Science of the Total Environment 878 (2023) 163084

Contents lists available at ScienceDirect

Science of the Total Environment

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

Self-organizing map algorithm for assessing spatial and temporal patterns of pollutants in environmental compartments: A review

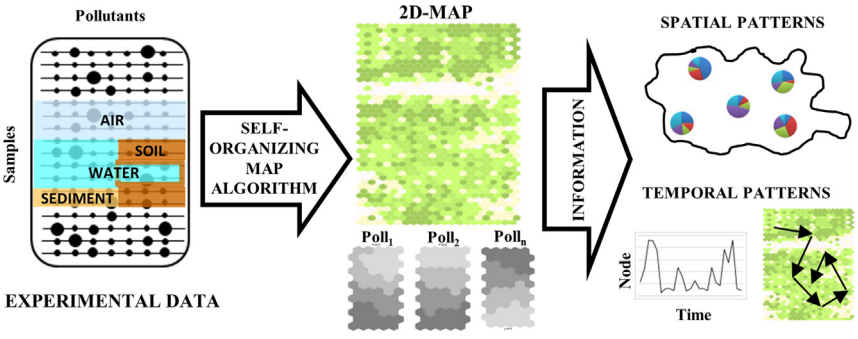
Sabina Licen [a,*], Aleksander Astel [b], Stefan Tsakovski [c]

[a] Department of Chemical and Pharmaceutical Sciences, University of Trieste, 34127 Trieste, Italy
[b] Department of Environmental Chemistry, Pomeranian University in Słupsk, ul. Arciszewskiego 22b, 76-200, Słupsk, Poland
[c] Chair of Analytical Chemistry, Faculty of Chemistry and Pharmacy, University of Sofia "St. Kliment Ohridski", 1 J. Bourchier Blvd., Sofia 1164, Bulgaria

HIGHLIGHTS

- The Self-Organizing Map (SOM) algorithm operation principle is presented.
- The way of extracting information from SOM's output features is described.
- The SOM application for disclosing pollution patterns in the environmental compartments is presented.
- Advice for extracting valuable environmental information from the model results is presented.
- Advice on reporting SOM model details in a paper to attain comparability and reproducibility is presented.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Philip K. Hopke

Keywords:
Air
Sediment
Soil
Water
Pollution
Self-Organizing Map
Clustering and Factorial methods

ABSTRACT

The evaluation of the spatial and temporal distribution of pollutants is a crucial issue to assess the anthropogenic burden on the environment. Numerous chemometric approaches are available for data exploration and they have been applied for environmental health assessment purposes. Among the unsupervised methods, Self-Organizing Map (SOM) is an artificial neural network able to handle non-linear problems that can be used for exploratory data analysis, pattern recognition, and variable relationship assessment. Much more interpretation ability is gained when the SOM-based model is merged with clustering algorithms. This review comprises: (i) a description of the algorithm operation principle with a focus on the key parameters used for the SOM initialization; (ii) a description of the SOM output features and how they can be used for data mining; (iii) a list of available software tools for performing calculations; (iv) an overview of the SOM application for obtaining spatial and temporal pollution patterns in the environmental compartments with focus on model training and result visualization; (v) advice on reporting SOM model details in a paper

Abbreviations: ALK, total alkalinity; ANOVA, analysis of variance; BMU, best matching unit; BOD, biological oxygen demand; BPNN, back propagation neural network; BTEX, benzene, toluene, ethylbenzene, xylene; CART, classification and regression tree; CCA, canonical correspondence analysis; Chl-a, chlorophyll *a*; CMB, chemical mass balance; COD, chemical oxygen demand; CPANN, counter propagation artificial neural network; DME, distribution matching error; DO, dissolved oxygen; DOC, dissolved organic carbon; EC, elemental carbon; Eh, oxidation-reduction potential; FA, factorial analysis; FC, fecal coliform; GPU, graphics processing unit; GUI, graphical user interface; HCA, hierarchical cluster analysis; HDT, Hasse diagram technique; KM, k-means clustering analysis; KNN, k-nearest neighbor neural network; LDA, linear discriminant analysis; LVQ, learning vector quantization; ML-ANN, multi-layer artificial neural network; MLP, multi-layer perceptron neural network; OC, organic carbon; OCPs, organochlorine pesticides; PAHs, polycyclic aromatic hydrocarbons; PARAFAC, parallel factor analysis; PCA, principal component analysis; PCBs, polychlorinated biphenyls; PLS, partial least squares regression; PLS-DA, partial least squares discriminant analysis; PM, particulate matter; PMF, positive matrix factorization; POPs, persistent organic pollutants; QE, quantization error of the map; SAL, salinity; SD, Secchi depth; SIMCA, soft independent modeling by class analogy; SOM, self-organizing map; SPM, suspended particulate matter; TDS, total dissolved solids; TE, topographic error of the map; TH, total hardness; TIC, total inorganic carbon; TIN, total inorganic nitrogen; TN, total nitrogen; TOC, total organic carbon; TOM, total organic matter; TP, total phosphorus; TPH, total petroleum hydrocarbon; TS, total sulfides; TSS, total suspended solids; TVOC, total volatile organic carbons; U-matrix, unified distance matrix; VOC, volatile organic compound; WD, water depth; WQI, water quality index; WT, water temperature.

* Corresponding author.
E-mail addresses: slicen@units.it (S. Licen), aleksander.astel@apsl.edu.pl (A. Astel), stsakovski@chem.uni-sofia.bg (S. Tsakovski).

http://dx.doi.org/10.1016/j.scitotenv.2023.163084
Received 1 December 2022; Received in revised form 23 February 2023; Accepted 22 March 2023
Available online 28 March 2023

to attain comparability and reproducibility among published papers as well as advice for extracting valuable information from the model results is presented.

Contents

1. Introduction

The evaluation of the spatial and temporal distribution of pollutants is an important issue to assess the anthropogenic burden on the environment. Modern analytical techniques with a high level of automation allow to process of many samples for multi-pollutant analysis purposes in a short time interval. Moreover, real-time or quasi-real-time instruments/sensors allow the collection of high-frequency data (Chapman et al., 2020; Dupont et al., 2020). Thus, the most modern direction in environmental analysis is to gather desired information from data sets of a big-data domain, containing information about a variety of physio-chemical characteristics (put as variables in column-wise order) related to each collected sample (put as rows in a data array).

Numerous different chemometric approaches are available for data mining and they have been applied for environmental health assessment purposes. The possible multivariate analysis methods available are usually split into unsupervised and supervised methods. Examples of the former comprises PCA, HCA, KM, CMB, FA, PMF, HDT, and SOM, while the latter LDA, PLS, PLS-DA, CART, MLP, KNN, and SIMCA (Dupont et al., 2020; Govender and Sivakumar, 2020; Hopke, 2015; Mas et al., 2010; Sun et al., 2020; Ye et al., 2020).

Among the unsupervised methods, the SOM (Kohonen, 1987, 2001) is an artificial neural network that can be used for exploratory data analysis and pattern recognition and can handle non-linear problems. Concerning other unsupervised methods, the SOM is also able to deal with big data sets with the possibility of visually exploring the outcomes of the model in versatile 2D maps in which similar samples are mapped close together on a grid (Vesanto, 1999). The SOM is often used in association with other algorithms, such as KM, PCA and HCA, for further elaborating its outcomes. In some applications it is followed by the use of supervised methods.

The SOM, being resistant to missing data as is often the case in environmental studies, was effectively applied by many research groups, however, according to our best knowledge, none of the comprehensive reviews on the SOM concerning environmental pollution assessment were published before except for (Chon, 2011) who focused on the SOM applied to environmental ecology. The first part of the present review comprises a brief discussion of the SOM's main aspects. A description of algorithm initialization parameters, as well as suggestions for their choice, are presented. The SOM output features are described, and suggestions for extracting useful information are included. The SOM peculiarities in comparison with other multivariate analysis techniques are also discussed. The core of the paper is driven by the analysis of how the SOM method is applied to face the environmental pollution issue and how the outcomes of the model are used to achieve the conclusions. Finally, advice on reporting SOM model details in a paper to attain comparability and reproducibility among published papers as well as advice for extracting valuable information from the model results is presented.

2. Document selection

We chose the Scopus database for data mining as (Mongeon and Paul-Hus, 2016) indicated that Scopus has a greater number of indexed journals than the Web of Science (WoS). Moreover, as explained in (Gavel and Iselid, 2008), 85 % of the WoS sources are also included in Scopus but only 51 % of Scopus sources are included in the WoS.

Thus to have an overall picture of the use of the SOM to model pollutant distribution in environmental science we searched the Scopus database (last accessed on 6th December 2021) using the following search parameters: TITLE-ABS-KEY (self AND organizing AND map) AND (pollutant)

Table 1
Number of selected documents split by environmental compartment/s.

Compartment/s	N° of publications
Air	24
Sediment	13
Soil	7
Water	31
Water + Sediment	5
Sediment + Soil	2
Water + Sediment + Soil	1

AND NOT (ecology); DOCTYPE (Article); SUBJAREA (Environmental Science); PUBYEAR (All years to 2021). As a result of the query 172 entries were found.

As the focus of this review is the SOM modeling of pollutant behavior in the different environmental compartments (air, sediment, soil, water) 83 papers were selected from the above-mentioned list, choosing papers containing a sufficiently detailed description of the specific application of the SOM algorithm and possible association with other multivariate techniques as well as the comparison of different data mining approaches. We discarded (i) publications focused on the presence of pollutants in biota; (ii) publications containing modeling of matrix physical parameters without association with pollutants, and (iii) publications focused on epidemiological studies.

The split by different environmental compartment/s of the documents present in the final list is reported in Table 1.

A brief bibliometric analysis has been performed by mining the metadata downloaded from the Scopus website using the bibliometrix package (Aria and Cuccurullo, 2017) in the R software environment (Team, 2016). Two main results are worth to be highlighted: (i) the publication trend shows an annual growth rate of 19 %; (ii) 68 % of the papers are spread in the 40 % of the sources (journals), the remaining papers are present one for each remaining source. The detail about the results is reported in the Supplementary Material.

3. Self-organizing map algorithm

The SOM algorithm is part of the neural network family and it can be used for exploratory data analysis and pattern recognition. The SOM algorithm allows a multivariate analysis of the data using a self-learning approach. It is an unsupervised technique, and hence none of an a priori knowledge about data grouping or classification is necessary. The technique was designed to obtain a nonlinear dimensionality reduction in which similar samples (inputs) are grouped and mapped together in a bi-dimensional representation (Kohonen, 2013).

In the following paragraphs, the main aspects of the learning process and output exploration will be summarized. An extended discussion about the SOM algorithm features can be found in (Himberg et al., 2001; Kohonen, 1987, 2001, 2013) and references therein.

3.1. Self-organizing map algorithm learning process

A representation of the algorithm learning process is presented in Fig. 1a. The experimental data (samples) are represented by a matrix of n-dimensional input vectors defined by their variable values. The goal is to obtain a matrix of n-dimensional output vectors, significantly fewer than the input ones, that is a model of the experimental data and still, represents its variability and the relationships among the variables. The number of the output vectors has to be chosen by the user, the details are discussed in par. 3.2. The output vectors are called nodes (or neurons, prototypes, units), and

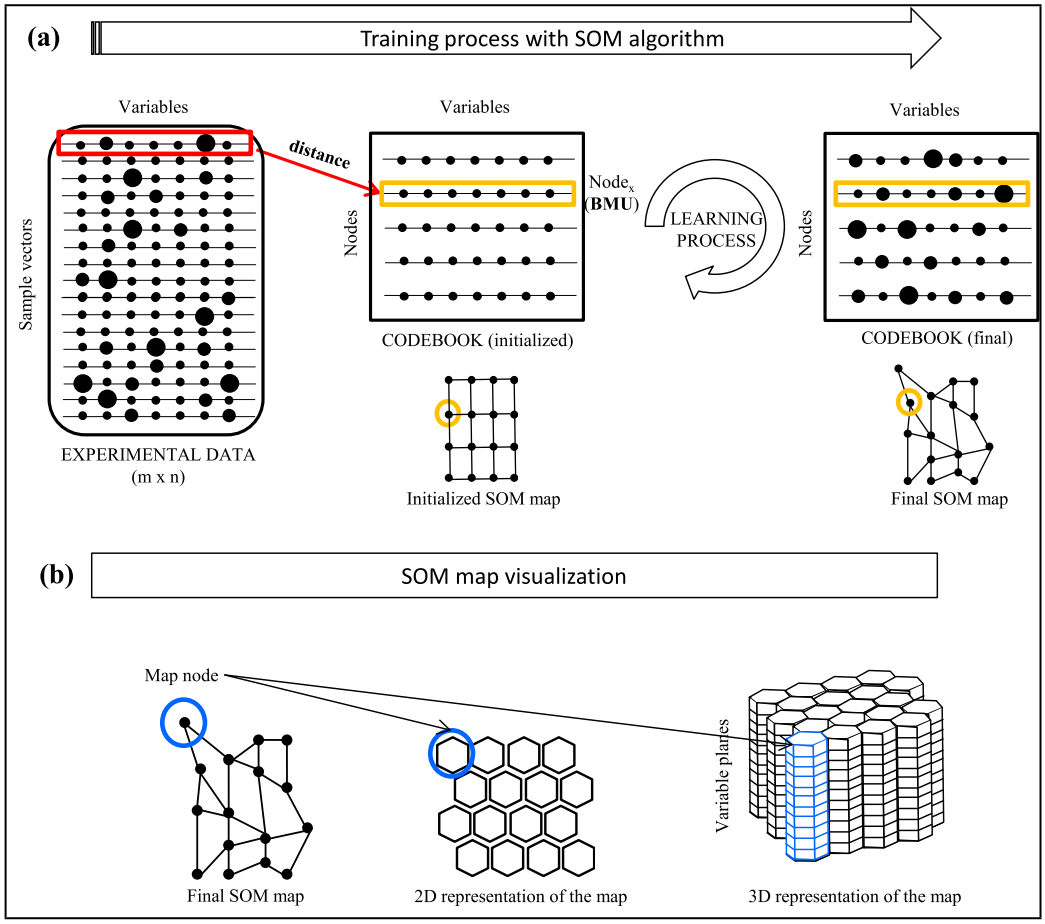


Fig. 1. (a) the Self-Organizing Map (SOM) algorithm learning process; (b) SOM map visualization (BMU = Best Matching Unit).

the matrix containing the values for each modeled variable in the output vectors is usually referred to as a codebook. When the calculation is run, each input vector is presented to the output matrix. Using a "winner-take-all" selection rule, the algorithm matches the input vector with the more similar (i.e. less distant) output vector. The latter is called the best matching unit (BMU). The BMU iteratively updates its values (variable weights) to be more similar to the input vector using sequential or batch algorithms. Moreover, also the weights of the nodes on the map that are in the neighborhood of the BMU are simultaneously updated. In this way, the initialized model "learns" from the experimental data. The learning rate and neighborhood radius are essential constants in the learning process. There are many learning rate functions while the linear, inverse of time, and power series are mostly used in the SOM. The learning rate, as well as neighborhood radius, are factors that decrease monotonically throughout the learning process to ensure convergence (Vesanto and Alhoniemi, 2000). Once each input vector is presented to the output matrix, the first run (epoch) of the process is finished. Usually, this process has to be iterated several times to obtain a model that accurately represents the experimental data variability.

The final output matrix is usually represented in a bi- or tri-dimensional space (Fig. 1b) in which the similar nodes are mapped close together retaining the inherent topological characteristics of the input vectors.

