Journal of Geochemical Exploration 242 (2022) 107084

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

A new version of the Langelier-Ludwig square diagram under a compositional perspective

Matthias Templ [a], Caterina Gozzi [b,*], Antonella Buccianti [b]

[a] Zurich University of Applied Sciences (ZHAW), Rosenstrasse 3, Winterthur, CH-8400 Zurich, Switzerland
[b] University of Florence, Dept. of Earth Sciences, Via G. La Pira 4, 50121 Firenze, Italy

ARTICLE INFO

Keywords:
Langelier-Ludwig diagram
Compositional data
Groundwater chemistry

ABSTRACT

The Langelier-Ludwig square diagram is a commonly used diagnostic tool in groundwater chemistry. Suitable groupings of cations and anions are selected and plotted as percentages of milliequivalents with the sums of the selected cations and anions plotted on the y- and x-axes, respectively. It displays relative ratios rather than absolute concentration whereby each axis ranges from 0 to 50 meq%. However, the sample space in which data are represented in a Langelier-Ludwig square diagram is indeed given by the simplex. Incorrect conclusions may be drawn when the compositional nature of compositional data is not taken into account, i.e., a change in one value in one component changes all other values due to due to the constant sum constraint of the measured chemical elements. Correlations are thus influenced by the presence of negative bias in the covariance structure and linear or nonlinear patterns on the square diagram can be misinterpreted. A new version of the Langelier-Ludwig square diagram based on a well-chosen coordinate representation of cations and anions is proposed. The advantage of the revised diagram is that all the information is contained in the log-ratios describing the intricate relationship between chemical species in aqueous solutions. It is shown that the geochemical interpretation of this new diagram – based on the relative dominance of major ions and distance from the (robust) barycenter of the data – provides a better and unbiased understanding of water-environment interactions. To further aid interpretation, (robust) tolerance ellipses show the correlation structure in the new version of the Langelier-Ludwig square diagram, and clustering algorithms can be applied to divide the data into groups beforehand. A bunch of different plotting options and interactive representations complete the implementation in free open-source software. It is recommended to replace the classic Langelier-Ludwig diagram with the new version.

1. Introduction

The chemical composition of water is a multi-component system, and several graphical-numerical methods have been proposed in the literature to describe this complexity. Most proposals have attempted to relate the chemical concentration of major ions to weathering and dissolution processes, and to the minerals from which the ions are derived in a particular lithologic context. Chemical reactions controlled by thermodynamic equilibria affect natural waters (acid-base equilibrium, complexation, solubility, oxidation, etc.) and determine the distribution of ionic species. This is done by taking into account their properties, which are well described by the position in the periodic table and the value of the ionic potential (Railsback, 2003). Typically, different diagrams are used to distinguish groups of similar data to reconstruct the presence of geochemical facies and potential mixing pathways and allow their subsequent spatial representation. In this framework, the chemistry of the world's surface water, expressed as a function of rainfall and seawater chemistry, rock weathering, and evaporation, was represented by Gibbs (1970) using a binary diagram comprising TDS (Total Dissolved Solids, in mg/L) and the ratios $Na^+/(Na^+ + Ca^{2+})$ or $Cl^-/(Cl^- + HCO_3^-)$. The Stiff diagram (Stiff, 1970) instead shows the concentrations (in milliequivalents) of the major ions (both cations and anions) as a polygonal shape describing the relative abundance of the various species with respect to the total. A similar representation can also be achieved by the double bar graphs proposed by Collins (1923), where the bars with respect to the horizontal zero line represent the proportional abundance of compounds (anions and cations separately) in equivalents per liter for different types of water. Finally, the Piper diagram (Piper,

* Corresponding author.
E-mail address: caterina.gozzi@unifi.it (C. Gozzi).

https://doi.org/10.1016/j.gexplo.2022.107084
Received 13 June 2022; Received in revised form 9 August 2022; Accepted 16 September 2022
Available online 26 September 2022

1944) consists of two trilinear diagrams in the lower section representing the relative concentrations of cations and anions, and a diamond-shaped diagram in the middle combining the cation and anion proportions. Various proposals and variants of the Piper diagram were described by Durov (1948). Recently, it has been modified to represent large volumes of hydrochemical analyses (Merino et al., 1944) and a revision in light of the compositional data analysis approach (Aitchison, 1982) has been proposed by Shelton et al. (2018). A new type of major ion plot based on conversion of data to *ilr* coordinates was also adopted by Engle and Rowan (2014), which replicates the type of information contained in Durov diagrams.

The idea of interpreting the water analysis on the basis of the main components in a simple graphical representation goes back to Hill (1940) and Langelier and Ludwig (1942). The last authors proposed the well-known "square" diagram. According to Langelier and Ludwig (1942), the main chemical constituents of water can be divided into four groups, each of which has chemically similar properties: c_1, consisting of the alkali cations sodium and potassium, c_2, consisting of the hardness cations calcium and magnesium, a_1, consisting of the strongly acidic anions nitrate, chloride, and sulfate, and a_2, consisting of the weakly acidic anions carbonate and bicarbonate. The electrical neutrality of salt solutions requires that $c_1 + c_2 = a_1 + a_2$, giving the square diagram the special properties of a non-Euclidean space. Implications of interpreting the Langelier-Ludwig diagram in the light of the Compositional Data Theory were also revealed by Buccianti and Magli (2011) by comparing confidence regions inside the constrained space with real ones. The greatest differences between metrics occur when approaching the limit of the simplex, where sample attribution to different quadrants could be misplaced, thereby affecting the geochemical interpretation (Buccianti and Magli, 2011). These problems demonstrate the need for revised versions of the LL (Langelier-Ludwig) diagram that take into account the proper sample space while still allowing an easy geochemical interpretation of the results. Despite this, very little research effort has been conducted in this regard for the LL diagram, while new versions were proposed for the Piper and Durov diagrams (Shelton et al., 2018; Engle and Rowan, 2014).

In this paper, the LL diagram is considered in two different ways: 1) as a 50 closed simplex in which it is mandatory to follow the rules of compositional data analysis to identify natural groups, anomalous data, and robust confidence ellipses in it; and 2) as a corresponding compositional version of the binary diagram with axes represented by isometric log-ratio coordinates, mapping a real space in which the identification of linear or nonlinear patterns or mixture paths follows the Euclidean geometry. The application examples concern the geochemistry of groundwaters collected in different areas of central Italy. To perform both analyses, the new methods of this work were implemented in the free and open source *R* software environment (R Development Core Team, 2022), specifically in the `robCompositions` package (Templ et al., 2011).

2. Materials and methods

2.1. Geochemical data set

The Tuscany region in central Italy is bordered to the north and south by the orogenic belt of the Northern Apennines, formed by the pressure phase from the Cretaceous to the Miocene associated with the collision of the European and African plates. An early compressional phase led to the overthrust of the Tuscan units and the Ligurian units on the Umbrian-Marchean units (Oligocene-Miocene) and to the uplift of the Apennine chain (Miocene-Pliocene/Pleistocene). Subsequently, an extensional phase (upper Miocene to upper Pleistocene) overlaid the thrust structures and resulted in several NW-SE-oriented graben systems that were filled by Neogene sediments (Lavecchia, 1990). The crustal thinning and the high geothermal gradient associated with this extensional phase led to the emplacement of intrusive bodies and volcanic edifices. Consequently, hydrothermal activity and various geothermal fields developed in Tuscany (e.g., Larderello and Mt. Amiata) with several thermal springs and CO_3^{2-}-rich gas vents discharging in central–southern Tuscany (Minissale, 2004). As a result of this tectonic evolution, the hydrogeological setting of the region is primarily characterized by i) an impermeable cover of marls and clays with some permeable layers allowing a limited flow of water, ii) a carbonate–evaporitic complex that contains the exploited geothermal reservoirs, and iii) a crystalline rock complex with a lower permeability (Fig. 1). The most typical lithologies are represented by Paleozoic metamorphic rocks (e.g., phyllitic to quartzitic and micaschist rocks), Mesozoic (Triassic evaporitic anyhydrites) and Cenozoic carbonate and evaporitic formations, which are overlaid by flysch series, and cross-cut or covered by Neogene to Quaternary granite intrusions and volcanic conduits, respectively.

