/10.1093/annhyg/mem045

Hinsby K, de Consso Melo MT, Dahl M (2008) European case studies supporting the derivation of natural background levels and groundwater threshold values for the protection of dependent ecosystems and human health. Sci Total Environ 401(1–3):1–20. https://doi.org/10.1016/j.scitotenv.2008.03.018

Hirsch RM, Slack JR, Smith RA (1982) Techniques of trend analysis for monthly water quality data. Water Resour Res 18(1).

Hogarty KY, Hines CV, Kromrey JD, Perron JM, Mumford AKR (2005) The quality of factor solutions in exploratory factor analysis: the influence of sample size, communality, and overdetermination. Educ Psychol Meas 65(2):202–226. https://doi.org/10.1177/0013164404267287

Hornung RW, Reed LD (1990) Estimation of average concentration in the presence of nondetectable values. Appl Occup Environ Hyg 5:46–51

Hu LT, Bentler PM (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct. Equ Model 6(1):1–55. https://doi.org/10.1080/10705519909540118

ISPRA (2018) Linee Guida per la Determinazione dei Valori di Fondo. ISPRA. www.isprambiente.gov.it

ISS (2008) Metodiche di pretrattamento di campioni di acque di falda prelevati in siti contaminati.

Judd AG (1980) The use of cluster analysis in the derivation of geotechnical classifications. Bull Assoc Eng Geo 27:193–211.

Kaiser HF (1960) The application of electronic computers to factor analysis. Educ Psychol. Meas 20:141–151.

Kao YH, Wang SW, Liu CW, Wang PL, Wang CH, Maji SK (2011) Biogeochemical cycling of arsenic in coastal salinized aquifers: evidence from sulfur isotope study. Sci Total Environ 409(22):4818–4830. https://doi.org/10.1016/j.scitotenv.2011.08.003

Kazmierczak J, Postma D, Dang T, Hoang H. Van, Larsen F, Hass AE, Hoffmann AH, Fensholt R, Pham NQ, Jakobsen R (2022) Groundwater arsenic content related to the sedimentology and stratigraphy of the Red River delta, Vietnam. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2021.152641

Kendall MG (1948). Rank correlation methods. Griffin, London.

Knaack D, Ivey G, Macphee C, Peterzon J, Price L, Williams M, Leybourne MI (2021) The effect of filtration size on the geochemistry of groundwater samples from a massive sulfide deposit at the Bathurst Mining Camp, New Brunswick, Canada. Geochem Explor Environ Anal 21(4). https://doi.org/10.1144/geochem2021-057

Kumanova X, Leka G, Nilsson B, Jacks G (2015) Assessing the mobility of metals in an aquatic environment: River Fani and River Mati, Albania. Environ Earth Sci 74(7):6293–6301. https://doi.org/10.1007/s12665-015-4653-7

Lakshminarayan K, Harp SA, Samad T (1999) Imputation of missing data in industrial databases. Appl Intell. https://doi.org/10.1023/a:1008334909089

Li J, Heap AD (2014) Spatial interpolation methods applied in the environmental sciences: a review. Environ Model Softw 53:173–189. https://doi.org/10.1016/j.envsoft.2013.12.008

Li X, Liang G, He B, Ning Y, Yang Y, Wang L, Wang G (2025) Recent advances in groundwater pollution research using machine learning from 2000 to 2023: A bibliometric analysis. Environ Res 267. Academic Press Inc. https://doi.org/10.1016/j.envres.2024.120683

Lindberg RD, Runnells DD (1984) Ground water redox reactions: an analysis of equilibrium state applied to Eh measurements and geochemical modeling. Science 225(4665):925–927. https://doi.org/10.1126/science.225.4665.925

Luo R, Cheng Y, Song, T, Wu Z, Wang Y, Li P (2025) Contrasting kinetics of arsenic release from As-bearing ferrihydrite coprecipitates by distinct sulfate-reducing bacteria. Ecotoxicol. Environ. Saf 293. https://doi.org/10.1016/j.ecoenv.2025.118035