3.2. The SOM initialization parameters

The choice of dimensions of the final SOM map, i.e. the number of nodes and the aspect ratio is a key factor for obtaining a good model of the experimental data. A too-small number of nodes can hide some important differences among the experimental data. On the other hand, a big number of nodes can lead to a poorly significant difference among adjoining nodes. The aspect ratio influences the topology preservation of the experimental data. Other initialization parameters have to be chosen, such as the similarity measure. A list of initialization parameters with a brief description and suggestions for the user choice is presented in Table 2. In general, the best way to obtain a good model for a specific purpose is to try different runs of the algorithm, change one or more parameters, and explore the results (Clark et al., 2020).

3.3. The SOM outputs

The SOM output visualization allows the user to (i) disclose relationships among experimental variables; (ii) highlight the presence of possible similarity groups (clusters) in the experimental data; (iii) establish the appropriateness of the selection of the initialization parameters for the specific problem. In Table 2 a list of the main output parameters of the SOM is proposed, together with a brief description and suggestions for extracting useful information from them.

In Fig. 2 a visual representation of the model outcome information extraction in relation to the experimental data (BMUs, hits, quantization errors, see Table 2) is presented.

Fig. 3 depicts the visualization features of the SOM model. The component planes can be used to visualize possible correlations among the variables (Skwarzec et al., 2009). The Unified distance matrix (U-matrix) can be used to identify the possible presence of different clusters in the data (Ultsch and Lötsch, 2017). Refer to Table 2 for other details.

3.4. Self-organizing map algorithm advantages

The use of the SOM algorithm, with respect to other unsupervised techniques, shows several benefits. First of all, it is a non-linear mapping technique, and hence this results with two advantages: (i) it can be also applied to non-normally distributed data, and (ii) it can reveal non-linear relationships among the variables (Astel et al., 2007). The model output contains the same number of variables as the input data, therefore the model interpretation is simpler with respect to other algorithms that produce new variables related to the experimental ones (i.e. FA, PCA).

Table 2
Description of the main Self-Organizing Map input parameters and output features and their significance.

Input		
Parameter	Description	Significance and suggestions
Shape of the map	The shape of the map can be rectangular, i.e. flat like a sheet, or toroidal. The latter allows joining the map edges.	The rectangular shape can lead to a partial loss of topological preservation of the experimental data because the nodes at the edges have a lower number of surrounding neighbors compared to the central ones (Ultsch and Herrmann, 2007). The toroidal map shows the same number of neighbors around all the nodes, but it is more time-consuming in terms of calculation and more challenging to visualize.
Shape of map nodes	The shape of the map nodes can be rectangular (also called "square") or hexagonal.	The hexagonal shape allows for having more nearest neighbors around a node thus it allows greater topological preservation of the experimental data structure (Kalteh et al., 2008).
Number of nodes	It is the number of modeled vectors that will be used for representing the experimental data. The number of map nodes has to be chosen by the user.	The choice of the number of nodes depends on the level of compression of the experimental data that has to be achieved. The comparison between the number of samples present in the experimental data and the number of nodes is useful to understand how much the data have been compressed in the model. It is usually suggested not to have too many nodes without hits (Kohonen, 2013). This condition can happen if the number of nodes is too large. On the other hand, choosing a very small number of nodes can be considered similar to applying a k-means clustering algorithm because of losing most of the structural features of the experimental data (Ultsch and Herrmann, 2007). Some heuristic rules are available (Vesanto, 1999), suggesting a number of nodes that is five times the square root of the number of samples. If the user needs a more compressed map, it is suggested to reduce four times the above-mentioned number ("small map"). Otherwise, for a more detailed map, it is suggested to multiply four times the above-mentioned number ("big map").
Map dimension ratio	It is the ratio of the grid side lengths representing the node mapping. It has to be chosen by the user.	The most commonly used method is to consider the square root of the ratio between the two largest eigenvalues of the experimental data (Vesanto and Alhoniemi, 2000). The use of the eigenvalues allows to shape the map mimicking the experimental data structure.
Map initialization matrix	It is the codebook from which the learning process starts See Fig. 1(a). It can be built randomly or based on data analysis of the experimental data. In the former method the weight of nodes is initialized either by randomly assigning	In general, a random initialization requires a greater number of epochs and can lead to different outcomes for every run. A "seed" must be set to attain reproducible results. Data analysis-based initialization usually requires a smaller

Table 2 (*continued*)

Input		
Parameter	Description	Significance and suggestions
	small values to the weights or using the vector weights of randomly sampled samples from the experimental data. In the data analysis method the initialization matrix is usually built using the eigenvectors related to the two largest eigenvalues of the experimental data. Other methods can be used e.g. genetic algorithms (Ballabio and Vasighi, 2012).	number of epochs because it starts from nodes that already roughly represent the experimental data structure. Moreover with this initialization method reproducible results are obtained (if the other SOM parameters are not changed).
Similarity measure	It is a measure used to find the best matching unit (BMU) during the learning process and adjust the node values. The most commonly used is Euclidean distance (others possibilities are e.g. sum of squares or Manhattan distance).	Different similarity measures can lead to different clustering of the experimental data on the map. Special attention should be paid on high dimensional data sets with 10 or more variables, where BMU finding could become unstable because of negligible difference in distances between BMU and experimental samples.
Neighborhood function	It is the kernel function that regulates the smoothing process of the map during the training. The most commonly used is the Gaussian function. Other functions, such as bubble, cut Gaussian and, Epanechnikov can be used.	Different functions shape the final map with different smoothing, the more the smoothing the better the map, but the more the calculation time. See (Clark et al., 2020) for details.
Training algorithm	It is the underlying engine for building neural network models with the goal of training features or patterns from the input data. In the SOM usually, one of two algorithms is used: sequential or batch training algorithm, STA or BTA, respectively. An STA constructs the nodes in the SOM to represent the entire data set and their weights are optimized at each iteration step. In BTA, instead of using a single data vector at a time, the whole data set is presented to the map before any adjustment is made.	(Kohonen, 2013) suggests that BTA is recommendable for practical applications, because it does not involve any learning-rate parameter, and its convergence is an order of magnitude faster.
Learning rate function	It is a hyper-parameter used to govern the pace at which the neural network updates or learns the value of a parameter estimate. Can be linear, the inverse of time and power series. It decreases over time to enable convergence.	The learning rate function should be analyzed experimentally, however, a linear value is commonly set as initial.
Neighborhood radius	It is a distance around the node that defines a set of BMU neighborhood nodes that are adjusted during the training of the neural network net. It decreases over time to enable finer adjustment of the net.	There is no suggested quality value to refer to, however, usually a radius higher than 1 is set for the first phase of the training net, that is defined as "rough". The value is decreased to one for net "fine" tuning. A value equal to one means that only the first BMU neighbors are updated (Himberg et al., 2001).
Number of epochs	It is the number of times the experimental data are presented to the algorithm to outline the final map. It has to be chosen by the user.	The number of epochs has to be adequate to obtain a map that is "stable" i.e. substantial changes are not perceived if the number of epochs is further increased. Too many epochs can lead to overfitting the training data. Calculating the DME for maps trained with different number of epochs can allow the establishment of the minimum number of epochs necessary for adequately modeling the experimental data structure (Ponmalai and Kamath, 2019). Some heuristic rules are also available (Vesanto, 1999). It is suggested to set the number of epochs as the maximum value between two and (N/m*10 + N/m*40)/8 where N = the number of nodes and m = the number of samples in the experimental data.

Table 2 (*continued*)

Output		
Feature	Description	Significance and suggestions
Codebook	It is the matrix containing the values for each modeled variable in the output vectors (i.e. the weight of nodes), thus it has the same number of rows as the nodes and the same number of columns as the variables of the experimental data (Fig. 1a)	It represents the recurrent variable profiles present in the experimental data. It also allows the representation of the component planes.
Component plane	A component plane describes the distribution of the values of a modeled variable on the map by different coloring (*heatmap*). It represents the weight of the nodes for a variable only (see Fig. 3a).	The visual exploration and comparison of the component planes allows for the disclosure of possible relationships among the variables. Variables that show a similar distribution of the values on the map are usually correlated. Variables that present an opposite distribution of the values are usually anti-related.
Unified Distance Matrix (U-matrix)	The U-matrix shows the distances between the nodes by different coloring (see Fig. 3b).	It allows the identification of possible cluster separation of the data. The U-matrix can be imagined as a mountain-valley representation (Ultsch and Lötsch, 2017) in which the mountains represent greater distances thus they indicate the separation between possible clusters of data.
Best Matching Unit (BMU) list	Each experimental vector is associated in terms of similarity to a node (BMU). This list contains the number of a node which each sample is associated with.	Experimental data samples that are associated to the same BMU can be considered generally very similar to each other and represented by the vector weights of the BMU. Nevertheless, as the algorithm is forced to associate each sample to a node, also a possible outlier sample is associated to a BMU. Thus the level of similarity between the sample and the BMU has to be verified by checking the related quantization error. On the other hand, a node may be a BMU of none of the samples.
Hits	The hits represent the number of samples associated with each node (see Fig. 2).	The hits distribution on the map, together with the U-matrix, allows identifying possible sample grouping.
Quantization error list	The quantization error represents the difference in similarity between a sample and its BMU (see Fig. 2). The more the error the less the similarity. This list contains the error associated to each sample in relation to its BMU.	Basic statistics can be applied to the list for evaluating the error value distribution. Relatively high error values (e.g. higher than the 95th percentile or 98th percentile (Licen et al., 2019)) can show the presence of possible outlier samples.
Quantization error of the map (QE)	The QE corresponds to the average of the quantization error list. It is a measure of the quality	The QE usually decreases as the number of nodes increases. There is no suggested quality value to

Table 2 (*continued*)

Output		
Feature	Description	Significance and suggestions
	of the map.	refer to, but it can be used for comparing different maps obtained from the same dataset. See (Clark et al., 2020) for details.
Topographic error (TE)	The TE represents the degree of topology preservation with respect to the experimental data. It is a measure of the quality of the map.	The TE usually increases as the number of nodes increases. There is no suggested quality value to refer to, but it can be used for comparing different maps obtained from the same dataset. See (Clark et al., 2020) for details.
Distribution matching error (DME)	It is an estimation of the convergence of the probability distribution of the SOM nodes on the training data. It is a measure of the quality of the map (Yin and Allinson, 1995).	The Kolmogorov-Smirnov (KS) test is performed for each variable and the DME is the proportion of variables that fail the test. The KS test has two main advantages: it is a non-parametric test and it can be applied for comparing series with a different number of elements such as the experimental data and the codebook. See (Ponmalai and Kamath, 2019) for more details.

The visualization of the output produces a high-dimensional data projection in a bi-dimensional space while retaining the topological structure of the input data, thus mapping together similar samples (Wienke et al., 1995). This feature allows the extraction of meaningful information also from large and complex data (Licen et al., 2020a). For comparison, when analyzing large datasets using techniques such as PCA, the graphical representation of the results has to be optimized using density scatterplots for allowing visual exploration.

Another advantage of the SOM algorithm is that it is relatively resistant to missing data (Astel et al., 2007). During the learning process, the similarity measure can be also calculated for sample vectors containing missing values, because the corresponding variables are selectively excluded from the calculations (Clark et al., 2020). Moreover, the SOM algorithm can be successfully used for a posteriori estimation of missing values as shown in Folguera et al. (2015) who reported an application concerning environmental datasets.

The SOM algorithm is also able to deal with data noise (Ponmalai and Kamath, 2019; Vesanto, 1999). This feature is particularly important when modeling data recorded by instruments. The positive side effect of this aspect is that the model in not affected too much by outliers (Muñoz and Muruzábal, 1998; Vesanto, 1999). A variable noise reduction needs a two-step optimization that is represented in the SOM algorithm by the joined use of a vector quantization algorithm and a neighborhood function (Yin, 2008). The weight of the nodes that are updated during the learning process can be considered local averages of the data, thus similar sample vectors except for the noise are associated to the same node. For the above-mentioned property a two-step clustering that applies SOM followed by KM is more effective in terms of computational time and clustering quality than applying KM only (Misra et al., 2020).

The SOM shows an easy way to identify possible outliers. The outliers can be either isolated in a single node showing a high distance from the neighbor nodes or they can be scattered onto the map (Muñoz and Muruzábal, 1998). Outlier detection can be achieved by inspecting the quantization error values, for example establishing a threshold above which the related experimental data have to be checked (Licen et al., 2019).

3.5. Available software and tools

Several software and tools are available for performing the SOM calculation. The most used in the reviewed papers were: the SOM toolbox (Vesanto and Alhoniemi, 2000) for MATLAB environment and the kohonen package (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018) that works in the R environment (Team, 2016). Some studies used Kohonen and CPANN toolbox (Ballabio and Vasighi, 2012) for MATLAB environment. Few studies used Statistica Neural Networks, SPSS, Australian SiroSOM® software, Statistica 8.0.

Although not present in the reviewed papers, among the tools prepared in R environment, it is worth adding to the list some recent packages: (i) SOMbrero package (Olteanu and Villa-Vialaneix, 2015) which provides methods for handling numerical data, contingency tables, and dissimilarity matrices and a GUI for training and visualization; (ii) somoclu package (Wittek et al., 2017), which is a general toolbox for training SOMs with support for cluster and GPU computations, and interfaces to Python, MATLAB and R; (iii) SOMEnv package (Licen et al., 2021) which is based on the kohonen package (Wehrens and Kruisselbrink, 2018) with embedded Vesanto heuristic rules (Vesanto and Alhoniemi, 2000) and a GUI for training and visualization of SOM, with some features dedicated to high frequency data elaboration; (iii) aweSOM package (Boelaert et al., 2021) which is based on kohonen package (Wehrens and Kruisselbrink, 2018) and has a GUI for training and visualization of the SOM on numeric, categorical or mixed data. For a comparison of the tools see Table 3.

4. The SOM application to pollutant profiling in environmental compartments

The general scheme of the SOM method application covered by the reviewed papers is proposed in Fig. 4. As a rule, the data were pre-processed using a normalization technique and then the SOM-based model was run. In

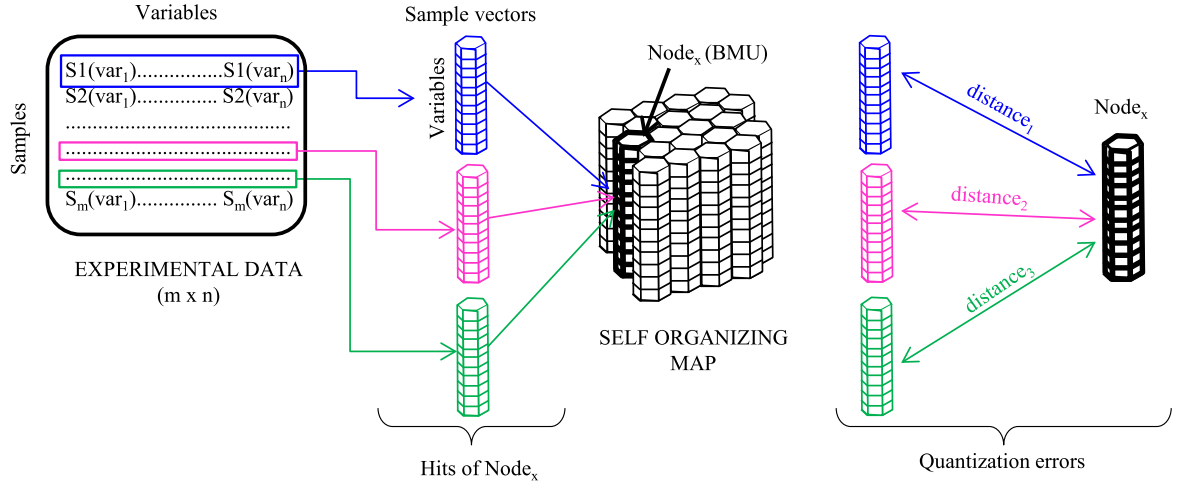


Fig. 2. Hits and quantization error extraction representation (BMU = Best Matching Unit).