Two different databases were considered for the analysis. The first one refers to the main different lithological and environmental conditions covering most of the region (blue dots), while the second relates to a small area around the city of Arezzo (purple dots; Fig. 1). It is expected that the first database is more heterogeneous compared to the second, which allows searching for natural groups through cluster analysis. The geochemical analysis of the waters from Tuscany region was investigated in Nisi et al. (2016a) with the aim to find a baseline from a compositional point of view. The database, which covers most of the regional area, contains the chemical composition of 2033 samples for which pH values, electrical conductivity (EC), and coordinates are known in addition to the concentration in meq/L of Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, HCO_3^-, Cl^-, SO_4^{2-}. Hydrochemical facies are mainly (about 70%) dominated by Ca-HCO_3 with Total Dissolved Solids (TDS) varying from 29 to 9400 mg/L (median: 667 mg/L). The same scheme is proposed for the more spatially limited Arezzo database, which contains the chemistry of about 367 samples. No values below the detection limit or zero values are reported in either data set. In case of their presence, it is advisable to refer to Martin-Fernandez et al. (2015), Templ et al. (2016) and Chen et al. (2018) to approach the problem from the compositional perspective.

2.2. The classic Langelier-Ludwig diagram

The sample space of the data represented in a LL diagram is a simplex closed to 50 and constructed from the major constituents of natural waters such as Na^+, K^+, Ca^{2+}, Mg^{2+}, HCO_3^-, CO_3^{2-}, Cl^-, SO_4^{2-}. Sometimes nitrate NO_3^- may also be included, while HCO_3^- and CO_3^{2-} together are considered the alkalinity of the water, i.e., the measure of the ability of the water body to neutralize acids and bases (buffering capacity) and thus maintain a fairly stable pH. The coordinates of sample points in the diagram are calculated by default construction on a forced ion balance basis (Langelier and Ludwig, 1942). Accordingly, the parameters to obtain the square are given by the standard equations R_1 (Eq. (1)), which represents the relationships between cations and R_2 (Eq. (2)) between anions:

$$R_1(Na^+ + K^+) = \frac{(Na^+ + K^+)}{Na^+ + K^+ + Ca^{2+} + Mg^{2+}} \times 50, \quad (1)$$

$$R_2\left(HCO_3^- + CO_3^{2-}\right) = \frac{\left(HCO_3^- + CO_3^{2-}\right)}{HCO_3^- + CO_3^{2-} + Cl^- + SO_4^{2-}} \times 50. \quad (2)$$

Consequently:

$$R_3\left(Ca^{2+} + Mg^{2+}\right) = 50 - (Na^+ + K^+), \quad (3)$$

$$R_4\left(Cl^- + SO_4^{2-}\right) = 50 - \left(HCO_3^- + CO_3^{2-}\right). \quad (4)$$

The classical diagram is constructed considering the horizontal axis R_2 (or R_4) and the vertical axis R_1 (or R_3) as shown in Fig. 2. Some geochemical interpretations can also be made here, depending on where

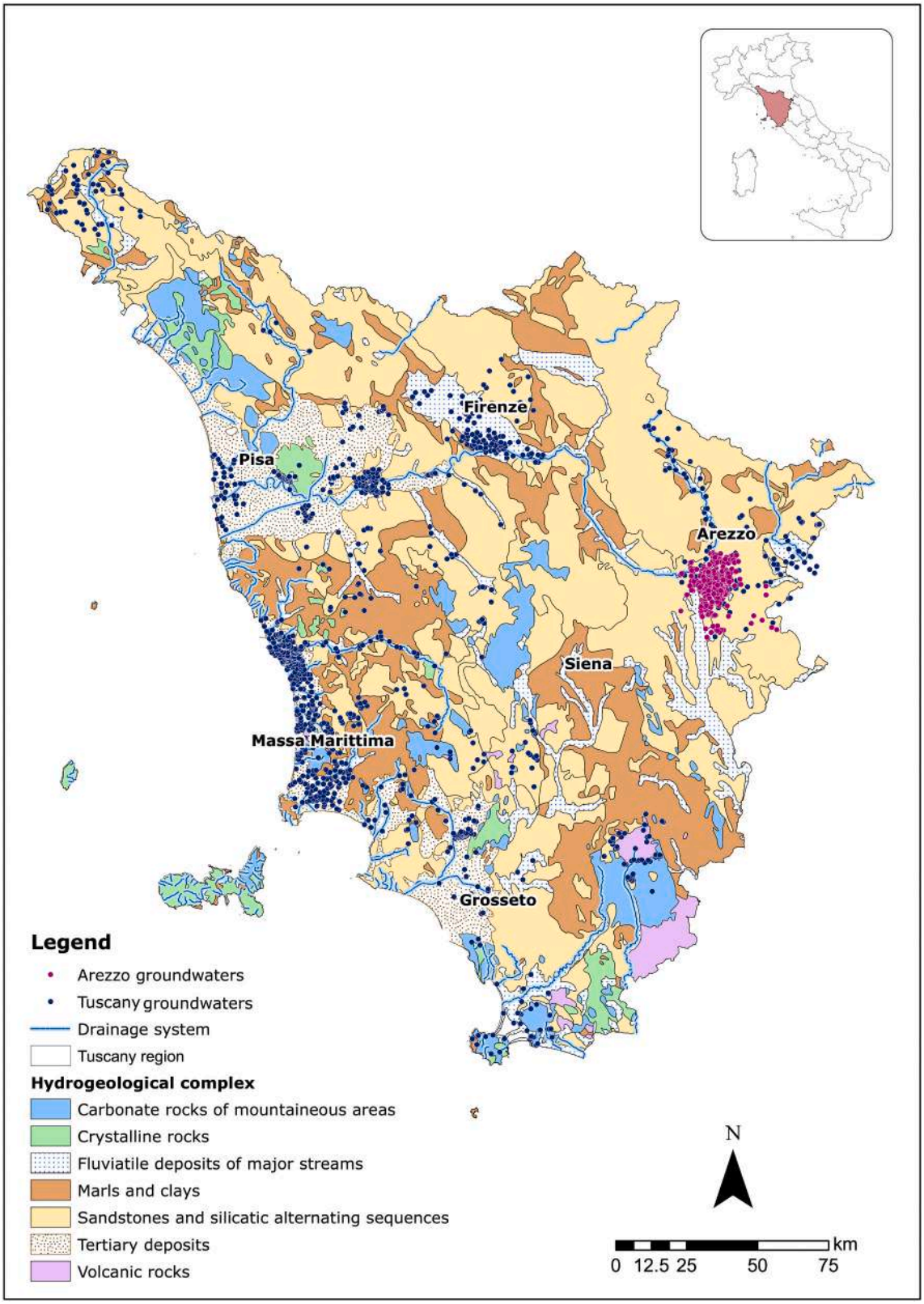


Fig. 1. Schematic map of the main hydrogeological complexes in the Tuscany region and location of groundwater samples. The layer of hydrological complexes was derived from ISPRA Ambiente (2017).

the cases fall in the different parts of the plane (e.g. Minissale et al., 2000).