MacCallum RC, Widaman KF, Preacher KJ, Hong S (2001) Sample size in factor analysis: the role of model error. Multivar Behav Res 36(4):611–637. https://doi.org/10.1207/S15327906MBR3604_06

MacCallum RC, Widaman KF, Zhang S, Hong S (1999) Sample size in factor analysis. Psychol Methods 4(1):84–99. https://doi.org/10.1037/1082-989X.4.1.84

Mann HB (1945) Nonparametric tests against trend. Econometrica 13(3):245–259

Mann HB, Whitney DR (1947) On a test of wether one of two random variables is stochastically larger than the other. Ann Math Stat 18:50–60

Marchetto A (2024) Rkt: mann-kendall test, seasonal and regional kendall tests - R package version 1.7. https://CRAN.R-project.org/package=rkt.

Marsh HW, Hau KT, Wen Z (2004) In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Struct Equ Modeling 11(3):320–341. https://doi.org/10.1207/s15328007sem1103_2

McArthur JM, Banerjee DM, Hudson-Edwards KA, Mishra R, Purohit R, Ravenscroft P, Cronin A, Howarth RJ, Chatterjee A, Talukder T, Lowry D, Houghton S, Chadha DK (2004) Natural organic matter in sedimentary basins and its relation to arsenic in anoxic ground water: the example of West Bengal and its worldwide implications. Appl Geochem 19(8):1255–1293. https://doi.org/10.1016/j.apgeochem.2004.02.001

McGill R, Tukey JW, Larsen WA (1978) Variations of box plots. Am Stat 32(1):12–16. https://doi.org/10.1080/00031305.1978.10479236

McLeod AI (2022) Kendall: Kendall rank correlation and mann-kendall trend test - r package version 2.2.1. https://CRAN.R-project.org/package=Kendall.

Menafoglio A, Guadagnini L, Guadagnini A, Secchi P (2021) Object oriented spatial analysis of natural concentration levels of chemical species in regional-scale aquifers. Spat Stat 43:100494. https://doi.org/10.1016/j.spasta.2021.100494

Mendizabal I, Baggelaar PK, Stuyfzand PJ (2012) Hydrochemical trends for public supply well fields in The Netherlands (1898–2008), natural backgrounds and upscaling to groundwater bodies.