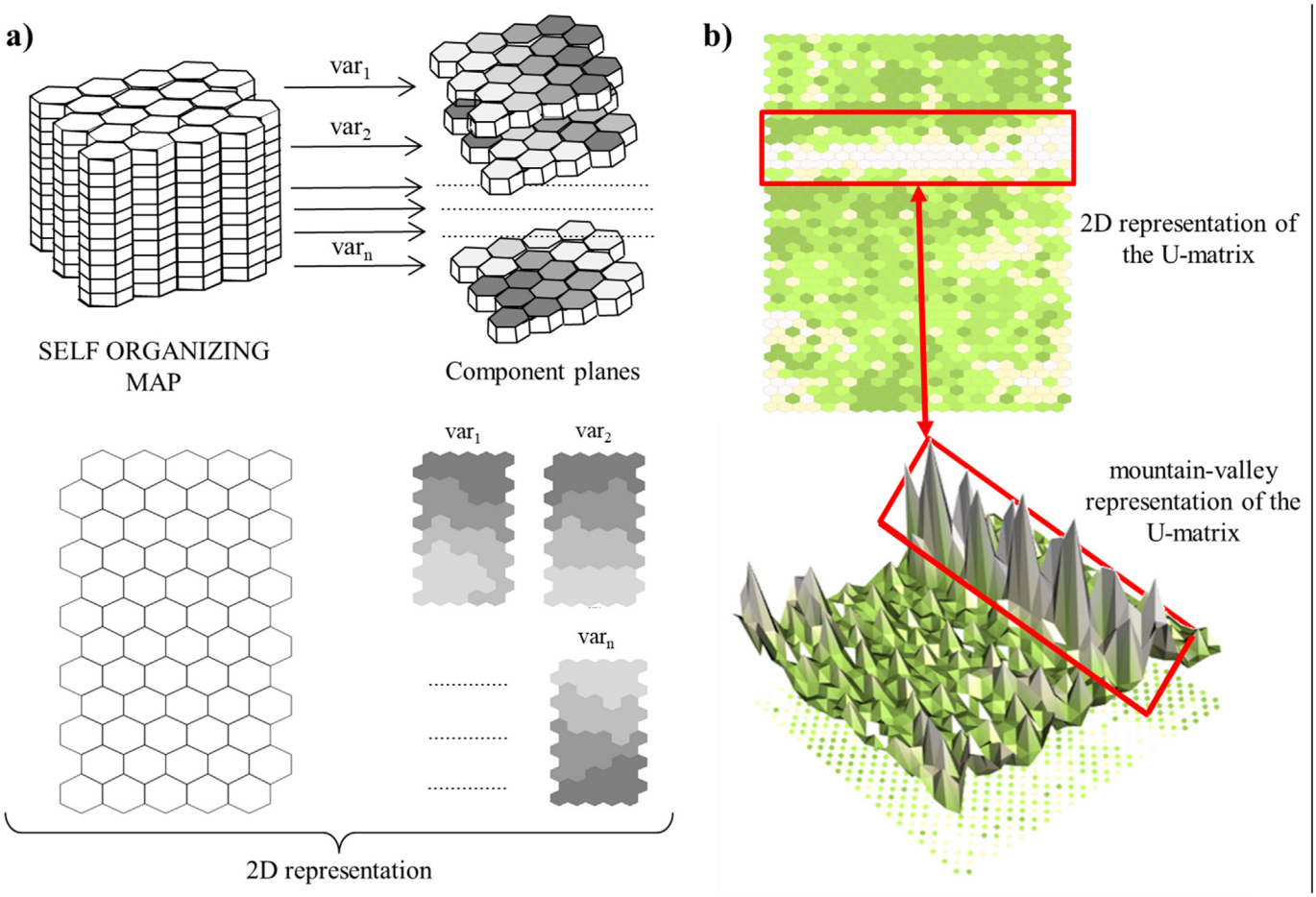


Fig. 3. (a) Self-Organizing Map, component planes, and (b) Unified distance matrix (U-matrix) representation.

numerous studies variables different from pollutants were also considered, such as meteorological factors, land use, matrix physical properties, distances from possible pollution sources, etc. These variables were used either for model training or improving model interpretation as covariates. In several papers, the SOM was used for data dimensionality reduction and the output matrix (codebook) underwent a 2nd level abstraction using another multivariate technique for revealing possible clusters in the samples. Some studies reported on the independent use of different multivariate techniques applied to the data and related results in comparison with SOM model outcomes.

The number of processed samples spanned from a few dozen to tens of thousands, the latter usually in the case of data recorded by high-frequency instruments. The dimensionality of environmental monitoring data sets could differ from 1D when only pollutants and sampling stations (sites) were included to 2D or 3D in data sets where time (seasons), depth, or certain fractions of the investigated environmental compartment were considered. The SOM has the ability to represent models for each dimension (time, depth) in 2D and 3D data sets. This "unfolding" provides an opportunity for the creation of SOM-based models such as "time vs pollutants" for each site or "depth vs pollutants" for each site.

In Table 4 a list of all the discussed papers is presented. Information about the environmental compartment, considered variables, combined multivariate techniques, and the scope of the study are reported. The papers are listed by environmental matrix and by the first author. As regards the air compartment the papers were grouped according to the use of meteorological data in the model building: (Set 1) studies that considered pollutant data only; (Set 2a) studies using both pollutant and meteorological data as training data; (Set 2b) studies using meteorological data as training data, and pollutant concentrations as associated data; (Set 2c) studies using pollutant concentrations as training data, and meteorological data as associated data.

In the following paragraphs, the paper content is discussed focusing on: (i) the different ways of extracting information from the model, with special attention to visualization methods; (ii) the coupling of SOM with a 2nd level abstraction technique; (ii) the comparison to other multivariate analyses independently applied to the experimental data. Finally, approaches related to different applications, such as single/multi-component compartments and spatial/temporal variations, are discussed.

4.1. Information extraction from the SOM model

After training, the SOM outcomes were explored for identifying similar samples. The visualization of the hits as sample names depicted on the SOM map was often used. This method allows to assess of sample grouping and finding qualitative correlations with sample categories (site, collection time, depth) (Carrillo et al., 2021; Geng et al., 2021; Jiang et al., 2021; Ki et al., 2017; Liu et al., 2019; Noh et al., 2016; Olkowska et al., 2014; Tudesque et al., 2008; Zhu et al., 2020). This method cannot be applied if the number of samples is too large. In the latter case, the nodes were depicted with a size proportional to the number of hits (Åkesson et al., 2015; Hossain Bhuiyan et al., 2021; Licen et al., 2019; Orak et al., 2020; Pandey et al., 2015). The comparison between the hits and the component planes easily reveals the most polluted samples. In some papers, the weight of the nodes was drawn on the SOM map by use of pie charts (Chang et al., 2020; Ladwig et al., 2017; Liao et al., 2020; Tsuchihara et al., 2020). The afore-mentioned visualization is useful because it collects a variety of information in one figure, but it is not applicable if the number of nodes is more than one hundred because the figure could be scarcely readable.

Table 3
Description of the main available software tools for calculating and visualizing Self-Organizing Map models.

Tool	SOM toolbox (Vesanto and Alhoniemi, 2000)	Kohonen and CP-ANN toolbox (Ballabio and Vasighi, 2012)	kohonen package (Wehrens and Buydens, 2007; Wehrens and Kruisselbrink, 2018)	SOMbrero package (Olteanu and Villa-Vialaneix, 2015)	somoclu library (Wittek et al., 2017)	SOMEnv (Licen et al., 2021)	aweSOM (Boelaert et al., 2021)
Software environment	MATLAB	MATLAB	R software	R software	Multiplatform (Python, R software, MATLAB)	R software	R software
Shape of the map	Rectangular or toroidal	Rectangular or toroidal	Rectangular or toroidal	Rectangular	Rectangular or toroidal	Rectangular	Rectangular
Shape of the nodes	Rectangular or hexagonal	Rectangular	Rectangular or hexagonal	Rectangular or hexagonal	Rectangular or hexagonal	Hexagonal	Hexagonal
Number of nodes	Heuristic rules available	User choice	User choice	User choice	User choice	Heuristic rules available	User choice
Map dimension ratio	Heuristic rules available	Squared maps only (ratio = 1)	User choice	User choice	User choice	Heuristic rules available	User choice
Codebook initialization	Random, random samples, eigenvalues based	Random, eigenvalues based, genetic algorithm	Random or user choice	Random, random samples, eigenvalues based	Random, eigenvalues based	Eigenvalues based	Random, eigenvalues based
Similarity measure	Euclidean	Euclidean	Sum of squares, Euclidean, Manhattan, Tanimoto	Euclidean, maximum, Manhattan, Canberra, Minkowski, Letremy	Euclidean	Euclidean	Sum of squares
Neighborhood function	Bubble, gaussian, cut-gaussian, Epanechnikov	Gaussian	Bubble, gaussian	Gaussian, Letremy	Bubble, gaussian	Bubble, gaussian	Bubble
Number of epochs	Heuristic rules available	User choice	User choice	User choice	User choice	Heuristic rules available	User choice
Training algorithm	Sequential or batch	Sequential or batch	Sequential or batch	Sequential	Batch	Batch	Sequential
Neighborhood radius	Default values or user choice	Default values or user choice	Default values or user choice	Default values or user choice	Default values or user choice	Default values or user choice	Default values or user choice
Graphical User Interface (GUI)	No	Yes	No	Yes	No	Yes	Yes

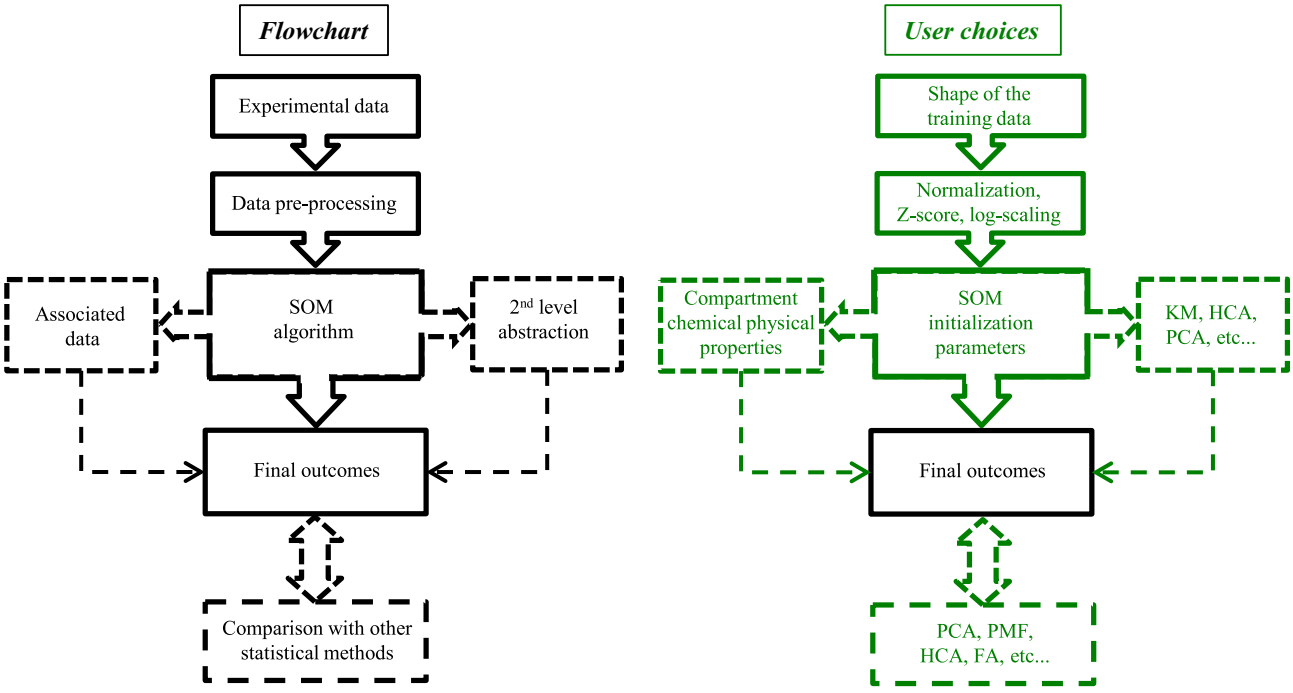


Fig. 4. General scheme of the Self-Organizing Map (SOM) method flowchart (left) and corresponding possible user choices (right). The elements surrounded by dashed lines are optional (KM = k-means clustering analysis; HCA = hierarchical cluster analysis; PCA = principal component analysis; PMF = positive matrix factorization; FA = factorial analysis).

Different techniques were used for finding possible correlations among the variables: (i) visual exploration of the component planes (Alvarez-Guerra et al., 2008; Carafa et al., 2011; Lee et al., 2019; Wang et al., 2020); (ii) a correlation matrix evaluation and comparison with SOM output (Dai et al., 2018; Guo et al., 2020; Xiao et al., 2021); (iii) the inspection of the loading plot obtained by an independent PCA approach and comparison with SOM output (Guo et al., 2020; Kebonye et al., 2021; Lee et al., 2021; Noh et al., 2016); (iv) the application of the SOM algorithm on the transposed codebook (Cheng et al., 2017; Rogowska et al., 2014; Tobiszewski et al., 2010; Tsakovski et al., 2010b); (v) the use of Spearman correlation coefficient (Åkesson et al., 2015). The (i) approach is straightforward and thus, especially useful when comparing SOM models obtained by different runs starting from the same input data. The (ii) and (iii) methods are interesting since they compare the results of different independent techniques. The (iv) and (v) approaches are convenient for disclosing non-linear relationships among the variables.

4.2. Second-level abstraction

Several studies presented a combined approach using SOM and KM algorithms, some studies used HCA, and a few studies used PCA.

Regarding KM or HCA, as a rule, the cluster split was shown on the SOM map by bold lines or different colors for visual comparison to component planes. None of the studies tried different clustering methods applied to the codebook. The preferred way was to use one clustering method on the codebook and compare the results to a clustering method applied to the experimental data. An analysis of the possibly different information mined using diverse clustering algorithms applied to the codebook would be interesting, especially considering the easy way of representation of the cluster split on the SOM map. Two studies used fuzzy-KM (Lee et al., 2019; Orak et al., 2020). The use of the latter is interesting because it allows assigning the nodes to more than one cluster, thus highlighting spatial or temporal transition "states" that are usually present in complex matrixes.

The variable profiles of the clusters were shown by different plots such as bar plots (Astel and Małek, 2008; Chen et al., 2016; Tsakovski et al., 2009), boxplots (Astel et al., 2016; Guo et al., 2020; Khedairia and Khadir, 2012; Licen et al., 2020b; Yang et al., 2012), radar plots (Amiri and Nakagawa, 2021; Kim et al., 2019; Nakagawa et al., 2020; Tsuchihara et al., 2020). In some of the papers, the statistical difference among the clusters was assessed by significance tests, such as Kruskal–Wallis test (Astel et al., 2016; Chea et al., 2016; Olkowska et al., 2014; Torres-Martínez et al., 2021; Tsakovski et al., 2010b), Mann-Whitney test (Lee et al., 2019; Li et al., 2020a; Skwarzec et al., 2012), ANOVA (Carafa et al., 2011; Dai et al., 2018; Li et al., 2020b). The use of a significance test can confirm the reliability of the grouping and highlight the cluster differences.