2.3. Coordinate representation of compositions for the new diagram

Compositional data describe the (say D) parts of a whole and are usually represented as vectors of proportions, percentages,

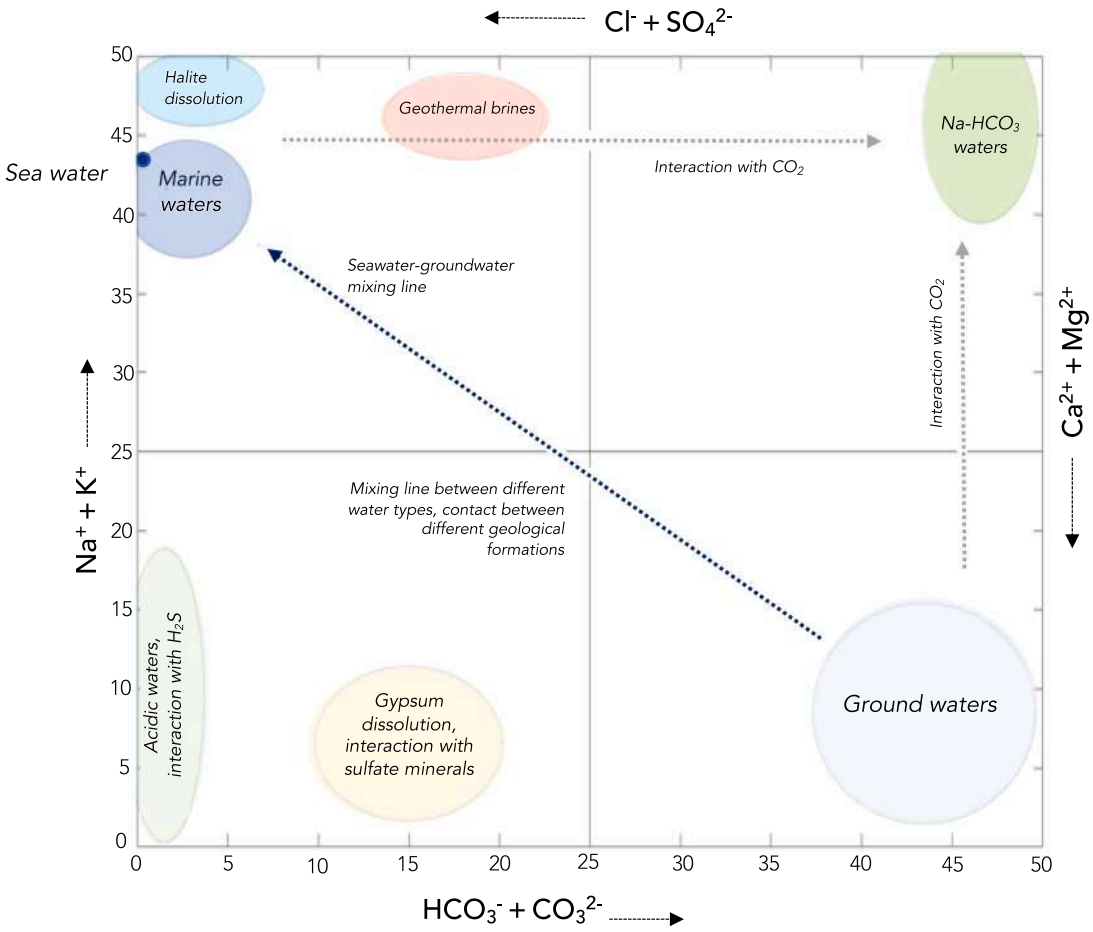


Fig. 2. Classical Langelier-Ludwig diagram with indications for some geological interpretations. The blue dot represents a typical seawater composition. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

concentrations, or frequencies. In the case of groundwater chemical composition, a composition represents the magnitude of chemical elements in mg/L of a groundwater sample. Not all values of a part are possible, i.e. not every value is possible to ensure that the sum from the equation below holds. For example, the higher Ca-HCO_3, the lower the contributions of the other parts. Therefore, it is easy to see that a negative bias (Pawlowsky-Glahn et al., 2007; Filzmoser and Hron, 2008a; Templ and Templ, 2020) is introduced in the correlation between parts, which one can show that it leads to unreliable and biased results when analyzing data with non-compositional methods (Templ and Templ, 2021). More precisely, a D-part simplex is defined as (see Filzmoser et al., 2018):

$$\mathbb{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D)^T, x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\} \tag{5}$$

with n compositional vectors $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})^T$ in $\mathbb{S}^D$, $i = 1, \ldots, n$.

It is easy to see that the simplex $\mathbb{S}^D (\mathbb{S}^D \notin \mathbb{R}^D)$ is characterized by its own geometry, the Aitchison geometry. Note that each composition vector $\mathbf{x}$ can be rescaled by a constant c and then the compositions $\mathbf{x}$ and $\mathbf{x} = c\mathbf{x}$ are compositionally equivalent. It can be shown that the exact value of κ does not matter in modern analysis of compositional data, and it can also be different for each composition (Filzmoser et al., 2018).

Awareness of the problems associated with the (classical, non-compositional) statistical analysis of compositional data goes back to Pearson (1897), who was concerned with spurious correlations, and Chayes (1960), who discovered the basic constraints affecting the variance-covariance structure of a compositional data matrix. Aitchison (1982) recognized that compositions provide information about relative, not absolute, values of parts or components. The consequence is that any statement about a composition can be stated in terms of ratios of components (Aitchison, 1986).

Log-ratios are more manageable in mathematical calculations and provide a one-to-one mapping to a real space. Various log-ratios transformations have been proposed in the literature. Around 2000, the realization of the algebraic-geometric structure of the sample space of compositions (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001) led to the staying-in-the-simplex approach based on the principle of working in coordinates (Mateu-Figueras et al., 2011). Following this tendency, compositions are represented by orthonormal coordinates living in a real Euclidean space corresponding to the simplex sample space of compositions.

There are several ways to define orthonormal bases in the simplex. Our approach uses a sequential binary partition (SBP) of a compositional vector (Egozcue and Pawlowsky-Glahn, 2005). The motive is that such bases can easily lead to interpretive terms for grouped parts of the composition. The Cartesian coordinates of a composition in such a basis are called balances, while the composition vectors that make up the balance are named elements of the basis. An SBP is a hierarchy of the parts of a composition. In the first order of the hierarchy, all parts are divided into two groups. In the following steps, each group is again divided into two groups. The process continues until all groups have a single part. For the kth order partition, it is possible to define the balance between the two subgroups formed at that level: if $i_1, i_2, \ldots, i_r$ are the r parts of the first subgroup (coded by $+1$) and $j_1, j_2, \ldots j_s$ the s parts of the second (coded by -1), the balance is defined as the normalized log-ratio of the geometric mean of each group of parts:

$$z_r = \sqrt{\frac{rs}{r+s}} ln \frac{(x_{i1}, x_{i2} \ldots x_{ir})^{1/r}}{(x_{j1}, x_{j2} \ldots x_{js})^{1/s}} = ln \frac{(x_{i1}, x_{i2} \ldots x_{ir})^{a_+}}{(x_{j1}, x_{j2} \ldots x_{js})^{a_-}}, \quad (6)$$

where

$$a_+ = +\frac{1}{r}\sqrt{\frac{rs}{r+s}}, \quad a_- = -\frac{1}{s}\sqrt{\frac{rs}{r+s}}. \quad (7)$$

The term a_+ refers to parts in the numerator, a_- to parts in the denominator, and the values of r and s correspond to the k-th order partition. It is important to note that in an SBP process, the change of sign codes from + to − and vice versa in each step of the partition causes only the change of sign of the associated balance and that the order in which the balances are arranged is arbitrary. If all selected variables are measured, the sign matrix given in Table 1a can be used to encode the sequential binary partition to obtain the orthonormal coordinates of the LL diagram. In case CO_3^{2-} is not available in the dataset, as in this case study, the SPB encoding reported in Table 1b should be considered instead.

The possibility to work in the real space using the coordinate representation from Eq. (6) facilitates the identification of groups in the data and their modeling and, consequently, the application of inferential methods. The b_5 and b_2 balances of the Table 1b were used to visualize the corresponding real space of the simplex. More precisely, expressing $\mathbf{X}$, a compositional data matrix with n observations and D compositional parts, in coordinates by applying Eq. (6) using b_5 and b_2 yields the $n \times 2$ matrix $\mathbf{Z}$ with observations $\mathbf{z}_1, \ldots, \mathbf{z}_n$.