J Hydrol 450–451:279–292. https://doi.org/10.1016/j.jhydrol.2012.04.050
Mirzaei R, Sakizadeh M (2016) Comparison of interpolation methods for the estimation of groundwater contamination in Andimeshk-Shush Plain, Southwest of Iran. Environ Sci Pollut Res 23(3):2758–2769. https://doi.org/10.1007/s11356-015-5507-2
Molinari A, Guadagnini L, Marcaccio M, Guadagnini A (2012) Natural background levels and threshold values of chemical species in three large-scale groundwater bodies in Northern Italy. Sci Total Environ 425:9–19. https://doi.org/10.1016/j.scitotenv.2012.03.015
Molinari A, Guadagnini L, Marcaccio M, Guadagnini A (2019) Geostatistical multimodel approach for the assessment of the spatial distribution of natural background concentrations in large-scale groundwater bodies. Water Res 149:522–532. https://doi.org/10.1016/j.watres.2018.09.049
Molinari FC, Boldrini G, Severi P, Dugoni G, Caputo DR, Martinelli G (2007) Risorse Idriche Sotterranee della Provincia di Ferrara.
Mukherjee A, Coomar P, Sarkar S, Johannesson KH, Fryar AE, Schreiber ME, Ahmed KM, Alam MA, Bhattacharya P, Bundschuh J, Burgess W, Chakraborty M, Coyte R, Farooqi A, Guo H, Ijumulana J, Jeelani G, Mondal D, Nordstrom DK, Podgorski Joel, Polya David A., Scanlon Bridget R., Shamsudduha Mohammad, Tapia Joseline, Vengosh A (2024) Arsenic and other geogenic contaminants in global groundwater. Nat Rev Earth Environ 5(4):312–328. https://doi.org/10.1038/s43017-024-00519-z
Muller D, Blum A, Hart A, Hookey J, Kunkel R, Scheidleder A, Tomlin C, Wendland F (2006) BRIDGE - background criteria for the identification of groundwater Thresholds.
Neidhardt H, Berner ZA, Freikowski D, Biswas A, Majumder S, Winter J, Gallert C, Chatterjee D, Norra S (2014) Organic carbon induced mobilization of iron and manganese in a West Bengal aquifer and the muted response of groundwater arsenic concentrations. Chem Geol 367:51–62. https://doi.org/10.1016/j.chemgeo.2013.12.021
Orecchia C, Giambastiani BMS, Greggio N, Campo B, Dinelli E (2022) Geochemical characterization of groundwater in the confined and unconfined aquifers of the Northern Italy. Appl Sci. https://doi.org/10.3390/app12157944
Parrone D, Ghergo S, Preziosi E (2019) A multi-method approach for the assessment of natural background levels in groundwater. Sci Total Environ 659:884–894. https://doi.org/10.1016/j.scitotenv.2018.12.350
Podgorski J, Berg M (2020) Global threath of arsenic in groundwater. Science 368:845–850
Rapti-Caputo D, Martinelli G (2009) The geochemical and isotopic composition of aquifer systems in the deltaic region of the Po River plain (northern Italy). Hydrogeol J 17(2):467–480. https://doi.org/10.1007/s10040-008-0370-6
Reimann C, Garrett RG (2005) Geochemical background—concept and reality. Sci Total Environ 350(1–3):12–27. https://doi.org/10.1016/J.SCITOTENV.2005.01.047
Reimann C, Filzmoser P, Garrett RG (2005) Background and threshold: critical comparison of methods of determination. Sci Total Environ 346(1–3):1–16. https://doi.org/10.1016/j.scitotenv.2004.11.023
Rotiroti M, Bonomi T, Sacchi E, McArthur JM, Jakobsen R, Sciarra A, Etiope G, Zanotti C, Nava V, Fumagalli L, Leoni B (2021) Overlapping redox zones control arsenic pollution in Pleistocene multi-layer aquifers, the Po Plain (Italy). Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2020.143646
Rotiroti M, Di Mauro B, Fumagalli L, Bonomi T (2015) COMPSEC, a new tool to derive natural background levels by the component separation approach: application in two different hydrogeological contexts in northern Italy. J Geochem Explor 158:44–54. https://doi.org/10.1016/j.gexplo.2015.06.017
Rotiroti M, Sacchi E, Fumagalli L, Bonomi T (2014) Origin of arsenic in groundwater from the multilayer aquifer in Cremona (Northern Italy). Environ Sci Technol 48(10):5395–5403. https://doi.org/10.1021/es405805v
Schiavo M, Giambastiani BMS, Greggio N, Colombani N, Mastrocicco M (2024) Geostatistical assessment of groundwater arsenic contamination in the Padana Plain. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2024.172998
Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15(1):72–101
Sracek O, Berg M, Müller B (2018) Redox buffering and de-coupling of arsenic and iron in reducing aquifers across the Red River Delta, Vietnam, and conceptual model of de-coupling processes. Environ Sci Pollut Res Int 25(16):15954–15961. https://doi.org/10.1007/s11356-018-1801-0
Stefani M, Vincenzi S (2005) The interplay of eustasy, climate and human activity in the late Quaternary depositional evolution and sedimentary architecture of the Po Delta system. Mar Geol 222–223(1–4):19–48. https://doi.org/10.1016/j.margeo.2005.06.029
Tang L.,Gao W, Lu Y, Tabelin CB, Liu J, Li H, Yang W, Tang C, Feng X, Jiang J, Xue S (2024) The formation of multi-metal(loid)s contaminated groundwater at smelting site: Critical role of natural colloids. J Hazard Mater 471. https://doi.org/10.1016/j.jhazmat.2024.134408
US EPA (2006) Data quality assessment: statistical methods for practitioners - EPA QA/G9-S. www.epa.gov.quality.
US EPA (2013) Field measurement of oxidation-reduction potential (ORP).
Wadoux AMJC, Minasny B, McBratney AB (2020) Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews 210. https://doi.org/10.1016/j.earscirev.2020.103359
Wagner BJ (1992) Simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modelling. Journal of Hydrology 35.
Wang J, Zeng XC, Zhu X, Chen X, Zeng X, Mu Y, Yang Y, Wang Y (2017) Sulfate enhances the dissimilatory arsenate-respiring prokaryotes-mediated mobilization, reduction and release of insoluble arsenic and iron from the arsenic-rich sediments into groundwater. J Hazard Mater 339:409–417. https://doi.org/10.1016/j.jhazmat.2017.06.052
Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244. https://doi.org/10.1080/01621459.1963.10500845
Washington BN, Groffman PM, Duncan JM, Band LE, Miller AJ (2024) Long-term changes in riparian connectivity and groundwater chemistry in an urban watershed. J Environ Qual. https://doi.org/10.1002/jeq2.20654
Weinman B, Goodbred SL, Zheng Y, Aziz Z, Steckler M, van Geen A, Singhvi AK, Nagar YC (2008) Contributions of floodplain stratigraphy and evolution to the spatial patterns of groundwater arsenic in Araihazar, Bangladesh. Bull Geol Soc Am 120:1567–1580. https://doi.org/10.1130/B26209.1
Wilcoxon F (1945) Individual comparisons by ranking methods. Bull. https://doi.org/10.2307/3001968
World Health Organization (WHO) (2017) Guidelines for Drinking-water Quality (4th ed., incorporating 1st addendum). WHO Press.
Yadav A, Raj A, Yadav B (2024) Enhancing local-scale groundwater quality predictions using advanced machine learning approaches. J Environ Manage. https://doi.org/10.1016/j.jenvman.2024.122903
Yan G, Yang L, Hou Q (2024) Natural background levels and driving factors of aluminum in shallow groundwater of an urbanized delta: Insight from eliminating anthropogenic-impacted groundwaters. Water Environ Res 96(7). https://doi.org/10.1002/wer.11076