PCA was applied to the codebook for different scopes. In (Carafa et al., 2011) a biplot obtained by PCA was used to identify the variables that emphasize the differences among clusters previously obtained by applying HCA to the codebook. (Yu et al., 2021) applied PCA to the transposed codebook for identifying the variables that highlighted the differences among the SOM map areas thus inferring a clustering rule based on the specific effects of the variables on the water quality.

4.3. Comparison of the SOM outcomes to different multivariate analyses

Multivariate techniques diverse from SOM were applied to the experimental data either for adding complementary information or for comparison with SOM outcomes. When utilizing different multivariate techniques, it is preferable to display the applied method in a flowchart for elucidating the different scopes. Some examples are reported in (Alvarez-Guerra et al., 2011; Chang et al., 2020; Cheng et al., 2017; Kebonye et al., 2021; Wesolowski et al., 2006; Yang et al., 2012; Yu et al., 2021; Zhu et al., 2020).

Some studies used PCA for assessing variable correlation (see par. 4.1). Several studies assessed the source apportionment to be compared with SOM clustering. They used PMF (He et al., 2021; Hossain Bhuiyan et al., 2021; Li et al., 2021; Li et al., 2020b; Tao et al., 2017; Wang et al., 2020), FA (Chen et al., 2016; Jampani et al., 2018; Kumar et al., 2021; Wang et al., 2015b), and PCA (Dai et al., 2018; Mao et al., 2021; Noh et al., 2016). (Yang et al., 2012) used HCA and LDA before the SOM to analyze the whole dataset characteristics aiming to set the optional clustering pattern in spatial and temporal ways separately. Using the SOM and KM, the clustering obtained by HCA was validated and visualized on the SOM map.

In the following paragraphs, the studies reporting a performance comparison between the SOM and other multivariate techniques are presented.

4.3.1. PCA

(Veses et al., 2014) aimed to determine the spatial distribution of heavy metals and total PAHs in a river basin. The sediment sample classification provided by PCA was not as useful as the one provided by the SOM, revealing itself as a powerful tool to be integrated with the first steps of sediment quality assessments. The authors prepared a useful summary table for presenting the benefits and drawbacks of the use of the two statistical analyses for the specific scope of the study. (Gontijo et al., 2021) analyzed the sediment content of humic and fulvic acids in an artificial reservoir. The authors found that both techniques could be efficiently used to cluster data and interpret results. However, they highlighted the advantages of SOM over PCA, such as the visualization features and the capability to retain data topology and deal with noisy and missing data. (Xiao et al., 2021) calculated a SOM model considering samples collected both in the wet and dry seasons. A PCA model was obtained for each season. Both approaches allowed identifying anthropogenic and natural sources. The SOM approach proved to be effective because it presented meaningful results even if the season were processed together. In (Amiri and Nakagawa, 2021) two separate SOM-based models were built using data collected during the dry and wet seasons, respectively in a coastal aquifer. The SOM output enabled more accurate clustering without overlaps with respect to PCA.

4.3.2. HCA

(Yotova et al., 2018) focused on toxic elements present in soil and their phytoavailability in an industrial area with copper mining factories and smelters. Both HCA and SOM provided similar grouping of the sampling sites. The SOM was able to discriminate some specific sites with intermediate pollution characteristics. (Jampani et al., 2018) analyzed groundwater physicochemical properties in long-term wastewater irrigated systems. HCA was used for variable clustering and the results were compared to the SOM component planes. The SOM gave insight into some variable behavior that was misinterpreted by HCA.

4.3.3. HCA and KM

(Lu et al., 2006) used SOM followed by KM to classify the PM_{10} distribution in Taiwan and identify "air-quality basins" with different impacts. Geometric means and geometric standard deviations in each of the five air-quality basins were found significantly different from each other for the SOM method by the Waller–Duncan k-ratio *t*-test. The grouping obtained by HCA and KM was less reliable. (Pearce et al., 2014) elaborated eight years of daily multi-pollutant concentrations collected at different monitoring stations in Atlanta identifying daily recurrent patterns. The outcomes were compared with KM and HCA finding consistent results but with the SOM adding more insight into between-class relationships.

4.3.4. HCA and PCA

(Alvarez-Guerra et al., 2008) independently mined a sediment dataset both with HCA and PCA. They concluded that the powerful visualization tools of the SOM facilitated the task of establishing an order of priority between the groups of sites depending on their need for further investigations or remediation actions in subsequent management steps. (Vignati et al., 2013) study dealt with heavy metal contamination in sediments. SOM results confirmed the information obtained by HCA and PCA. SOM allowed a better overview of how the different sediment characteristics were reflected across the sites. (Yotova et al., 2021) proposed an approach

Table 4
List of reviewed papers and main information therein.

Document	Environmental compartment/s	Sampling dimensions	2nd level abstraction	Independent multivariate technique/s	Pollutant/s and other variables	Main objective
(Alvarez-Guerra et al., 2011)	Air (Set1)	2D (stations, years)	KM	–	NO_2, SO_2, PM_{10}, O_3, CO, NOx	Classification of Spanish air quality monitoring stations
(Gulson et al., 2007)	Air (Set1)	2D (stations, years)	–	–	Elements in $PM_{2.5}$, Pb isotopes	Fingerprint identification of different sources of $PM_{2.5}$ in rural and urban sites in Australia
(Lu et al., 2006)	Air (Set1)	2D (stations, seasons)	KM	HCA	PM_{10}	Classification of the seasonal and spatial distribution of PM_{10} in Taiwan
(Neme and Hernández, 2011)	Air (Set1)	2D (stations, years)	–	–	NO_2, NO, O_3, SO_2, CO, Pb, $PM_{2.5}$, PM_{10}	Assessment of the correlation and evolution of pollutant patterns during eight years in Mexico City
(Pearce et al., 2016)	Air (Set1)	2D (stations, years)	–	–	CO, NO_2, NOx, O_3, SO_2, $PM_{2.5}$	Characterization of the spatial distribution of pollutants in Atlanta (Georgia)
(Tsakovski et al., 2012b)	Air (Set1)	1D (stations)	KM	PCA	inorganic trace elements, ionic species, OC, EC, PAHs, azarenes (in PM_{10})	Identification of different pollution sources in Krakow (Poland)
(Wesolowski et al., 2006)	Air (Set1)	2D (stations, years)	–	PCA, MLP	NO_2, SO_2, PM, PAHs	Analysis of seasonal pollution patterns in Gdansk (Poland) and comparison
(Astel et al., 2013)	Air (Set2a)	2D (stations, seasons)	KM	–	BTEX, meteorological data	Assessment of multi-year spatial and temporal variations of pollutant signatures in Trieste(Italy)
(de Oliveira et al., 2019)	Air (Set2a)	2D (stations, years)	KM	–	NO_2, CO, CO_2, O_3, meteorological data	Identification of spatial and daily variation of air pollution signatures in Sao Paulo (Brazil)
(Nathan and Lary, 2019)	Air (Set2a)	1D (stations)	–	–	Hydrocarbons, meteorological data	Identification and regional scale distribution of pollutant patterns from different sources in Texas
(Zhong et al., 2017)	Air (Set2a)	2D (stations, years)	KM	–	Fluoride in PM, meteorological data	Assessment of the spatial distribution of pollution level and inhalation exposure of fluoride near a mining and smelting facility in China
(Crawford et al., 2016)	Air (Set2b)	2D (stations, years)	–	–	$PM_{2.5}$ (H, Na, Al, Si, P, S, Cl, K, Ca, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Br, Pb therein), meteorological data	Identification of multi-year synoptic weather types related to different pollutant concentrations in the Sidney Region (Australia)
(Khedairia and Khadir, 2012)	Air (Set2b)	2D (stations, years)	KM	MLP	NO, CO, O_3, PM_{10}, NOx, NO_2, SO_2, meteorological data	Identification of typical meteorological conditions in an Algeria region and their relation to pollutant concentrations
(Jiang et al., 2016)	Air (Set2b)	2D (stations, years)	–	–	O_3, meteorological data	Identification of synoptic weather types related to ozone concentration exceedances in Sidney (Australia)
(Liao et al., 2020)	Air (Set2b)	2D (stations, seasons)	–	–	$PM_{2.5}$, meteorological data	Assessment of multi-year spatial and temporal variation of $PM_{2.5}$in dry-season related to synoptic circulation types in the Pearl River Delta Region (China)
(Bougoudis et al., 2014)	Air (Set2c)	2D (stations, years)	–	ML-ANN	O_3, NO, NO_2, CO, SO_2, meteorological data	Assessment of multi-year air quality clusters, with extreme pollution events identification in Attica region (Greece)
(Chang et al., 2020)	Air (Set2c)	2D (stations, years)	–	BPNN	$PM_{2.5}$, meteorological data	Assessment of spatio-temporal variations of $PM_{2.5}$ concentration profiles in northern Taiwan
(Romanić et al., 2018)	Air (Set2c)	2D (stations, seasons)	–	–	PCBs, OCPs, meteorological data	Assessment of seasonality variation of pollutant patterns in Zagreb (Croatia)
(Licen et al., 2018)	Air (Set2c)	1D (stations)	KM	–	Odor, benzene, H_2S, CO, meteorological data	Identification of pollution events related to industrial sources at a residential site in Trieste (Italy)
(Licen et al., 2019)	Air (Set2c)	2D (stations, years)	KM	–	PM size fractions, PM_{10}, meteorological data	Identification of the PM size fraction pattern of different industrial sources at two residential sites in Trieste (Italy)
(Licen et al., 2020a)	Air (Set2c)	2D (stations, years)	KM	–	PM size fractions, PM_{10}, meteorological data	Multi-annual variation of PM size fraction pattern of industrial sources at a residential site in Trieste (Italy)
(Licen et al., 2020b)	Air (Set2c)	1D (stations)	KM	–	odor, TVOC, meteorological data	Assessment of the pollution pattern profiles of different industrial sources in Viggiano (Italy)

(Pearce et al., 2014)	Air (Set2c)	2D (stations, years)	–	–	CO, NO_2, NOx, O_3, SO_2, $PM_{2.5}$, meteorological data	Multi-year analysis of pollution day types in Atlanta (Georgia) for developing a multipollutant metric of air quality
(Wu et al., 2017)	Air (Set2c)	2D (stations, years)	–	–	$PM_{2.5}$, meteorological data	Assessment of annual cycle spatial distribution of $PM_{2.5}$ concentration signatures in East Asia
(Alvarez-Guerra et al., 2008)	Sediment (estuaries)	1D (stations)	KM	HCA, PCA	TOC, As, Cr, Cu, Fe, Mn, Ni, Pb, Zn, PAHs, ecotoxicity	Identification of spatial pollutant patterns and relation with sediment physical properties (Spain)
(Arias et al., 2008)	Sediment (dock)	3D (stations, depth, fractions)	–	–	Cu, Co, Mn, Ni, Cr, Pb, Zn, Fe, Mg, Ca, K, Na	Assessment of metal contamination and relation with their mobility in the matrix (Spain)
(Chen et al., 2016)	Sediment (coastal)	1D (stations)	–	HCA, FA	TS, TN, TP, E_h, TOC, Cu, Pb, Zn, Cd, Cr, Hg, As	Assessment of the spatial distribution of heavy metal contamination in surface sediment (China)
(Dai et al., 2018)	Sediment (lake)	2D (stations, depths)	KM	PCA	Cr, Cu, Cd, Pb, Zn	Assessment of the geoaccumulation patterns of pollutants (China)
(Gontijo et al., 2021)	Sediment (reservoir)	1D (stations)	–	PCA	Humic and fulvic acids, isotopes	Identification of the origin and quality of sedimentary organic matter (Brazil)
(Ladwig et al., 2017)	Sediment (lakes)	2D (stations, depths)	–	PCA, KM	K, Ca, Ti, Rb, Zr, Sr, Mn, Fe, Cr, Cu, Zn, Pb	Identification of the evolution of pollutant patterns related to lake management history (Germany)
(Li et al., 2020b)	Sediment (river, suspended particulate matter)	2D (stations, years)	KM	PMF	PAHs	Assessment of the influence of anthropogenic activities on pollutant spatial and temporal patterns (Central Europe)
(Rogowska et al., 2014)	Sediment (coastal)	2D (stations, depth)	KM	–	Cd, Co, Cr, Cu, Fe, Hg, Mg, Mo, Ni, Pb, V, Zn, PAHs, PCBs	Evaluation of the pollution pattern and ecotoxicity due to a Second World War shipwreck (Poland)
(Tsakovski et al., 2009)	Sediment (lake)	1D (stations)	KM	–	Cr, Cu, Ni, V, Fe, Al, Li, Cd, Pb, As, Hg, PAHs, PCBs, pesticides	Assessment of relationship between pollutant spatial distribution and ecotoxicity (Poland)
(Tsakovski et al., 2012a)	Sediment (lagoon)	2D (stations, compartments)	KM	–	Zn, Cu, Mn, Pb, Cd	Assessment of the relationship between heavy metal spatial distribution and ecotoxicity (Spain)
(Veses et al., 2014)	Sediment (river)	1D (stations)	KM	PCA	Cd, Cu, Ni, Pb, Zn, Hg, As, Cr, PAHs	Assessment of freshwater sediment quality in a large area (Spain)
(Vignati et al., 2013)	Sediment (river)	2D (stations, years)	KM	PCA, HCA	Cd, Co, Cr, Cu, Ni, Pb, V, Zn, TiO_2, Fe_2O_3, MnO, $CaCO_3$, TOC	Identification of the spatial and temporal patterns of pollutants (Romania/Ukraine)
(Wang et al., 2015a)	Sediment (river)	1D (stations)	KM	FA, PMF	PAHs	Assessment of pollutant pattern spatial distribution and relation with health risk (Taiwan)
(Wang et al., 2020)	Sediment (lake)	2D (stations, years)	KM	PMF	Cr, Mn, Ni, Cu, Zn, As, Cd, Hg, Pb	Assessment of anthropogenic sources of heavy metals (China)
(Carrillo et al., 2021)	Sediment/soil (biological reserve wetland)	2D (stations, seasons)	–	PCA, HCA	Cu, Cd, Co, Mo, Ni, V, As, Ba, Pb, Zn, Cr	Assessment of the spatial distribution of contamination and ecological risk index levels (Ecuador)
(Cheng et al., 2017)	Sediment/soil (reservoir catchment)	1D (stations)	KM	–	As, Cd, Cr, Cu, Mn, Ni, Pb, Zn, soil depth	Identification of pollutant patterns related to flooded levels and land-use types in Manwan (China)
(He et al., 2021)	Soil (surface and subsurface soil of different land uses)	2D (stations and depths)	KM	PMF	Cd, Cr, Cu, Pb, Zn	Evaluation of the spatial distribution and sources of toxic elements and relation with land use (China)
(Hossain Bhuiyan et al., 2021)	Soil (agricultural top soil)	1D (stations)	KM	PMF	Fe, Mn, Cr, Co, Ni, Cu, Zn, As, Pb, Cd	Identification of sources and spatial pattern of pollutants in agricultural soils (Bangladesh)
(Kebonye et al., 2021)	Soil (topsoil near a mine)	1D (stations)	KM, PCA	–	Cd, As, Pb, Sb, Zn, oxidizable carbon, pH	Mapping of pollutant hotspots in soils near a Pb—Ag ore mine (Czech Republic)
(Li et al., 2021)	Soil (topsoil near a lead- smelting factory)	1D (stations)	KM	PMF	Pb, Zn, Cu, As, Cr, Mn, Ni, Fe	Assessment of pollutant patterns in a residential area near a factory (China)
(Tao et al., 2017)	Soil (topsoil at the regional scale)	1D (stations)	–	PMF	PAHs	Identification of the spatial distribution and sources of PAHs in northern China
(Yang et al., 2014)	Soil (topsoil at the regional scale)	1D (stations)	–	–	Cu, Pb, Zn, Cd, Ni, Cr, Hg, As, and Mn	Identification of different environmental quality categories and anomaly detection (China)
(Yotova et al., 2018)	Soil (topsoil near industrial areas)	1D (stations)	HCA	–	As, Cd, Cr, Cu, Mn, Ni, Pb, Zn, pH, TOM, $CaCO_3$	Assessment of the spatial distribution of contaminants in an industrial area and relation to phytoavailability (Bulgaria)
(Åkesson et al., 2015)	Water (groundwater)	1D (stations)	PCA	–	pesticide pollution degree, well filter, water age group, redox state, aquifer confinement, aquifer type, land use	Evaluation of pesticide patterns in public supply wells in Sweden