2.4. Further enhancements

2.4.1. Robust tolerance ellipses

The proposed new visualization is first improved by computing and visualizing the robust confidence ellipse around the barycenter for one or more data sets distinguished by an available criterion (geography, lithology, etc., or after grouping with a clustering algorithm). Standard multivariate procedures assume that the majority of $\mathbf{Z}$ observations are generated by a multivariate normal distribution with mean vector μ and covariance matrix Σ. For a location estimator $\mathbf{t}$ and a covariance estimator $\mathbf{C}$, the squared Mahalanobis distances between the observations expressed in coordinates and the respective location estimator $\mathbf{t}$,

$$MD(\mathbf{z}_i)^2 = (\mathbf{z}_i - \mathbf{t})' \mathbf{C}^{-1} (\mathbf{z}_i - \mathbf{t}), \quad \text{for } i = 1, \ldots, n, \quad (8)$$

approximately follow a χ^2 distribution with $D - 1$ degrees of freedom, χ^2_{D-1} (Filzmoser et al., 2018). Ellipses expressing the covariance of a data set, $\mathbf{Z}$, consisting of the points with constant Mahalanobis distances, i.e. constant multivariate distance to the location estimate $\mathbf{t}$. Note that certain quantile of this distribution, like the 0.975-th quantile, ${\chi_{D-1;0.975}}^2$, can be used as a cut-off value to identify multivariate outliers as observations (Filzmoser and Hron, 2008b).

The classical arithmetic mean (vector) and the (Pearson) sample covariance matrix are unsuitable as estimators of $\mathbf{t}$ and $\mathbf{C}$ in Eq. (8) in the presence of outliers, since they are then themselves spoiled by the outliers. Thus, $\mathbf{t}$ and $\mathbf{C}$ should be robust estimators, like the MCD estimators $\mathbf{t}_{MCD}$ and $\mathbf{C}_{MCD}$ or the MM-estimator $\mathbf{t}_{MM}$ and $\mathbf{C}_{MM}$ (Maronna et al., 2006). The MCD is an iterative algorithm to find those k out of n observations with the smallest determinant of the covariance matrix; k is typically set to $0.5 \cdot n$ (50 % of the data can then be outliers), but can be increased up to n. In the latter most extreme (non-robust) case, the estimator equals the Pearson covariance estimate. While the MCD estimator assigns 0/1 weights to observations, the MM estimator assigns continuous weights to observations. The lower the weight, the more likely an observation is considered an outlier. The MM-estimator is the most efficient robust estimator and also allows up to 50 % outliers. Further details on the MM-estimator can be found in Maronna et al. (2006).

The calculations were performed in the corresponding coordinate representation obtained by applying the isometric log-ratio transformation (Eq. (6) and Table 1a) to the main chemical components of water. They can then be transformed back to the simplex. The subsequent step was to plot the results of a cluster analysis in a squared diagram when the data are not grouped according to certain criteria. For each group, the robust 95% confidence ellipse around the barycenter was also reported.

2.4.2. Clustering

The proposed new visualization is secondly improved by the possibility of finding groups of compositions with a clustering algorithm. The implementation includes two methods, k-means and model-based clustering. k-means clustering can be chosen because of its simplicity and fast computational speed. However, model-based clustering methods generally provide better clustering results (Templ et al., 2008). Model-based clustering is computationally more complex than other clustering algorithms such as k-means because the covariance of each cluster must also be estimated at each step of the iterative procedure. However, this is not a burden for typical groundwater chemistry data sets because these datasets tend to be *small*, i.e., do not consist of several thousand compositions.

Model-based clustering uses a statistical model for the shape of the clusters. In the standard model, the distribution of a compositional cluster is assumed to have the density of a multivariate normal distribution on the simplex (Eq. (5)) with a given location and covariance. The procedure can be either applied on:

a) the isometric coordinate representation of all input variables (Na^+, K^+, Ca^{2+}, Mg^{2+}, HCO_3^-, CO_3^{2-}, Cl^-, SO_4^{2-}). Hereby, the clustering is applied on $\mathbf{z}_1, \ldots, \mathbf{z}_6$ related to the balances b_1 to b_6.
b) $\mathbf{Z} = \{\mathbf{z}_5, \mathbf{z}_2\}$, see Eq. (6). Hereby, the clustering is applied on b_5 and b_2, thus aiming to mark clusters that are visually visible in the compositional LL-diagram.

Both representations have their advantages. If the clusters visually visible in the compositional LL diagram are to be marked with an automatic procedure, the second approach is preferable. If the information of all log-ratios of the variables shall be included and the result of the clustering on these coordinate representations shall be displayed in

Table 1
Sign matrices used to encode a sequential binary partition and to form an orthonormal basis for the Langelier-Ludwig diagram.
(a) All variables measured.
(b) CO_3^{2-} not available.

(a)

Ions	b_1	b_2	b_3	b_4	b_5	b_6	b_7
Na^+	+1	+1	+1	0	0	0	0
K^+	+1	+1	-1	0	0	0	0
Ca^{2+}	+1	-1	0	+1	0	0	0
Mg^{2+}	+1	-1	0	-1	0	0	0
HCO_3^-	-1	0	0	0	+1	+1	0
CO_3^{2-}	-1	0	0	0	+1	-1	0
Cl^-	-1	0	0	0	-1	0	+1
SO_4^{2-}	-1	0	0	0	-1	0	-1

(b)

Ions	b_1	b_2	b_3	b_4	b_5	b_6
Na^+	+1	+1	+1	0	0	0
K^+	+1	+1	-1	0	0	0
Ca^{2+}	+1	-1	0	+1	0	0
Mg^{2+}	+1	-1	0	-1	0	0
HCO_3^-	-1	0	0	0	+1	0
Cl^-	-1	0	0	0	-1	+1
SO_4^{2-}	-1	0	0	0	-1	-1

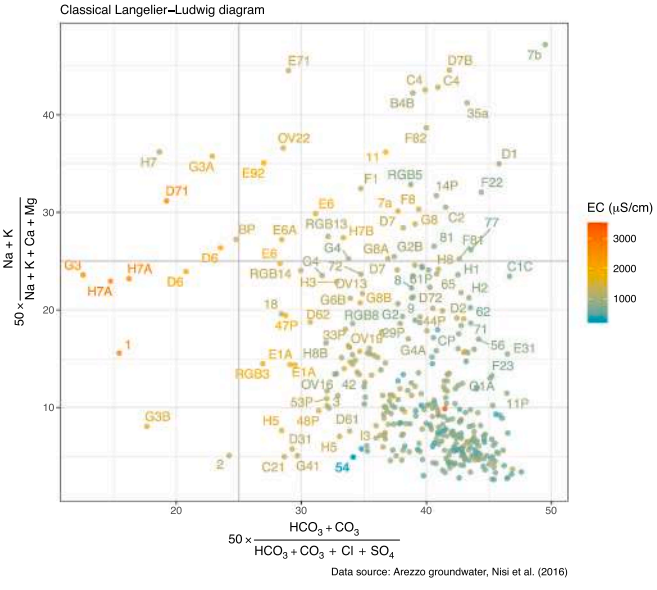
the LL diagram, the first approach is advantageous and aims at showing additional aspects that were not included in the LL diagram before.

It is assumed that the data consists of k clusters, generated by multivariate normal densities with expectations $\boldsymbol{\mu}_j$ and covariance matrices $\boldsymbol{\Sigma}_j$, for $j = 1, \ldots, k$. Especially, if D gets larger, many parameters need to be estimated from the available data, which can lead to instability. For this reason, the covariance estimations can be simplified by imposing constraints on the covariance structures of the clusters, ranging from allowing only the same covariances of the same size and shape in each cluster to k clusters with different shapes, sizes, and orientations. The algorithm itself optimizes both together, the optimal set of constraints and the optimal number of clusters based on a Bayesian information criterion proposed by Fraley and Raftery (1998), but - if desired - both can be overwritten manually. For a detailed description of the algorithm for model-based clustering, we refer the reader to Fraley and Raftery (1998) and for its application to cluster analysis of compositional data to Templ et al. (2008).