Yan G, Yang L, Liu L, Liu R, Huang G (2025) A regional investigation on natural background levels and contamination assessment of shallow groundwater contaminants in a coastal urbanized area. Water Environ Res 97(5). https://doi.org/10.1002/wer.70097

Yang S, Luo D, Tan J, Li S, Song X, Xiong R, Wang J, Ma C, Xiong H (2024) Spatial mapping and prediction of groundwater quality using ensemble learning models and SHapley additive explanations with spatial uncertainty analysis. Water. https://doi.org/10.3390/w16172375

Ying SC, Schaefer MV, Cock-Esteb A, Li J, Fendorf S (2017) Depth stratification leads to distinct zones of manganese and arsenic contaminated groundwater. Environ Sci Technol 51(16):8926–8932. https://doi.org/10.1021/acs.est.7b01121

Zahra A, Ali M, Ali N, Khan A, Zairov R, Sinyashin O, Wang Y, Zafar S, Khan FA (2025) A comprehensive analysis of the impact of arsenic, fluoride, and nitrate–nitrite dynamics on groundwater quality and its health implications. J Hazard Mater. https://doi.org/10.1016/j.jhazmat.2025.137093

Zanotti C, Caschetto M, Bonomi T, Parini M, Cipriano G, Fumagalli L, Rotiroti M (2022) Linking local natural background levels in groundwater to their generating hydrogeochemical processes in Quaternary alluvial aquifers. Sci Total Environ 805:150259. https://doi.org/10.1016/j.scitotenv.2021.150259

Zhang J, Xiao C, Yang W, Liang X, Zhang L, Wang X, Dai R (2024) Improving prediction of groundwater quality in situations of limited monitoring data based on virtual sample generation and Gaussian process regression. Water Res. https://doi.org/10.1016/j.watres.2024.122498