(continued on next page)

Table 4 (*continued*)

Document	Environmental compartment/s	Sampling dimensions	2nd level abstraction	Independent multivariate technique/s	Pollutant/s and other variables	Main objective
(Amiri and Nakagawa, 2021)	Water (groundwater)	2D (stations, seasons)	KM	PCA	Mg^{2+}, Ca^{2+}, K^+, Na^+, NO_3^-, SO_4^{2-}, Cl^-, F^-, HCO_3^-, CO_3^{2-}, pH, TDS, Eh, EC, Al, As, B, Br, Ba, Rb, Si, Sr, U, Zn	Assessment of spatiotemporal variations of groundwater quality in a coastal aquifer (Iran)
(Astel and Małek, 2008)	Water (rainfall)	2D (stations, years)	KM	PCA	Cl^-, NO_3^-, SO_4^{2-}, F^-, NH_4^+, Na^+, K^+, Ca^{2+}, Fe^{2+}, Mg^{2+}, Mn^{2+}, Zn^{2+}	Characterization of pollution sources using rainfall sampling (Poland)
(Astel et al., 2016)	Water (coastal lakes)	2D (stations, years)	KM	–	SAL, WT, DO, Chl-a, pH, Cond., BOD, COD, Cl^-, Br^-, NO_2^-, NO_3^-, PO_4^{3-}, SO_4^{2-}, Na^+, K^+, Ca^{2+}, Mg^{2+}, NH_4^+, Ni, Cu, Zn, Fe, Mn	Spatiotemporal variation of pollutant patterns in intermittently open-closed coastal lakes in Poland
(Carafa et al., 2011)	Water (river)	1D (stations)	HCA, PCA	–	PAHs, pesticides, BTEX, NH_4^+, As, Ba, Be, Cd, Co, Cr, Cu, Hg, Ni, Pb, Sb, Se, Zn	Assessment and spatial distribution patterns of water toxicity (Spain)
(Chea et al., 2016)	Water (river)	2D (stations, seasons)	HCA	–	WT, Cond., TSS, pH, DO, Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, HCO_3^-, NO_3^-, TP, COD, total ammonia	Assessment of spatial variability of water quality in the Mekong Basin (Asia)
(Geng et al., 2021)	Water (lake)	2D (stations, years)	KM	–	pH, COD, BOD, NH_3, TP, TN	Assessment of the difference in water quality between urban and suburban rivers (China)
(Guo et al., 2020)	Water (lakes)	2D (stations, years)	HCA	PCA	WT, WD, SD, Cond., DO, pH, Tur, Cl^-, ALK, TH, TP, TN, N-NH_4, N-NO_3, Chl-a, TSS, COD, Cu, Zn, As, Ni, Cd, Pb, Co, Cr, Mn, Fe, Al	Assessment of water quality and pollutant contamination and relation to eutrophication of three lakes in China
(Jampani et al., 2018)	Water (groundwater)	2D (stations, years)	–	FA, HCA	pH, Cond., DO, HCO_3, H_2SO_4, Cl^-, Mg^{2+}, Ca^{2+}, F^-, PO_4^{3-}, NO_3^-, TDS, SO_4^{2-}, Na^+, K^+	Assessment of multi-annual water pollution levels in a river basin in China
(Jiang et al., 2021)	Water (river)	1D (stations)	KM	–	TDS, Mg^{2+}, Ca^{2+}, K^+, Na^+, NH_4^+, NO_3^-, SO_4^{2-}, Cl^-, F^-, HCO_3^-, CO_3^{2-}, As, Mn, Fe, V, Cu	Assessment of the difference in water quality between urban and suburban rivers (China)
(Ki et al., 2017)	Water (river)	2D (stations, years)	–	–	pH, DO, EC, BOD, COD, TOC, TN, TP, SS	Identification of water pollution hotspots in a tributary river network (South Korea)
(Lee et al., 2019)	Water (groundwater)	1D (stations)	fuzzy- KM	–	pH, Eh, DO, Cond., Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-, NO_3^-, SO_4^{2-}, HCO_3^-, SiO_2	Assessment of water quality in a complex urban groundwater network (South Korea)
(Lee et al., 2021)	Water (groundwater)	2D (stations, years)	HCA	PCA	NO_3^-, pH, SO_4^{2-}, Cl^-, F^-, Al, Mn, Pb, Zn, Fe, As, Cu, turbidity	Assessment of the land use effect on shallow groundwater contamination (South Korea)
(Li et al., 2020a)	Water (groundwater)	2D (station, 2 campaigns)	KM	–	TDS, pH, K^+, Na^+, Ca^{2+}, Mg^{2+}, NH_4^+, HCO_3^-, Cl^-, SO_4^{2-}, NO_3^-, F^-, Fe^{2+}, Mn, CO_2	Evaluation of spatiotemporal variation of groundwater quality in Beijing (China)
(Mao et al., 2021)	Water (groundwater)	1D (stations)	KM	PCA	Ca^{2+}, Mg^{2+}, Cl^-, Na^+, K^+, HCO_3^-, NH_4^+, NO_3^-, Fe, Mn, pH, Eh	Identification of spatial air quality patterns and nitrogen pollution (China)
(Nakagawa et al., 2020)	Water (groundwater)	2D (station, 2 campaigns)	HCA	–	Fe, Mn, dissolved SiO_2, NO_2^-, NO_3^-, SO_4^{2-}, Cl^-, F^-, Na^+, K^+, Mg^{2+}, Ca^{2+}, pH	Assessment of the earthquake effect on the groundwater quality (Japan)
(Noh et al., 2016)	Water (artificial reservoir)	2D (stations, years)	KM	PCA	WT, DO, Cond., pH, Chl-a, SPM, DOC, SO_4^{2-}, NO_3^-, Hg species, TN, TP	Identification of the correlation between the Hg species variation and specific water quality levels in artificial reservoirs South Korea
(Olkowska et al., 2014)	Water (river)	2D (station, seasons)	KM	–	pH, Cond., F^-, Cl^-, Br^-, NO_2^-, NO_3^-, SO_4^{2-}, PO_4^{3-}, Li^+, Na^+, NH_4^+, Mg^{2+}, Ca^{2+}, Be, Co, Cu, Mo, Ag, Cd, Sb, Tl, Pb, U, Ba, B, V, Zn, Cr, As, Se, Mn, Al, Ni, Sn, Cs, Rb, Sr, formaldehyde, PAHs, PCBs, TOC, COD	Evaluation of the impact of heavy industry and heat and electricity production on water quality (Poland)
(Orak et al., 2020)	Water (river)	2D (stations, years)	fuzzy-KM	–	NH_3, NO_3^-, NO_2^-, PO_4^{3-}, DO, BOD, T	Identification of water quality classes (Turkey)
(Souid et al., 2020)	Water (groundwater)	1D (stations)	HCA	–	pH, SAL, Cond., TDS, HCO_3^-, F^-, Cl^-, Br^-, NO_3^-, NO_2^-, SO_4^{2-}, Na^+, K^+, Ca^{2+}, Mg^{2+}, Li^+	Identification of salinization processes in groundwater in Jerba Island (Tunisia)
(Tobiszewski et al., 2010)	Water (river)	1D (stations)	KM	–	WT, PAHs, chlorinated solvents, NO_3^-, PO_4^{3-}, SO_4^{2-}	Identification of pollution sources and their spatial patterns (Poland)
(Torres-Martínez et al., 2021)	Water (coastal wells)	2D (stations and depths)	KM	–	T, pH, DO, total dissolved solids (TDS), and EC, Ca^{2+}, Mg^{2+}, Na^+, K^+, HCO_3^-, Cl^-, NO_3^-, SO_4^{2-}, isotopes	Determination of nitrate and sulfate pollution sources and transformation related to seawater intrusion (Mexico)
(Tsakovski et al., 2010a)	Water (river)	2D (stations, years)	–	–	pH, DO, oxidation ability, BOD, COD, dissolved matter, non-dissolved matter, Cl^-, $SO4_2^-$-S, NH_4^+-N, NO_3^--N, Fe^{2+}	Identification of spatial and temporal water quality (Bulgaria)
(Tsakovski et al., 2010b)	Water (runoff)	1D (stations)	KM	–	PAHs, PCBs, Zn, Cl^-, NO_3^-, SO_4^{2-}, Na^+, NH_4^+, K^+, Ca^{2+}	Assessment of pollution sources of contamination in runoff water from roofs (Poland)
(Tsuchihara et al., 2020)	Water (groundwater)	1D (stations)	HCA	–	Mg^{2+}, Ca^{2+}, K^+, Na^+, NO_3^-, SO_4^{2-}, Cl^-, ^{222}Rn, isotopes	Identification of groundwater recharge sources (Japan)
(Tudesque et al., 2008)	Water (river)	2D (stations, years)	HCA	–	Cond., BOD, DO, pH, SM, T, Cl^-, NO_3^-, NO_2^-, SO_4^{2-}, PO_4^{3-}, HCO_3^-, Na^+, NH_4^+, K^+, Ca^{2+}, NH_3	Assessment of water quality changes in a river basin during three decades (France)
(Wang et al., 2015b)	Water (river)	1D (stations, years)	KM	FA	pH, DO, BOD, COD, TSS, NH_3-N, Cd, Pb, Cr, Cu	Identification of spatiotemporal variation of pollutants in dependence of natural and

(Xiao et al., 2021)	Water (groundwater)	2D (stations, years)	KM	PCA	Mg^{2+}, Na^{+}, K^{+}, HCO_3^-, Cl^-, F^-, NO_3^-, SO_4^{2-}, SiO_2, As, Cr, Cu, Fe, Li, Mn, Ni, Rb, Se, V, U, Ag, Cd, Pb, Tl, As, B, Ba, F	anthropogenic sources (Taiwan) Source identification and pollution assessment in industrial areas (China)
(Yang et al., 2012)	Water (river)	2D (stations, years)	KM	PCA, HCA	pH, TN, WT, DO, TP, NO_3^-, NH_4^+-N, BOD, COD, COD_{Mn}, TSS, Cu, Pb, Zn, Cd, CrVI	Identification of spatiotemporal patterns in a complex river network (China)
(Yotova et al., 2021)	Water (river)	2D (stations, years)	KM	PCA, HCA	DO, pH, EC, NH_4^+, NO_3^-, NO_2^-, PO_4^{3-}, BOD, TSS	Surface water quality of a transboundary river (Bulgaria, Greece)
(Yu et al., 2021)	Water (marine)	2D (stations, years)	PCA	–	VSS, NH_3, turbidity, TP, TN, TKN, TIN, T, SS, Si, SD, SAL, phaeopigments, pH, NO_2^-, PO_4^{3-}, FC, *E. coli*, DO, Chl-a, NH_3-N, BOD	Assessment of spatiotemporal variations of water quality in Hong Kong
(Zhu et al., 2020)	Water (groundwater)	2D (stations, years)	KM	–	Mg^{2+}, Ca^{2+}, K^+, Na^+, NO_3^-, SO_4^{2-}, Cl^-, HCO_3^-, pH, TDS, As, Cr, Cu, Fe, Mn, Ni, Pb, As, Cd, isotopes	Assessment of the influence of heavy metal migration from mining activities (China)
(Kim et al., 2019)	Water /sediments (river)	1D (stations)	KM	–	WT, TP, TN, TOC,, TIC, Chl-a, Ni, Zn, Co, Se, Fe, Al, As, Cd, Pb, Cr, Cu, Mn	Characterization of spatial heterogeneity of distribution of pollutants (South Korea)
(Kumar et al., 2021)	Water/sediment (river)	2D (stations and depths)	KM	FA	Mn, Zn Ni, Pb, Fe, Cu, Co, Cr, Cd	Assessment and relation of pollution patterns in water and sediment (Fiji)
(Liu et al., 2019)	Water /sediments (river)	2D (stations, years)	KM	–	TSS, COD, TN, NH_3-N, NO_3^--N, NO_2^--N, TP, DO, Ni, Zn, Cu, Cr, Cd, Pb, As, Hg	Classification of water and sediment quality and relation to the distribution of the benthic community of a river in China
(Pandey et al., 2015)	Water/sediments (river)	2D (stations, years)	–	PCA	pH, Cond., DO, TSS, BOD, COD, TH, AlK, PO_4^{3-}, NO_3^-, TDS, Cr, Mn, Fe, Ni, Cu, Zn, Cd, Pb	Identification of sources of metal speciation in sediments and relation with water quality (India)
(Skwarzec et al., 2012)	Water /sediments (sea)	2D (stations, years)	–	HCA	^{210}Po, ^{238}U, $^{239+240}Pu$	Identification of spatial and temporal variations of radionuclides (Poland)
(Olawoyin et al., 2013)	Water (surface and underground water)/soil/sediments	1D (stations)	–	–	SO_4^{2-}, PO_4^{3-}, Zn, Cd, Cr, Cu, Pb, Ni, Mn, Fe, carcinogenic PAHs, non-carcinogenic PAHs, TPH, BTEX, pH, Cond.	Categorization of the water, soil and sediment quality in petrolchemical regions (Nigeria)

combining the SOM technique with the WQI. The three factors of WQI calculated for each water river sample were used for building the model. The SOM showed comparable results to both techniques, but the SOM provided better separation of the sampling stations according to the water quality classes. Furthermore, the HCA dendrogram was overcrowded, thus difficult to interpret. On the contrary, the hits plotted on the SOM map were easily readable. (Pandey et al., 2015) considered the pollutant exchange between water and sediment in a river in the proximity of several different industrial sources. The output analysis proved that SOM was more efficient in the characterization of the spatiotemporal distribution of pollutants and source identification.

4.3.5. Supervised artificial neural networks

In some air quality studies, the SOM results were compared with supervised methods based on artificial neural networks. (Wesolowski et al., 2006) used SOM for confirming MLP results and found very good agreement in sample classification by season. (Khedairia and Khadir, 2012) used SOM followed by KM for identifying clusters representing different meteorological classes. Then the authors used MLP for modeling the relationship between air pollutant concentrations and meteorological parameters within each cluster. (Bougoudis et al., 2014) used ML-ANN to assess the SOM reliability finding a correct classification of above 80 % and misclassification issues for severe pollutant events. (Chang et al., 2020) used SOM to identify high pollutant events and BPNN for predicting the events. They found that the prediction of high pollution events was less effective without previously classifying the samples by SOM.