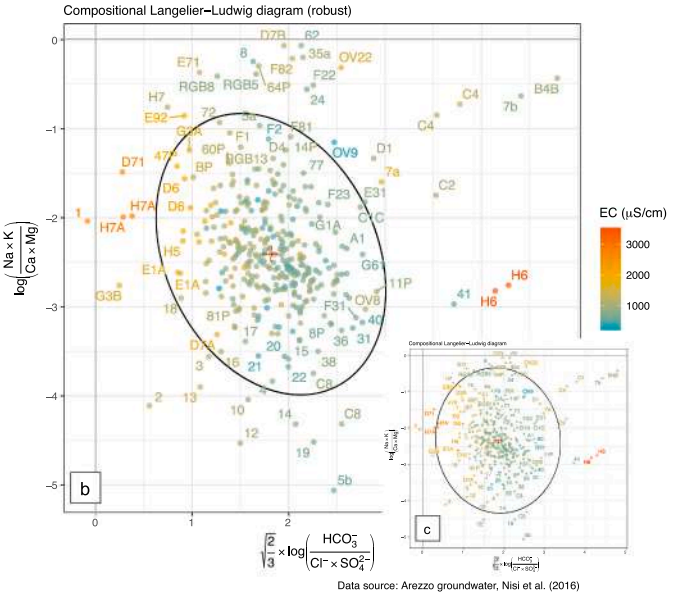
2.5. Plotting options

The heart of this contribution is to propose new visualizations for a compositional LL diagram. The visualizations - implemented in `robCompositions` - allow for flexibility and one can choose between many options:

- points-shapes / labels / points-shapes and labels / labels only if there is enough space, otherwise points-shapes
- tolerance ellipses (no / based on classical estimates / based on robust estimates), and level (default equals 97.5 %)
- pre-defined groups
- clustering (yes with many options / no)
- interactivity with mouse overlays supporting additional information and zooming (yes / no)
- coloring (with many options):
  - according to defined variables in the data set
  - according to a clustering result
  - according to the outlyingness of an observation, i.e. its robust weights / robust Mahalanobis distances.
  - according to the position in the compositional LL diagram and closeness to the center. Here, the points are projected onto a bivariate gradient color space that results in different colors in four quadrants of the plot. A green color is assigned if the values for both isometric coordinates, z_1 and z_2, are small. The color fuchsia is assigned to points with both having large values. Orange is selected when the value of z_1 is small and the value of z_2 is large,



(a) Classic Langelier-Ludwig diagram for Arezzo groundwater with partial point labeling and colour gradient related to conductivity (EC μS/cm) of the water sample.



(b) The compositional version of the Langelier-Ludwig diagram for Arezzo groundwater data with robust 97.5% robust tolerance ellipse and (c) with classical 97.5% tolerance ellipse.

Fig. 3. Classic and compositional Langelier-Ludwig diagrams for Arezzo groundwater.

and blue when the z_1 is large and the z_2 is small. Values closer to the extremes are saturated.
- plot (classical / robust) center
- transparency of points-shapes and labels

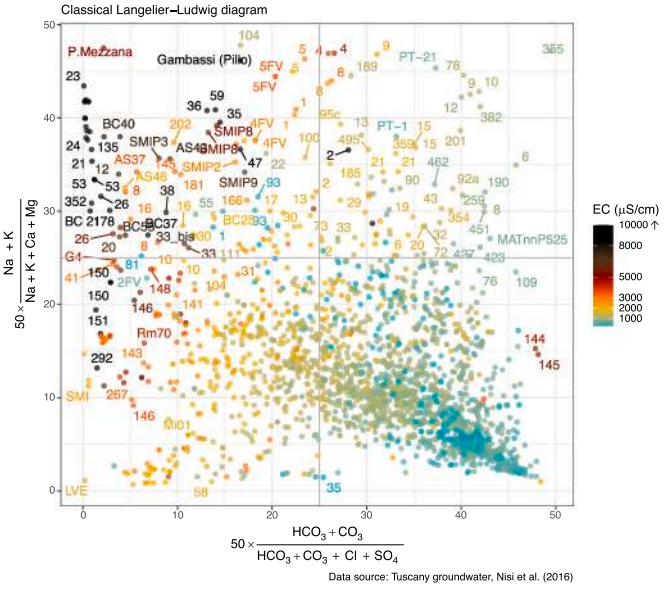
3. Results

3.1. Classic and compositional Langelier-Ludwig diagram with tolerance ellipses

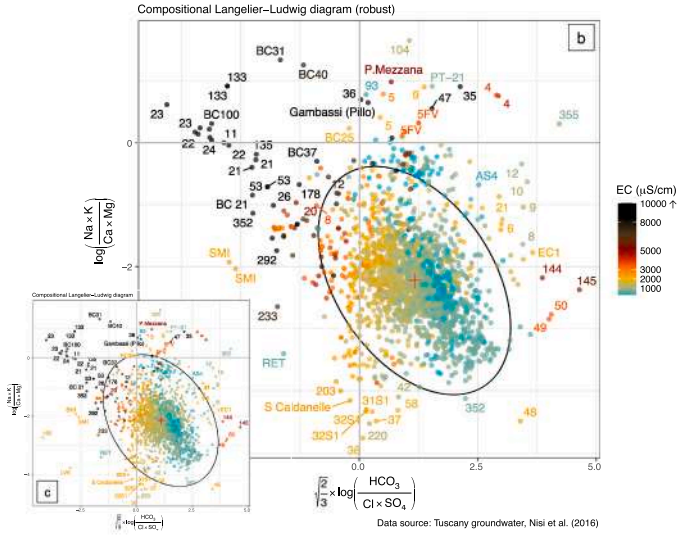
3.1.1. Arezzo groundwater

Firstly, for the sake of comparison, the classic LL diagram for Arezzo groundwater data is shown in Fig. 3a. Data points are partially labeled to allow for better interpretation and the color of points is related to conductivity, which is representative of the total dissolved solids content in groundwater and is related to the redox potential. Groundwater samples are typically characterized by a Ca(Mg)-HCO_3 hydrochemical facies followed by Na(K)-HCO_3, Ca(Mg)-SO_4 and Na(K)-Cl. In general, EC values tend to increase as one moves from the bottom-right corner of the plot, which represents a typical Ca(Mg)-HCO_3 groundwater composition, to the other quadrants. This is true except for two outliers located in the right lower part of the LL diagram with very high conductivity (up to 3000 µS/cm). Accordingly, the increasing EC values seem to well predict the transition to the other geochemical facies (Fig. 3a), and especially to Ca(Mg)-SO_4 and Na(K)-Cl. It is worth noting, that low EC values (i.e., <500 µS/cm) are extremely rare, while most of the samples have a median EC of 870 µS/cm. All of the presented plots have interactive mouse overlays that display information about each observation, but of course, this interactive feature cannot be displayed in a printed version of the plot.

Fig. 3b shows the compositional version of the LL diagram including a robust estimate of the covariance displayed by 97.5 % tolerance ellipses. The vertical and horizontal thick gray lines mark the values of zero on both axes. If the log-ratios are zero, then the ratios $HCO_3^-/(Cl^- \times SO_4^{2-})$ and $Na^+ \times K^+/(Ca^{2+} \times Mg^{2+})$ must be equal to 1. Accordingly, values above zero on the x-axis indicate a dominance of HCO_3^-, while values below zero on the y-axis indicate a dominance of $Ca^{2+} \times Mg^{2+}$. This situation characterizes all Arezzo groundwaters except for one sample (see label 1) in which $Cl^- \times SO_4^{2-}$ is predominant. Typically, samples with values of $HCO_3^-/(Cl^- \times SO_4^{2-})$ below average have larger conductivity. However, two outlying observations having very high conductivity and $HCO_3^-/(Cl^- \times SO_4^{2-})$ above average are detected (see label H6). These outliers are the same as those highlighted in the classical LL diagram (Fig. 3a). To illustrate the difference between the classical and the implemented robust covariance estimation, the same diagram is represented in Fig. 3c with the classical 97.5 % tolerance



(a) Classic Langelier-Ludwig diagram for Tuscany groundwater data with partial point labeling. The color of points is related to the electrical conductivity (EC μS/cm) of the water sample.



(b) The compositional version of the Langelier-Ludwig diagram for Tuscany groundwater data with partial point labeling, colour related to conductivity and robust 97.5% robust tolerance ellipse. (c) The same plot represented with classical 97.5% tolerance ellipse.

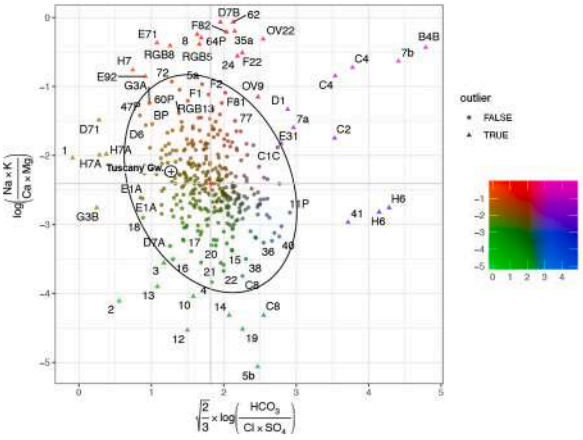
Fig. 4. Classic and compositional Langelier-Ludwig diagrams for Tuscany groundwater.

ellipse. The robust confidence ellipse collapses diagonally showing a negative correlation between the considered log-ratios. Thus, higher Cl^- and SO_4^{2-} concentrations are most likely associated with increased Na^+ and K^+. Differently, the classical confidence ellipse has a nearly circular appearance indicating that the two log-ratios are almost uncorrelated. Moreover, the plot shows that classical 97.5 % covariance estimation excludes a smaller number of observations, whereas the robust one identifies a greater amount of extreme values.

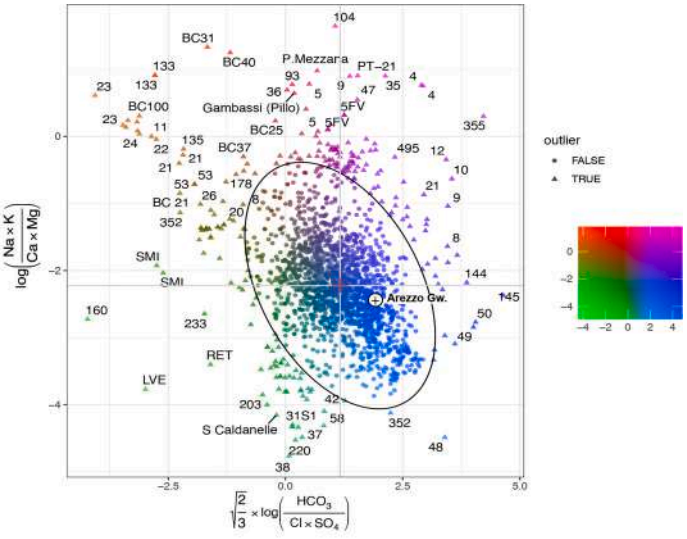
3.1.2. Tuscany groundwater

For comparison purposes, the classic LL diagram for the groundwater dataset of Tuscany is shown in Fig. 4a. EC gradient scale has been adjusted manually in order to be comparable with that of Fig. 3a. Most of the points are again in the lower right quadrant, indicating that groundwater samples are typically characterized by a Ca(Mg)-HCO_3 hydrochemical facies. Afterward, the most typical hydrofacies is Ca(Mg)-SO_4 followed by Na(K)-HCO_3 and Na(K)-Cl. Extremely high EC values (above 10,000 μS/cm) are mainly found for groundwater samples located in the upper left quadrant characterized by a composition near to that of marine waters (e.g., see labels 23, 24, 135) or close to that of geothermal brines (see labels 59, 35 and Gambassi). Several samples exceeding the 10,000 μS/cm are also found in the Ca(Mg)-SO_4 quadrant, close to the area typical of acidic waters. In contrast, only occasional points (including 2, 144 and 145) with EC above 5000 μS/cm are located in the right quadrants.

Fig. 4b shows the compositional version of the LL diagram with a robust estimate of the covariance indicated by 97.5 % tolerance ellipses. According to the plot, most waters are dominated by HCO_3^- on the x-axis and Ca^{2+} and Mg^{2+} on the y-axis. However, several samples display relative enrichment in the other ions. Particularly, in the upper-left corner of the diagram, a trend towards an increasing dominance of ($Cl^- \times SO_4^{2-}$) and ($Na^+ \times K^+$) is clearly visible. It is worth mentioning that, on the revised LL diagram, linear or nonlinear patterns or mixture paths can be interpreted using Euclidean geometry, whereas this is not the case for its classic version. In the same way as the Arezzo groundwaters, samples with values of $HCO_3^-/(Cl^- \times SO_4^{2-})$ below average tend to have larger conductivities. Nevertheless, some outlying observations with very high conductivity and $HCO_3^-/(Cl^- \times SO_4^{2-})$ above average are also detected (e.g. 4, 49 and 144). The robust 97.5 % confidence ellipse collapses diagonally in this case as well, demonstrating a negative



(a) The compositional version of the Langelier-Ludwig diagram for Arezzo groundwater data with partial point labeling, a four continuous colour scheme related to areas of the LL diagram, robust 97.5% tolerance ellipse and different shapes of points (outliers as triangles, non-outliers as filled circles). The circled + represents the barycenter of Tuscany groundwaters.



(b) The compositional version of the Langelier-Ludwig diagram for Tuscany groundwater data with partial point labeling, a four continuous colour scheme related to areas of the LL diagram, robust 97.5% tolerance ellipse and different shapes of points (outliers as triangles, non-outliers as filled circles). The circled + represents the barycenter of Arezzo groundwaters.

Fig. 5. Compositional Langelier-Ludwig diagrams for with bivariate gradient color scale indicating the closeness to the robust center.

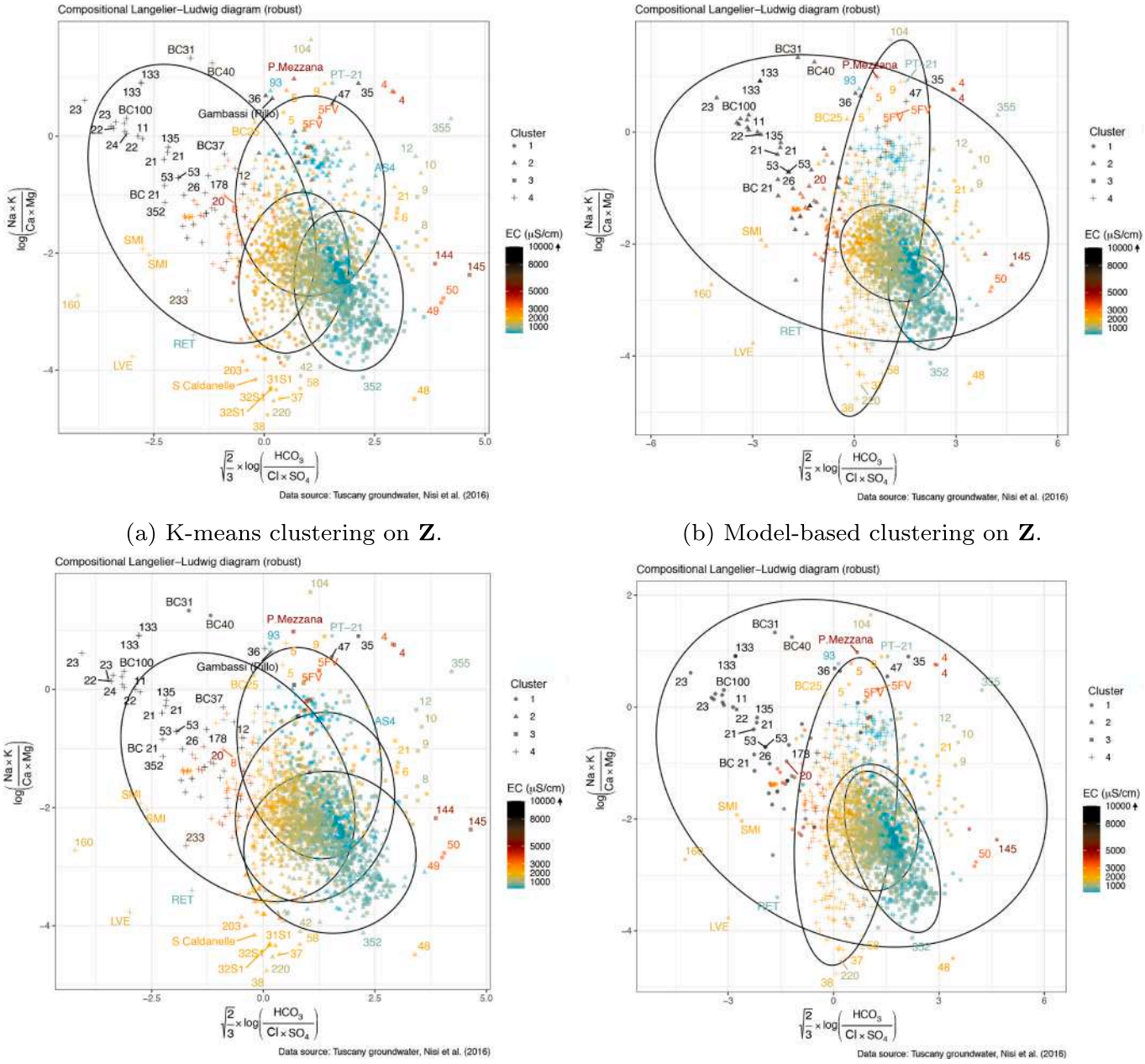
correlation between the considered log-ratios. Nevertheless, here the difference between the classical and robust 97.5 % confidence ellipses is less apparent compared to Fig. 4b-c, even though a greater number of extreme values is recognized by the robust confidence ellipse. An example of this is the group of blue samples (EC less than 1000 μS/cm) located near the upper-right corner of the plot that lies outside the robust confidence ellipse but within the classical one.