4.4. Single-compartment studies

4.4.1. Spatial studies

Most of the studies aiming to assess the spatial distribution of pollutants was focused on sediment and soil compartments, while only a few of them concerned air or water ecosystems.

The clustering results were generally used for identifying gradients of contamination in the investigated area and possible hotspots. In some studies the sample classification was reported on the corresponding points in a geographical map by different symbols or colors (Carafa et al., 2011; Cheng et al., 2017; He et al., 2021; Hossain Bhuiyan et al., 2021; Lee et al., 2019; Mao et al., 2021; Nathan and Lary, 2019).

As a rule, the whole experimental dataset was used for SOM training. This approach has the advantage that one or some sample categories such as site distances or collection depth are usually hidden to the SOM algorithm. Thus, observing the classification proposed by the model and considering the known categories, the reliability of the SOM model can be assessed.

Alternatively, some papers presented specific dataset splitting according to the scope of the study. In (Tao et al., 2017) the distribution of PAHs in surface soil was examined. The SOM was applied both to the absolute concentration PAH dataset and to the proportion of each PAH to the total content. The former model was used to classify the samples, the latter was used to detect possible different sources. (Jiang et al., 2021) collected samples of two tributaries of the same river, one that crosses an urban area and the other in a suburban area. A single SOM model for each tributary was built to independently detect sources that affected water quality. The purpose of (Rogowska et al., 2014) was to determine the possible effects of a shipwreck on marine sediment contamination both by heavy metals and POPs. Moreover, ecotoxicity tests were performed. A SOM-based model was built using all the collected data for a first exploration of the site and pollutant distribution, then a SOM-based model was built separately for each level of the sediment core. The comparison among the models' outcomes allowed the identification of the pollutant spatial patterns and the most impacted areas as well as to interpret of the ecotoxicity data. (Licen et al., 2019) collected PM fractions data at two monitoring sites near an industrial facility and built an independent SOM model coupled with KM for each site. The approach allowed identifying similar and different sources impacting the sites as well as the background air profile. Several

studies used Piper or trilinear diagrams for characterizing water samples however, only two of them plotted the cluster classification obtained by SOM onto the diagram for a combined interpretation (Mao et al., 2021; Tsuchihara et al., 2020). The latter can give more insight into water class classification because it adjoins the results of two different techniques in a single figure. (Tsuchihara et al., 2020) also plotted on a geographical map hexadiagrams representing groundwater characteristics colored by cluster split. Moreover, the cluster assignment in graphs representing the specific isotopic ratios for identifying the sources that affected groundwater quality was highlighted. (Dai et al., 2018) collected surface and core sediments for assessing the relation in heavy metal content at different depths. The cluster obtained were represented in terms of the geo-accumulation index for each metal at each depth. The relations among the concentrations at different depths were obtained by exploring the distances among the corresponding nodes on the SOM map. (Tsakovski et al., 2009) built a SOM-based model using pollutant concentrations in lake sediments. Mortality and ecotoxicity indexes were not an input for the model but they were plotted mimicking a SOM heatmap according to the sample assigned to the nodes. In this way, the accordance and relationship between the pollutant distribution and the indexes were reliably assessed.

4.4.2. Spatiotemporal studies

Most studies aiming to assess the spatial distribution of pollutants was focused on air and water compartments, while only a few of them concerned soil or sediments. The SOM model was used to disclose spatial pattern in a similar way respect to the studies discussed in par. 4.4.1.

Typically, the whole experimental dataset was used for SOM training, thus the time component was hidden to the SOM algorithm. Thus, observing the classification proposed by the model and considering the known date/time of sample collection, the reliability of the SOM model in the identification of temporal patterns can be assessed. Several approaches for detecting temporal patterns in the data were used. Visualization techniques are usually preferred because they highlight information better than a table or a text description. In the following lines, the main visualization types found in the selected documents are presented.

Several papers annotated years (Geng et al., 2021; Yotova et al., 2021; Yu et al., 2021), months (Astel et al., 2013; Wesolowski et al., 2006; Yang et al., 2012), or seasons (Liu et al., 2019) on the SOM map. In (Licen et al., 2018, 2020a; Tudesque et al., 2008; Yu et al., 2021) the temporal evolution of representative compartment classes was visualized using arrow trajectories on the SOM map. (Crawford et al., 2016; Jiang et al., 2016; Liao et al., 2020; Wang et al., 2015b) presented the spatiotemporal variation of compartment classes in a combined visualization on a geographical map. (Olkowska et al., 2014) showed the distribution of the hits in clusters for different seasons in separate SOM maps. (Chea et al., 2016) represented the wet-dry season variation by boxplot of cluster profiles. In (Orak et al., 2020; Yang et al., 2012) the temporal variation in relation to clusters was depicted by scatter plots. (Chang et al., 2020) proposed a SOM map with nodes displaying a representation of yearly, seasonal, and daily scale pollutant concentrations presented by different types of graphs (bar plot, pie chart). (Pearce et al., 2014) represented on the SOM map the percentage of samples represented by each BMU according to years and season for assessing the differences in pollutant characterization when compared with the component planes. (de Oliveira et al., 2019) collected pollutant and meteorological data by using an equipped vehicle driven along five main urban roads of a Brazilian city. The SOM model was used for assessing variable correlations coupled to the recorded spatial positions of the vehicle during the trips for identifying the more impacted areas.

Few studies split the experimental data before SOM training. In (Amiri and Nakagawa, 2021) two separate models were built using data collected in a coastal aquifer during the dry and wet seasons, respectively. The obtained clusters were visualized on separate geographical maps for highlighting the temporal differences in water classes between the seasons. In (Li et al., 2020a) two SOM-based models were built dividing the groundwater samples into separated datasets according to the monitoring year. The possible different cluster classification of the same sampling site were explored for assessing water quality evolution over time. In (Licen et al., 2020a) particle counter data were collected for three years at an industrial site and a SOM-based model was separately built for each year. The comparison allowed following the evolution of the cluster profiles over time according to the changes in the management of the plant.

In most studies, a min-max or z-score normalization was applied. A different method was proposed by (Alvarez-Guerra et al., 2011) who used "limit value quotients" (LVq) i.e. the quotients obtained by dividing the pollutant concentration by its corresponding limit value present in the current legislation. In the study, the authors used the SOM coupled with KM to classify regulated pollutant distribution recorded by the Spain station network and foster air quality management in identified critical areas. (Nakagawa et al., 2020) aimed to assess the differences in groundwater quality before and after an earthquake. A concentration ratio was calculated by dividing the variable concentration measured just after the earthquake by an estimated concentration using 5-year data before the earthquake. The concentration ratio was used as input to the SOM calculations to evaluate the earthquake-induced effects on groundwater chemistry.

4.5. Multi-compartment studies

Few studies reported multi-compartment analysis. Most of them proposed a water/sediment combined analysis. Two illustrated a soil/sediment analysis and one a water/soil/sediment analysis.

(Liu et al., 2019) aimed to assess the main physicochemical factors that can affect the benthic community in a river basin. The whole water and sediment dataset was used to build the SOM model and, a separate model for the data regarding the benthic community was built as well. The SOM model on compartment data proved to be very effective in site classification. The comparison between the models was useful for identifying the water and sediment pollutants affecting the benthic community. (Pandey et al., 2015) considered the pollutant exchange between water and sediment in a river in the proximity of several different industrial sources. All the samples were used to build the SOM model. The SOM proved to be more efficient in the characterization of the spatial and temporal (monthly) distribution of pollutants and source identification than conventional chemometric tools (PCA and HCA). (Kim et al., 2019) studied a river estuary and its inner and outer bays. They collected physicochemical data for the water compartment and heavy metal data for the sediment one for each sampling site. The data were used together for building the model. The sites were clustered and reliable results were obtained in association with the spatial distribution of fish populations. (Kumar et al., 2021) analyzed the heavy metal content of water and sediments in a river. A SOM-based model was built for each compartment. The comparison of the component planes and the distribution of the sites on the SOM map were used for classifying the degree of pollution at each site. (Skwarzec et al., 2012) identified regions and seasons of increased inflow of radionuclides in two river catchments optimizing the radiochemical monitoring network. The SOM algorithm allowed the grouping of the sampling sites and following their evolution according to the sampling time for assessing hotspots and radionuclides sources.

(Cheng et al., 2017) proposed a SOM model built from a dataset composed of toxic metals content of soil (inundated and not-flooded) and sediment samples. The authors decided to test the SOM classification capabilities by feeding the algorithm with the whole dataset. The obtained model was effective in detecting site classification and variable correlation. Then, a model for each category was built. The latter was necessary to disclose severely polluted areas for the different categories that were not pointed out by the first model. (Carrillo et al., 2021) built a separate SOM model for each compartment (sediment and soil). The results independently obtained were compared. This method allowed the identification of the most impacted sites.

(Olawoyin et al., 2013) was the only study concerning three compartments. The authors analyzed river delta samples of water, soil, and sediments for assessing the possible effects of petrochemical plants. A SOM model was built separately for each category of variables and area. The

outcomes were compared both for variable behavior and sampling site classification. The SOM visualization capabilities allowed highlighting of zones of priority that might require additional investigations.

5. Conclusions

The SOM allows obtaining proper data classification without previous knowledge about the data (Simeonova et al., 2010) and it shows useful advantages with respect to other popular chemometric tools (Astel et al., 2007) as described in par. 3.4. Much more interpretation ability is gained when the SOM model is merged with clustering algorithms and assessment of the cluster's significance. The SOM outputs can be easily visualized on 2D dimensions to identify data patterns and relationships among variables describing complex systems in various environmental compartments. Moreover, while the SOM model is poorly affected by outliers, it is a useful method to detect and visualize them (Muñoz and Muruzábal, 1998; Muruzábal and Muñoz, 1997).

To date, several free software tools are available. Thus, this review aims to encourage researchers who have to deal with environmental pollution issues to use SOM for environmental pollution spatial and temporal distribution assessment. Since every study presented in a paper has to be reproducible, we suggest, as a guide for the authors, to: (i) clearly describe the SOM input dataset (number of samples, number of variables and, the data unfolding as described in par.4); (ii) clarify the data pretreatment used (min-max normalization, *Z*-score, log-scaling, other); (iii) list the choice/value of all the input parameters presented in Table 2; (iv) make explicit the software and tools that have been used for calculation; (v) use flowcharts for explaining how different multivariate techniques were employed in the study.

Moreover we recommend to take advantage of the exploration of all the outputs for extracting valuable information, as illustrated in Table 2. Finally, we encourage the users to try different visual representations for emphasizing the spatiotemporal variations of pollutants in the environmental compartments.

CRediT authorship contribution statement

conceptualization, S.L., A.A., S.T.; methodology, S.L., A.A., S.T.; investigation: S.L.; writing—original draft preparation, S.L., A.A., S.T.; writing—review and editing, S.L., A.A., S.T.; visualization, S.L.; supervision, S.L.

Data availability

Data will be made available on request.

Declaration of competing interest

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

Acknowledgments

This review was partially supported by the Pomeranian University in Słupsk (Grant no 7-4-3).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2023.163084.

References

Åkesson, M., Sparrenbom, C.J., Dahlqvist, P., Fraser, S.J., 2015. On the scope and management of pesticide pollution of swedish groundwater resources: the scanian example. Ambio 44, 226–238. https://doi.org/10.1007/s13280-014-0548-1.

Alvarez-Guerra, E., Molina, A., Viguri, J.R., Alvarez-Guerra, M., 2011. A SOM-based methodology for classifying air quality monitoring stations. Environ. Prog. Sustain. Energy 30, 424–438. https://doi.org/10.1002/ep.10474.

Alvarez-Guerra, M., González-Piñuela, C., Andrés, A., Galán, B., Viguri, J.R., 2008. Assessment of self-organizing map artificial neural networks for the classification of sediment quality. Environ. Int. 34, 782–790. https://doi.org/10.1016/j.envint.2008.01.006.

Amiri, V., Nakagawa, K., 2021. Using a linear discriminant analysis (LDA)-based nomenclature system and self-organizing maps (SOM) for spatiotemporal assessment of groundwater quality in a coastal aquifer. J. Hydrol. 603, 127082. https://doi.org/10.1016/j.jhydrol.2021.127082.

Aria, M., Cuccurullo, C., 2017. Bibliometrix: an R-tool for comprehensive science mapping analysis. J. Informetr. 11, 959–975. https://doi.org/10.1016/j.joi.2017.08.007.

Arias, R., Barona, A., Ibarra-Berastegi, G., Aranguiz, I., Elías, A., 2008. Assessment of metal contamination in dregded sediments using fractionation and self-organizing maps. J. Hazard. Mater. 151, 78–85. https://doi.org/10.1016/j.jhazmat.2007.05.048.

Astel, A., Małek, S., 2008. Multivariate modeling and exploration of environmental n-way data from bulk precipitation quality control. J. Chemom. 22, 738–746. https://doi.org/10.1002/cem.1156.

Astel, A., Tsakovski, S., Barbieri, P., Simeonov, V., 2007. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. Water Res. 41, 4566–4578. https://doi.org/10.1016/j.watres.2007.06.030.

Astel, A.M., Bigus, K., Obolewski, K., Glińska-Lewczuk, K., 2016. Spatiotemporal assessment of water chemistry in intermittently open/closed coastal lakes of southern Baltic. Estuar. Coast. Shelf Sci. 182, 47–59. https://doi.org/10.1016/j.ecss.2016.09.010.

Astel, A.M., Giorgini, L., Mistaro, A., Pellegrini, I., Cozzutto, S., Barbieri, P., 2013. Urban BTEX spatiotemporal exposure assessment by chemometric expertise. Water Air Soil Pollut. 224, 1503. https://doi.org/10.1007/s11270-013-1503-7.

Ballabio, D., Vasighi, M., 2012. A MATLAB toolbox for self organizing maps and supervised neural network learning strategies. Chemom. Intell. Lab. Syst. 118, 24–32. https://doi.org/10.1016/j.chemolab.2012.07.005.

Boelaert, J., Ollion, E., Sodoge, J., Megdoud, M., Naji, O., Lemba Kote, A., Renoud, T., Hym, S., 2021. CRAN - package aweSOM [WWW document]. URL https://cran.r-project.org/web/packages/aweSOM/index.html. (Accessed 5 March 2022).

Bougoudis, I., Iliadis, L., Spartalis, S., 2014. Comparison of self organizing maps clustering with supervised classification for air pollution data sets. IFIP Adv. Inf. Commun. Technol. 436, 424–435. https://doi.org/10.1007/978-3-662-44654-6_42.

Carafa, R., Faggiano, L., Real, M., Munné, A., Ginebreda, A., Guasch, H., Flo, M., Tirapu, L., der Ohe, P.C.V., 2011. Water toxicity assessment and spatial pollution patterns identification in a Mediterranean River Basin District. Tools for water management and risk analysis. Sci. Total Environ. 409, 4269–4279. https://doi.org/10.1016/j.scitotenv.2011.06.053.

Carrillo, K.C., Drouet, J.C., Rodríguez-Romero, A., Tovar-Sánchez, A., Ruiz-Gutiérrez, G., Viguri Fuente, J.R., 2021. Spatial distribution and level of contamination of potentially toxic elements in sediments and soils of a biological reserve wetland, northern Amazon region of Ecuador. J. Environ. Manag. 289. https://doi.org/10.1016/j.jenvman.2021.112495.