3.2. *Arezzo and Tuscany groundwater: distance from the robust compositional barycenter*

A further implementation regarding the plotting options is the possibility to color points according to the position in the compositional LL diagram and closeness to the center. The difference of Fig. 5a to 3b is the color coding and shape of symbols. Points are projected to a bivariate gradient color space, thus obtaining different colors for the four quadrants of the plot according to the values of the isometric coordinates, see Section 2.5.

This is especially useful in combination with a map, not shown here, using the same colors of groundwater samples as those displayed in Fig. 5a. In the bivariate gradient color space, the values closer to the extremes are saturated compared with those nearer the barycenter.

As for Arezzo ground waters, colors are mostly de-saturated, indicating an overall closeness to the barycenter. The few samples with saturated colors primarily tend towards shades of orange and fuchsia (dominance of $(Na^+ \times K^+)$ over $(Ca^{2+} \times Mg^{2+})$), while shades of green and blue are less common (Fig. 5a). In a similar manner, Fig. 5b presents the same results as Fig. 4b but with points projected to a bivariate



(a) K-means clustering on $\mathbf{Z}$.

(b) Model-based clustering on $\mathbf{Z}$.

(c) K-means clustering on all isometric log-ratio coordinates obtained through balances $b_1, \ldots, b_6$.

(d) Model-based clustering on all isometric log-ratio coordinates obtained through balances $b_1, \ldots, b_6$.

Fig. 6. Results of k-means and model-based clustering plotted on the LL diagram using robust tolerance ellipses.

gradient color scale. From the visual inspection of the LL diagram for Tuscany groundwater, it is apparent that colors are overall more saturated, suggesting that a greater number of samples are located further away from the barycenter. It is noteworthy that the interpretation depends on the barycenter of the considered data set. This means that the four quadrants obtained cannot be compared with those of the classical LL diagram, but their positions vary depending on the barycenter of the data. In order to better illustrate this point, the barycenter of Arezzo ground waters is reported as a circled + on the diagram of Tuscany ground waters and vice versa (Fig. 5a-b). From the diagrams, it can be seen that the centers of the two datasets are quite closer. However, with respect to Arezzo, the barycenter of Tuscany groundwater appears to be displaced more towards the upper-left side (see Fig. 5a). This indicates that Tuscany samples are, in general, slightly more enriched in ($Cl^- \times SO_4^{2-}$) and ($Na^+ \times K^+$) than those from Arezzo.

3.3. Clustering analysis on Tuscany groundwater

With the use of clustering algorithms, the proposed visualization of the LL diagram is further extended by allowing the detection of groups of compositions, as described in Section 2.4.2. The results of clustering **Z** are shown in Fig. 6a-b while those resulting from the clustering applied to all isometric log-ratio coordinates are illustrated in Fig. 6c-d. The shape of points corresponds to the cluster membership of a sample, the tolerance ellipses show the robust covariance in all four clusters and the color of points is related to the EC. The graphs at the top correspond directly to the visible information in the LL diagram, while the diagrams at the bottom incorporate all isometric log-ratio coordinates into the cluster analysis. This means that additional aspects describing the interlinks between the parts of the composition are taken into consideration. The k-means clustering (Fig. 6a-c) shows more clearly separated clusters, while the model-based clustering results (Fig. 6b-d) take into account the covariances of the individual clusters and can thus achieve more elliptical clusters. However, the results show that there is no clear clustering structure in the Tuscany groundwater data. In case a clear clustering structure exists, the resulting groups can also be mapped in order to visualize their geographic distribution and examine potential driving processes.

4. Discussion

4.1. What is the danger of interpreting the classic Langelier-Ludwig diagram?

One of the main dangers of interpreting the classic LL diagram is that the increase of one element would automatically decrease the other elements. Since compositional data are proportions, a variation in one component changes the relative amount of every other and this is especially true if one element is dominant, as commonly happens for Ca^{2+} and HCO_3^- in groundwaters. As a result, the classic LL diagram is influenced by variations in the values of these elements. There are many recent examples in literature in which the classic LL diagram is used for water classification purposes and attempts are made to identify end-members, linear or nonlinear patterns and correlations based on the diagram (e.g., Bhat et al., 2016; Cangemi et al., 2019; Safari et al., 2020). Nevertheless, in the light of our results, supported by the Theory of Compositional Data Analysis, this could lead to non-exhaustive interpretations or even misinterpretations due to the constant sum constraint. Similar problems are also documented for classical ternary or binary geochemical diagrams (e.g., Pawlowsky-Glahn et al., 2015; Buccianti, 2015). The new proposed version of the square diagram aims to solve these issues allowing to work on a real space.

Note that the compositional LL diagram does not permit an absolute classification of waters in hydrochemical facies but instead displays relative variations in major ion content with respect to the barycenter of the log-ratios (red +). In fact, it is impossible to find a single point on the compositional diagram that corresponds to the center of the classic LL diagram. If we consider for example two theoretical compositions having both $R_1 = 25$ (i.e. $Na^+ + K^+ = Ca^{2+} + Mg^{2+}$) and $R_2 = 25$ (i.e. $HCO_3^- = Cl^- + SO_4^{2-}$) but different concentrations, we will find that both of them will be plotted exactly in the same position in the middle of the classic LL diagram. However, in the compositional version of the diagram, these two compositions will not be plotted at the same location but the coordinates b_5 and b_2 of the two samples will differ according to the relative behavior among the considered variables. In summary, if $Na^+ + K^+ = Ca^{2+} + Mg^{2+}$ then R_1 will always be 25 due to the relative character of compositional data, differently for the same condition b_2 will have multiple values. The same applies for R_2 and b_5. The components of each subset are added together in the classic LL diagram, so that considering ions in pairs (e.g. $Na^+ + K^+$) implies that the information provided by the single components is lost. On the other hand, in the compositional version, the pairs of components are multiplied by one another, thus accounting for their proportional relations (Daunis-I-Estadella et al., 2006), better following the structure of the Law of Mass Action governing chemical reactions and weathering processes (Buccianti and Zuo, 2016). As a consequence, the point separating the different facies in the classical LL diagram can no longer be defined in the compositional diagram. Instead, new areas delineating the relative dominance of cations and anions could be delimited on the compositional diagram by determining whether the log-ratios are higher or lower than zero. It is worth noting that products and related geometric means can be highly affected by parts with small relative values, possibly causing misinterpretations (e.g. Rock, 1988; Gozzi and Buccianti, 2022). For example, imagine a composition with 4 parts with the values (0.001,2,5,7.5), then your geometric mean is 0.523. However, if you replace 0.001 with 0.00001, the geometric mean becomes 0.1654. This shows the sensitivity of the geometric mean regarding values on the border of the simplex, i.e. small contributions of a chemical element. However, this is rarely the case when working with the major components of natural waters.