Chang, F.-J., Chang, L.-C., Kang, C.-C., Wang, Y.-S., Huang, A., 2020. Explore spatio-temporal PM2.5 features in northern Taiwan using machine learning techniques. Sci. Total Environ. 736, 139656. https://doi.org/10.1016/j.scitotenv.2020.139656.

Chapman, J., Truong, V.K., Elbourne, A., Gangadoo, S., Cheeseman, S., Rajapaksha, P., Latham, K., Crawford, R.J., Cozzolino, D., 2020. Combining chemometrics and sensors: toward new applications in monitoring and environmental analysis. Chem. Rev. 120, 6048–6069. https://doi.org/10.1021/acs.chemrev.9b00616.

Chea, R., Grenouillet, G., Lek, S., 2016. Evidence of water quality degradation in lower mekong basin revealed by self-organizing map. PLoS One 11, e0145527. https://doi.org/10.1371/journal.pone.0145527.

Chen, H., Wang, J., Chen, J., Lin, H., Lin, C., 2016. Assessment of heavy metal contamination in the surface sediments: a reexamination into the offshore environment in China. Mar. Pollut. Bull. 113, 132–140. https://doi.org/10.1016/j.marpolbul.2016.08.079.

Cheng, F., Liu, S., Yin, Y., Zhang, Y., Zhao, Q., Dong, S., 2017. Identifying trace metal distribution and occurrence in sediments, inundated soils, and non-flooded soils of a reservoir catchment using self-organizing maps, an artificial neural network method. Environ. Sci. Pollut. Res. 24, 19992–20004. https://doi.org/10.1007/s11356-017-9559-3.

Chon, T.-S., 2011. Self-organizing maps applied to ecological sciences. Eco. Inform. 6, 50–61. https://doi.org/10.1016/j.ecoinf.2010.11.002.

Clark, S., Sisson, S.A., Sharma, A., 2020. Tools for enhancing the application of self-organizing maps in water resources research and engineering. Adv. Water Resour. https://doi.org/10.1016/j.advwatres.2020.103676.

Crawford, J., Griffiths, A., Cohen, D.D., Jiang, N., Stelcer, E., 2016. Particulate pollution in the Sydney region: source diagnostics and synoptic controls. Aerosol Air Qual. Res. 16, 1055–1066. https://doi.org/10.4209/aaqr.2015.02.0081.

Dai, L., Wang, L., Li, L., Liang, T., Zhang, Y., Ma, C., Xing, B., 2018. Multivariate geostatistical analysis and source identification of heavy metals in the sediment of poyang Lake in China. Sci. Total Environ. 621, 1433–1444. https://doi.org/10.1016/j.scitotenv.2017.10.085.

de Oliveira, R.H., Carneiro, C.C., de Almeida, F.G.V., de Oliveira, B.M., Nunes, E.H.M., dos Santos, A.S., 2019. Multivariate air pollution classification in urban areas using mobile sensors and self-organizing maps. Int. J. Environ. Sci. Technol. 16, 5475–5488. https://doi.org/10.1007/s13762-018-2060-9.

Dupont, M.F., Elbourne, A., Cozzolino, D., Chapman, J., Truong, V.K., Crawford, R.J., Latham, K., 2020. Chemometrics for environmental monitoring: a review. Anal. Methods 12, 4597–4620. https://doi.org/10.1039/d0ay01389g.

Folguera, L., Zupan, J., Cicerone, D., Magallanes, J.F., 2015. Self-organizing maps for imputation of missing data in incomplete data matrices. Chemom. Intell. Lab. Syst. 143, 146–151. https://doi.org/10.1016/J.CHEMOLAB.2015.03.002.

Gavel, Y., Iselid, L., 2008. Web of science and scopus: a journal title overlap study. Online Inf. Rev. 32, 8–21. https://doi.org/10.1108/14684520810865958.
Geng, M., Wang, K., Yang, N., Li, F., Zou, Y., Chen, X., Deng, Z., Xie, Y., 2021. Evaluation and variation trends analysis of water quality in response to water regime changes in a typical river-connected lake (Dongting Lake), China. Environ. Pollut., 268 https://doi.org/10.1016/j.envpol.2020.115761.
Gontijo, E.S.J., Herzsprung, P., Lechtenfeld, O.J., de Castro Bueno, C., Barth, J.A.C., Rosa, A.H., Friese, K., 2021. Multi-proxy approach involving ultrahigh resolution mass spectrometry and self-organising maps to investigate the origin and quality of sedimentary organic matter across a subtropical reservoir. Org. Geochem., 151 https://doi.org/10.1016/j.orggeochem.2020.104165.
Govender, P., Sivakumar, V., 2020. Application of k-means and hierarchical clustering techniques for analysis of air pollution: a review (1980–2019). Atmos. Pollut. Res. 11, 40–56. https://doi.org/10.1016/j.apr.2019.09.009.
Gulson, B., Korsch, M., Dickson, B., Cohen, D., Mizon, K., Michael Davis, J., 2007. Comparison of lead isotopes with source apportionment models, including SOM, for air particulates. Sci. Total Environ. 381, 169–179. https://doi.org/10.1016/j.scitotenv.2007.03.018.
Guo, C., Chen, Y., Xia, W., Qu, X., Yuan, H., Xie, S., Lin, L.-S., 2020. Eutrophication and heavy metal pollution patterns in the water suppling lakes of China's south-to-north water diversion project. Sci. Total Environ. 711, 134543. https://doi.org/10.1016/j.scitotenv.2019.134543.
He, Y., Han, X., Ge, J., Wang, L., 2021. Multivariate statistical analysis of potentially toxic elements in soils under different land uses: spatial relationship, ecological risk assessment, and source identification. Environ. Geochem. Health https://doi.org/10.1007/s10653-021-00992-1.
Himberg, J., Ahola, J., Alhoniemi, E., Vesanto, J., Simula, O., 2001. The Self-Organizing Map as a Tool in Knowledge Engineering, pp. 38–65 https://doi.org/10.1142/9789812811691_0002.
Hopke, P.K., 2015. Chemometrics applied to environmental systems. Chemom. Intell. Lab. Syst. 149, 205–214. https://doi.org/10.1016/j.chemolab.2015.07.015.
Hossain Bhuiyan, M.A., Chandra Karmaker, S., Bodrud-Doza, M., Rakib, M.A., Saha, B.B., 2021. Enrichment, sources and ecological risk mapping of heavy metals in agricultural soils of dhaka district employing SOM, PMF and GIS methods. Chemosphere, 263 https://doi.org/10.1016/j.chemosphere.2020.128339.
Jampani, M., Huelsmann, S., Liedl, R., Sonkamble, S., Ahmed, S., Amerasinghe, P., 2018. Spatio-temporal distribution and chemical characterization of groundwater quality of a wastewater irrigated system: a case study. Sci. Total Environ. 636, 1089–1098. https://doi.org/10.1016/j.scitotenv.2018.04.347.
Jiang, N., Betts, A., Riley, M., 2016. Summarising climate and air quality (Ozone) data on self-organising maps: a Sydney case study. Environ. Monit. Assess. 188, 103. https://doi.org/10.1007/s10661-016-5113-x.
Jiang, Y., Gui, H., Li, C., Chen, J., Chen, C., Wang, C., Zhao, H., Guo, Y., Xu, J., Li, J., Qiu, H., 2021. Evaluation of the difference in water quality between urban and suburban rivers based on self-organizing map. Acta Geophys. 69, 1855–1864. https://doi.org/10.1007/s11600-021-00631-4.
Kalteh, A.M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. Environ. Model. Softw. 23, 835–845. https://doi.org/10.1016/J.ENVSOFT.2007.10.001.
Kebonye, N.M., Eze, P.N., John, K., Gholizadeh, A., Dajčl, J., Drábek, O., Němeček, K., Borůvka, L., 2021. Self-organizing map artificial neural networks and sequential gaussian simulation technique for mapping potentially toxic element hotspots in polluted mining soils. J. Geochem. Explor. 222. https://doi.org/10.1016/j.gexplo.2020.106680.
Khedairia, S., Khadir, M.T., 2012. Impact of clustered meteorological parameters on air pollutants concentrations in the region of Annaba, Algeria. Atmos. Res. 113, 89–101. https://doi.org/10.1016/j.atmosres.2012.05.002.
Ki, S.J., Song, S., Kang, T.W., Kim, S., Kang, T., Baek, S.G., Baek, J.H., Kim, J.H., 2017. Addressing water pollution hotspots in the tributary monitoring network using a nonlinear data analysis tool. Desalin. Water Treat. 77, 156–162. https://doi.org/10.5004/DWT.2017.20681.
Kim, D.-K., Jo, H., Han, I., Kwak, I.-S., 2019. Explicit characterization of spatial heterogeneity based on water quality, sediment contamination, and ichthyofauna in a riverine-to-coastal zone. Int. J. Environ. Res. Public Health 16, 409. https://doi.org/10.3390/ijerph16030409.
Kohonen, T., 2013. Essentials of the self-organizing map. Neural Netw. 37, 52–65. https://doi.org/10.1016/J.NEUNET.2012.09.018.
Kohonen, T., 2001. Self-Organizing Maps. Springer series in Information Sciences. Springer-Verlag.
Kohonen, T., 1987. Adaptive, associative, and self-organizing functions in neural computing. Appl. Opt. 26, 4910. https://doi.org/10.1364/ao.26.004910.
Kumar, S., Islam, A.R.M.T., Hasanuzzaman, M., Salam, R., Khan, R., Islam, M.S., 2021. Preliminary assessment of heavy metals in surface water and sediment in nakuvadra-Rakiraki River, Fiji using indexical and chemometric approaches. J. Environ. Manag. 298. https://doi.org/10.1016/j.jenvman.2021.113517.
Ladwig, R., Heinrich, L., Singer, G., Hupfer, M., 2017. Sediment core data reconstruct the management history and usage of a heavily modified urban lake in Berlin, Germany. Environ. Sci. Pollut. Res. 24, 25166–25178. https://doi.org/10.1007/s11356-017-0191-z.
Lee, C.M., Choi, H., Kim, Y., Kim, M.S., Kim, H.K., Hamm, S.Y., 2021. Characterizing land use effect on shallow groundwater contamination by using self-organizing map and buffer zone. Sci. Total Environ. 800. https://doi.org/10.1016/j.scitotenv.2021.149632.
Lee, K.-J., Yun, S.-T., Yu, S., Kim, K.-H., Lee, J.-H., Lee, S.-H., 2019. The combined use of self-organizing map technique and fuzzy c-means clustering to evaluate urban groundwater quality in Seoul metropolitan city, South Korea. J. Hydrol. 569, 685–697. https://doi.org/10.1016/j.jhydrol.2018.12.031.
Li, J., Shi, Z., Wang, G., Liu, F., 2020. Evaluating spatiotemporal variations of groundwater quality in Northeast Beijing by self-organizing map. Water (Switzerland) 12, 1382. https://doi.org/10.3390/W12051382.
Li, J., Wang, G., Liu, F., Cui, L., Jiao, Y., 2021. Source apportionment and ecological-health risks assessment of heavy metals in topsoil near a factory, Central China. Expo. Health 13, 79–92. https://doi.org/10.1007/s12403-020-00363-8.
Li, R., Hua, P., Zhang, J., Krebs, P., 2020. Effect of anthropogenic activities on the occurrence of polycyclic aromatic hydrocarbons in aquatic suspended particulate matter: evidence from Rhine and elbe Rivers. Water Res. 179, 115901. https://doi.org/10.1016/j.watres.2020.115901.
Liao, Z., Xie, J., Fang, X., Wang, Y., Zhang, Y., Xu, X., Fan, S., 2020. Modulation of synoptic circulation to dry season PM2.5 pollution over the Pearl River Delta region: an investigation based on self-organizing maps. Atmos. Environ. 230, 117482. https://doi.org/10.1016/j.atmosenv.2020.117482.
Licen, S., Barbieri, G., Fabbris, A., Briguglio, S.C., Pillon, A., Stel, F., Barbieri, P., 2018. Odor control map: self organizing map built from electronic nose signals and integrated by different instrumental and sensorial data to obtain an assessment tool for real environmental scenarios. Sensors Actuators B Chem. 263, 476–485. https://doi.org/10.1016/j.snb.2018.02.144.
Licen, S., Cozzutto, S., Barbieri, G., Crosera, M., Adami, G., Barbieri, P., 2019. Characterization of variability of air particulate matter size profiles recorded by optical particle counters near a complex emissive source by use of self-organizing map algorithm. Chemom. Intell. Lab. Syst. 190, 48–54. https://doi.org/10.1016/j.chemolab.2019.05.008.
Licen, S., Cozzutto, S., Barbieri, P., 2020a. Assessment and comparison of multi-annual size profiles of particulate matter monitored at an urban-industrial site by an optical particle counter with a chemometric approach. Aerosol Air Qual. Res. 20, 800–809. https://doi.org/10.4209/aaqr.2019.08.0414.
Licen, S., Di Gilio, A., Palmisani, J., Petraccone, S., de Gennaro, G., Barbieri, P., 2020b. Pattern recognition and anomaly detection by self-organizing maps in a multi month e-nose survey at an industrial site. Sensors (Switzerland) 20. https://doi.org/10.3390/s20071887.
Licen, S., Franzon, M., Rodani, T., Barbieri, P., 2021. SOMEnv: an R package for mining environmental monitoring datasets by self-organizing map and k-means algorithms with a graphical user interface. Microchem. J. 165. https://doi.org/10.1016/j.microc.2021.106181.
Liu, X., Zhang, J., Shi, W., Wang, M., Chen, K., Wang, L., 2019. Priority pollutants in water and sediments of a river for control basing on benthic macroinvertebrate community structure. Water (Switzerland) 11, 1267. https://doi.org/10.3390/w11061267.
Lu, H.-C., Chang, C.-L., Hsieh, J.-C., 2006. Classification of PM10 distributions in Taiwan. Atmos. Environ. 40, 1452–1463. https://doi.org/10.1016/j.atmosenv.2005.10.051.
Mao, H., Wang, G., Rao, Z., Liao, F., Shi, Z., Huang, X., Chen, X., Yang, Y., 2021. Deciphering spatial pattern of groundwater chemistry and nitrogen pollution in poyang Lake Basin (eastern China) using self-organizing map and multivariate statistics. J. Clean. Prod. 329, 129697. https://doi.org/10.1016/j.jclepro.2021.129697.
Mas, S., de Juan, A., Tauler, R., Olivieri, A.C., Escandar, G.M., 2010. Application of chemometric methods to environmental analysis of organic pollutants: a review. Talanta 80, 1052–1067. https://doi.org/10.1016/j.talanta.2009.09.044.
Misra, S., Li, H., He, J., 2020. Robust geomechanical characterization by analyzing the performance of shallow-learning regression methods using unsupervised clustering methods. Mach. Learn. Subsurf. Charact., 129–155 https://doi.org/10.1016/B978-0-12-817736-5.00005-3.
Mongeon, P., Paul-Hus, A., 2016. The journal coverage of web of science and scopus: a comparative analysis. Scientometrics 106, 213–228. https://doi.org/10.1007/s11192-015-1765-5.
Muñoz, A., Muruzábal, J., 1998. Self-organizing maps for outlier detection. Neurocomputing 18, 33–60. https://doi.org/10.1016/S0925-2312(97)00068-4.
Muruzábal, J., Muñoz, A., 1997. On the visualization of outliers via self-organizing maps. J. Comput. Graph. Stat. 6, 355–382. https://doi.org/10.1080/10618600.1997.10474748.
Nakagawa, K., Yu, Z.Q., Berndtsson, R., Hosono, T., 2020. Temporal characteristics of groundwater chemistry affected by the 2016 Kumamoto earthquake using self-organizing maps. J. Hydrol. 582. https://doi.org/10.1016/j.jhydrol.2019.124519.
Nathan, B.J., Lary, D.J., 2019. Combining domain filling with a self-organizing map to analyze multi-species hydrocarbon signatures on a regional scale. Environ. Monit. Assess. 191, 337. https://doi.org/10.1007/s10661-019-7429-9.
Neme, A., Hernández, L., 2011. Visualizing patterns in the air quality in mexico city with self-organizing maps. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 6731 LNCS, pp. 318–327 https://doi.org/10.1007/978-3-642-21566-7_32.
Noh, S., Kim, C.-K., Lee, J.-H., Kim, Y., Choi, K., Han, S., 2016. Physicochemical factors affecting the spatial variance of monomethylmercury in artificial reservoirs. Environ. Pollut. 208, 345–353. https://doi.org/10.1016/j.envpol.2015.09.054.
Olawoyin, R., Nieto, A., Grayson, R.L., Hardisty, F., Oyewole, S., 2013. Application of artificial neural network (ANN)-self-organizing map (SOM) for the categorization of water, soil and sediment quality in petrochemical regions. Expert Syst. Appl. 40, 3634–3648. https://doi.org/10.1016/j.eswa.2012.12.069.
Olkowska, E., Kudłak, B., Tsakovski, S., Ruman, M., Simeonov, V., Polkowska, Z., 2014. Assessment of the water quality of Kłodnica River catchment using self-organizing maps. Sci. Total Environ. 476–477, 477–484. https://doi.org/10.1016/j.scitotenv.2014.01.044.
Olteanu, M., Villa-Vialaneix, N., 2015. On-line relational and multiple relational SOM. Neurocomputing 147, 15–30. https://doi.org/10.1016/j.neucom.2013.11.047.
Orak, E., Akkoyunlu, A., Can, Z.S., 2020. Assessment of water quality classes using self-organizing map and fuzzy C-means clustering methods in Ergene River, Turkey. Environ. Monit. Assess., 192 https://doi.org/10.1007/s10661-020-08560-3.
Pandey, M., Pandey, A.K., Mishra, A., Tripathi, B.D., 2015. Application of chemometric analysis and self organizing map-artificial neural network as source receptor modeling for metal speciation in river sediment. Environ. Pollut. 204, 64–73. https://doi.org/10.1016/j.envpol.2015.04.007.
Pearce, J.L., Waller, L.A., Chang, H.H., Klein, M., Mulholland, J.A., Sarnat, J.A., Sarnat, S.E., Strickland, M.J., Tolbert, P.E., 2014. Using self-organizing maps to develop ambient air

quality classifications: a time series example. Environ. Heal. A Glob. Access Sci. Source 13, 56. https://doi.org/10.1186/1476-069X-13-56.
Pearce, J.L., Waller, L.A., Sarnat, S.E., Chang, H.H., Klein, M., Mulholland, J.A., Tolbert, P.E., 2016. Characterizing the spatial distribution of multiple pollutants and populations at risk in Atlanta, Georgia. Spat. Spatiotemporal. Epidemiol. 18, 13–23. https://doi.org/10.1016/j.sste.2016.02.002.
Ponmalai, R., Kamath, C., 2019. Self-organizing maps and their applications to data analysis. Lawerence Livermore Natl. Lab. 46. https://doi.org/10.2172/1566795.
Rogowska, J., Kudłak, B., Tsakovski, S., Wolska, L., Simeonov, V., Namieśnik, J., 2014. Novel approach to ecotoxicological risk assessment of sediments cores around the shipwreck by the use of self-organizing maps. Ecotoxicol. Environ. Saf. 104, 239–246. https://doi.org/10.1016/j.ecoenv.2014.03.025.
Romanić, S.H., Vuković, G., Klinčić, D., Antanasijević, D., 2018. Self-organizing maps for indications of airborne polychlorinated biphenyl (PCBs) and organochlorine pesticide (OCPs) dependence on spatial and meteorological parameters. Sci. Total Environ. 628–629, 198–205. https://doi.org/10.1016/j.scitotenv.2018.02.012.
Simeonova, P., Lovchinov, V., Dimitrov, D., Radulov, I., 2010. Environmetric approaches for lake pollution assessment. Environ. Monit. Assess. 164, 233–248. https://doi.org/10.1007/s10661-009-0888-7.
Skwarzec, B., Kabat, K., Astel, A., 2009. Seasonal and spatial variability of 210Po, 238U and 239 + 240Pu levels in the river catchment area assessed by application of neural-network based classification. J. Environ. Radioact. 100, 167–175. https://doi.org/10.1016/j.jenvrad.2008.11.007.
Skwarzec, B., Strumińska-Parulska, D.I., Boryo, A., Kabat, K., 2012. Polonium, uranium and plutonium radionuclides in aquatic and land ecosystem of Poland. J. Environ. Sci. Health A Tox. Hazard. Subst. Environ. Eng. 47, 479–496. https://doi.org/10.1080/10934529.2012.646153.
Souid, F., Telahigue, F., Agoubi, B., Kharroubi, A., 2020. Isotopic behavior and self-organizing maps for identifying groundwater salinization processes in Jerba Island, Tunisia. Environ. Earth Sci. 79, 175. https://doi.org/10.1007/s12665-020-8899-3.
Sun, X., Wang, H., Guo, Z., Lu, P., Song, F., Liu, L., Liu, J., Rose, N.L., Wang, F., 2020. Positive matrix factorization on source apportionment for typical pollutants in different environmental media: a review. Environ Sci Process Impacts 22, 239–255. https://doi.org/10.1039/c9em00529c.
Tao, S.-Y., Zhong, B.-Q., Lin, Y., Ma, J., Zhou, Y., Hou, H., Zhao, L., Sun, Z., Qin, X., Shi, H., 2017. Application of a self-organizing map and positive matrix factorization to investigate the spatial distributions and sources of polycyclic aromatic hydrocarbons in soils from Xiangfen County, northern China. Ecotoxicol. Environ. Saf. 141, 98–106. https://doi.org/10.1016/j.ecoenv.2017.03.017.
Team, R.C, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.
Tobiszewski, M., Tsakovski, S., Simeonov, V., Namieśnik, J., 2010. Surface water quality assessment by the use of combination of multivariate statistical classification and expert information. Chemosphere 80, 740–746. https://doi.org/10.1016/j.chemosphere.2010.05.024.
Torres-Martínez, J.A., Mora, A., Mahlknecht, J., Kaown, D., Barceló, D., 2021. Determining nitrate and sulfate pollution sources and transformations in a coastal aquifer impacted by seawater intrusion—A multi-isotopic approach combined with self-organizing maps and a bayesian mixing model. J. Hazard. Mater. 417. https://doi.org/10.1016/j.jhazmat.2021.126103.
Tsakovski, S., Astel, A., Simeonov, V., 2010a. Assessment of the water quality of a river catchment by chemometric expertise. J. Chemom. 24, 694–702. https://doi.org/10.1002/cem.1333.
Tsakovski, S., Kudłak, B., Simeonov, V., Wolska, L., Garcia, G., Namieśnik, J., 2012a. Relationship between heavy metal distribution in sediment samples and their ecotoxicity by the use of the hasse diagram technique. Anal. Chim. Acta 719, 16–23. https://doi.org/10.1016/j.aca.2011.12.052.
Tsakovski, S., Kudlak, B., Simeonov, V., Wolska, L., Namiesnik, J., 2009. Ecotoxicity and chemical sediment data classification by the use of self-organising maps. Anal. Chim. Acta 631, 142–152. https://doi.org/10.1016/j.aca.2008.10.053.
Tsakovski, S., Simeonova, P.A., Simeonov, V.D., 2012b. Statistical modeling of air pollution. J. Environ. Sci. Health A Tox. Hazard. Subst. Environ. Eng. 47, 31–43. https://doi.org/10.1080/10934529.2012.629576.
Tsakovski, S., Tobiszewski, M., Simeonov, V., Polkowska, Z., Namieśnik, J., 2010b. Chemical composition of water from roofs in Gdansk, Poland. Environ. Pollut. 158, 84–91. https://doi.org/10.1016/j.envpol.2009.07.037.
Tsuchihara, T., Shirahata, K., Ishida, S., Yoshimoto, S., 2020. Application of a self-organizing map of isotopic and chemical data for the identification of groundwater recharge sources in Nasunogahara alluvial fan, Japan. Water (Switzerland), 12 https://doi.org/10.3390/w12010278.
Tudesque, L., Gevrey, M., Grenouillet, G., Lek, S., 2008. Long-term changes in water physicochemistry in the adour-Garonne hydrographic network during the last three decades. Water Res. 42, 732–742. https://doi.org/10.1016/J.WATRES.2007.08.001.
Ultsch, A., Herrmann, L., 2007. The architecture of emergent self-organizing maps to reduce projection errors. ESANN 2005 Proc. - 13th Eur. Symp. Artif. Neural Networks, pp. 1–6.
Ultsch, A., Lötsch, J., 2017. Machine-learned cluster identification in high-dimensional data. J. Biomed. Inform. 66, 95–104. https://doi.org/10.1016/j.jbi.2016.12.011.
Vesanto, J., 1999. SOM-based data visualization methods. Intell. Data Anal. 3, 111–126. https://doi.org/10.3233/IDA-1999-3203.
Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Trans. Neural Netw. 11, 586–600. https://doi.org/10.1109/72.846731.
Veses, O., Mosteo, R., Ormad, M.P., Ovelleiro, J.L., 2014. Freshwater sediment quality in Spain. Environ. Earth Sci. 72, 2917–2929. https://doi.org/10.1007/s12665-014-3195-8.
Vignati, D.A.L., Secrieru, D., Bogatova, Y.I., Dominik, J., Céréghino, R., Berlinsky, N.A., Oaie, G., Szobotka, S., Stanica, A., 2013. Trace element contamination in the arms of the Danube Delta (Romania/Ukraine): current state of knowledge and future needs. J. Environ. Manag. 125, 169–178. https://doi.org/10.1016/j.jenvman.2013.04.007.
Wang, Y.-B., Liu, C.-W., Kao, Y.-H., Jang, C.-S., 2015a. Characterization and risk assessment of PAH-contaminated river sediment by using advanced multivariate methods. Sci. Total Environ. 524–525, 63–73. https://doi.org/10.1016/j.scitotenv.2015.04.019.
Wang, Y.-B., Liu, C.-W., Lee, J.-J., 2015b. Differentiating the spatiotemporal distribution of natural and anthropogenic processes on river water-quality variation using a self-organizing map with factor analysis. Arch. Environ. Contam. Toxicol. 69, 254–263. https://doi.org/10.1007/s00244-015-0167-2.
Wang, Z., Shen, Q., Hua, P., Jiang, S., Li, R., Li, Y., Fan, G., Zhang, J., Krebs, P., 2020. Characterizing the anthropogenic-induced trace elements in an urban aquatic environment: a source apportionment and risk assessment with uncertainty consideration. J. Environ. Manag. 275. https://doi.org/10.1016/j.jenvman.2020.111288.
Wehrens, R., Buydens, L.M.C., 2007. Self- and super-organizing maps in R: the kohonen package. J. Stat. Softw. 21, 1–19.. 10.18637/jss.v021.i05.
Wehrens, R., Kruisselbrink, J., 2018. Flexible self-organizing maps in kohonen 3.0. J. Stat. Softw. 87. https://doi.org/10.18637/jss.v087.i07.
Wesolowski, M., Suchacz, B., Halkiewicz, J., 2006. The analysis of seasonal air pollution pattern with application of neural networks. Anal. Bioanal. Chem. 384, 458–467. https://doi.org/10.1007/s00216-005-0197-0.
Wienke, D., Xie, Y., Hopke, P.K., 1995. Classification of airborne particles by analytical scanning electron microscopy imaging and a modified kohonen neural network (3MAP). Anal. Chim. Acta 310, 1–14. https://doi.org/10.1016/0003-2670(95)00128-M.
Wittek, P., Gao, S.C., Lim, I.S., Zhao, L., 2017. Somoclu: an efficient parallel library for self-organizing maps. J. Stat. Softw. 78.. 10.18637/jss.v078.i09.
Wu, D., Zewdie, G.K., Liu, X., Kneen, M.A., Lary, D.J., 2017. Insights into the morphology of the East Asia PM2.5 annual cycle provided by machine learning. Environ. Health Insights 11. https://doi.org/10.1177/1178630217699611.
Xiao, J., Wang, L., Chai, N., Liu, T., Jin, Z., Rinklebe, J., 2021. Groundwater hydrochemistry, source identification and pollution assessment in intensive industrial areas, eastern chinese loess plateau. Environ. Pollut. 278. https://doi.org/10.1016/j.envpol.2021.116930.
Yang, C., Guo, R., Wu, Z., Zhou, K., Yue, Q., 2014. Spatial extraction model for soil environmental quality of anomalous areas in a geographic scale. Environ. Sci. Pollut. Res. 21, 2697–2705. https://doi.org/10.1007/s11356-013-2200-1.
Yang, Y., Wang, C., Guo, H., Sheng, H., Zhou, F., 2012. An integrated SOM-based multivariate approach for spatio-temporal patterns identification and source apportionment of pollution in complex river network. Environ. Pollut. 168, 71–79. https://doi.org/10.1016/j.envpol.2012.03.041.
Ye, Z., Yang, J., Zhong, N., Tu, X., Jia, J., Wang, J., 2020. Tackling environmental challenges in pollution controls using artificial intelligence: a review. Sci. Total Environ. 699, 134279. https://doi.org/10.1016/j.scitotenv.2019.134279.
Yin, H., 2008. The self-organizing maps: background, theories, extensions and applications. Stud. Comput. Intell. 115, 715–762. https://doi.org/10.1007/978-3-540-78293-3_17/COVER.
Yin, H., Allinson, N.M., 1995. On the distribution and convergence of feature space in self-organizing maps. Neural Comput. 7, 1178–1187. https://doi.org/10.1162/NECO.1995.7.6.1178.
Yotova, G., Varbanov, M., Tcherkezova, E., Tsakovski, S., 2021. Water quality assessment of a river catchment by the composite water quality index and self-organizing maps. Ecol. Indic. 120. https://doi.org/10.1016/j.ecolind.2020.106872.
Yotova, G., Zlateva, B., Ganeva, S., Simeonov, V., Kudłak, B., Namieśnik, J., Tsakovski, S., 2018. Phytoavailability of potentially toxic elements from industrially contaminated soils to wild grass. Ecotoxicol. Environ. Saf. 164, 317–324. https://doi.org/10.1016/j.ecoenv.2018.07.077.
Yu, J., Tian, Y., Wang, X., Zheng, C., 2021. Using machine learning to reveal spatiotemporal complexity and driving forces of water quality changes in Hong Kong marine water. J. Hydrol. 603. https://doi.org/10.1016/j.jhydrol.2021.126841.
Zhong, B., Wang, L., Liang, T., Xing, B., 2017. Pollution level and inhalation exposure of ambient aerosol fluoride as affected by polymetallic rare earth mining and smelting in Baotou, North China. Atmos. Environ. 167, 40–48. https://doi.org/10.1016/j.atmosenv.2017.08.014.
Zhu, G., Wu, X., Ge, J., Liu, F., Zhao, W., Wu, C., 2020. Influence of mining activities on groundwater hydrochemistry and heavy metal migration using a self-organizing map (SOM). J. Clean. Prod. 257. https://doi.org/10.1016/j.jclepro.2020.120664.