4.2. Benefits of the new compositional diagram

The presented methods described in Section 2.4 are available through the function `LLdiagram` implemented in the R package `robCompositions` (Templ et al., 2011; Filzmoser et al., 2018) as of version 2.3.3. The proposed new visualization of the LL diagram allows the computation of the robust confidence ellipse around the barycenter for one or more data sets and to distinguish cases according to a set of criteria (e.g., geography, lithology, or other environmental drivers). In most cases, when analyzing the water chemistry, the choice of the grouping criterion for the visualization is made a priori before the LL diagram is constructed. With the option of quickly changing the visualization based on a criterion, it will be easier to select the most suitable representation for the data. This improvement can facilitate a better understanding of water-environment interactions and the detection of triggering factors for water composition. Research advancements rely on this knowledge to develop water system resilience to anthropogenic and climate-related impacts (Gozzi et al., 2021).

Rather than choosing an external parameter to display samples, the implementation allows you to define groups based on a clustering algorithm (k-means and model-based clustering) and plot the results on the LL diagram. Clustering methods can either be applied to b_2 and b_5 or to the isometric coordinate representation of all variables (b_1 to b_6). As a first option, we focus only on the visible information provided by the LL diagram. Differently, using the second option a more comprehensive picture is given incorporating the information of all log-ratio coordinates with additional benefits to the interpretation. Furthermore, the implementation allows interactive diagrams through the R package `plotly` (Sievert, 2020). The interactive version supports zooming and mouse overlays and allows to display additional information about observations/points in the diagram. When hovering over a marker, one can

get a tooltip showing the needed information which could be very useful for a quick visual exploration of the data.

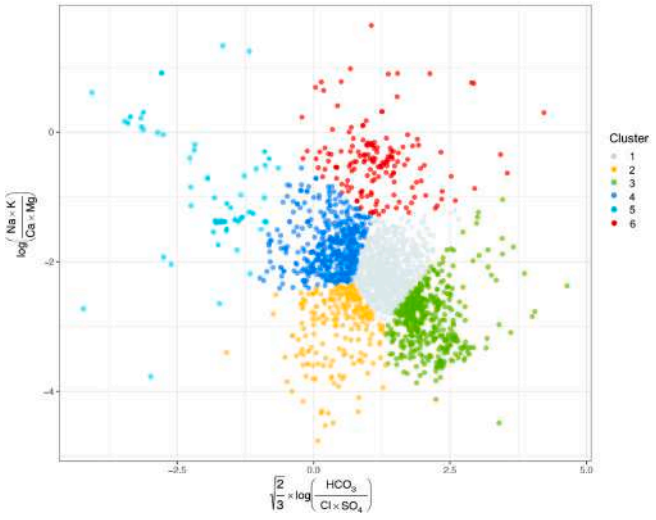
Another visualization tool (Fig. 5a) was presented in which points are colored using a bivariate gradient color space based on their position in the LL diagram and their proximity to the robust center of the data. This tool is especially useful for highly heterogeneous datasets and in combination with a map using the same color coding of the plot. However, one should be aware that the barycenter of the data changes based on the dataset and accordingly the bivariate color scale. Comparisons among different LL diagrams may be possible by plotting the barycenters of the datasets in the same graph, as illustrated by the example of the Arezzo and Tuscany ground waters (Fig. 5a and b). The benefit of this visualization is that it combines the information contained in the LL diagram with a measure of distance from the barycenter.

4.3. Geochemical interpretation based on the new diagram

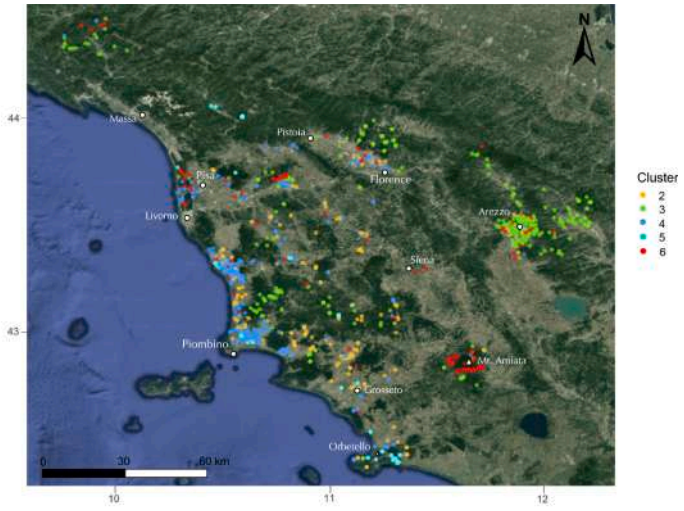
Considering the overall closeness to the barycenter (Fig. 5a) and the narrow variety of EC variations (Fig. 3), the results indicate that the composition of Arezzo ground waters is fairly homogeneous. The waters are mainly dominated by HCO_3^- and $Ca^{2+}(Mg^{2+})$ as a result of water-rock interaction processes with calcium/magnesium carbonate. However, values of b_2 close to zero suggest the presence in some samples of higher relative contents of $Na^+(K^+)$ with respect to the barycenter (see labels D7B, 62, 35a, etc). The origin of this relative increase could be related to the influence of CO_2-rich waters that favor cation exchange processes with clay minerals contained in sedimentary formations, as reported by Vaselli et al. (2005); Vaselli (2011). The presence of a deep CO_2-rich gas phase can be originated by thermo-metamorphic processes combined with mantle- and biogenic CO_2 inputs (e.g., Vaselli et al., 2005). The peculiar relative enrichment in HCO_3^- detected for site H6, may indicate that these are mineral-thermal waters circulating in deep aquifers with significantly higher salinity due to advanced water-rock interaction with carbonatic lithotypes. It is important to note that this difference is not evident from the classical LL diagram. Furthermore, values of b_5 close to zero suggest higher relative contents of $Cl^-(SO_4^{2-})$ with respect to the average (see labels 1, H7A, D71). This could be related to possible anthropogenic contributions of Cl^- and SO_4^{2-} (e.g., agricultural fertilizers, animal waste, and municipal and industrial sewage) since the presence of natural sources (i.e., evaporitic minerals) has not been ascertained in the study area (Buccianti et al., 2014).

Several environmental factors could explain the lack of clustering with defined boundaries in the data, including the geological heterogeneity of the Tuscany area, the nature of the hydrological complex that hosted the studied groundwaters (Fig. 1) and the occurrence of mixing processes. However, even if the groups are not clearly distinct from one another, applying k-means clustering (with k = 6) on **Z** seems to provide interesting results. In fact, the obtained clusters, illustrated in Fig. 7a, marked five water groups having different deviations from the barycenter, here approximated by the central cluster (n.1). Therefore, the external groups (i.e. 2–6) may identify waters characterized by different geochemical processes than those closer to the barycenter.

In order to verify the presence of potential spatial patterns, the results are represented in the map of Fig. 7b after filtering out samples belonging to cluster 1. Observing the map, samples belonging to cluster 2 appear widespread over the Tuscany area with higher occurrences in the northern area of Grosseto and near the coast north of Piombino. The relative enrichment in Ca(Mg)-SO_4 could derive from dissolution processes of either sulfate minerals in Triassic Formations or Miocene gypsum and/or anydrite layers (Cortecci et al., 2002) or, in rarer cases,



(a) K-means clustering on **Z** of Tuscany groundwater with six groups. Colors marked the cluster membership of the samples.



(b) Geographical representation of clusters 2-6. Colors mark the cluster membership of the samples.

Fig. 7. K-means clustering on **Z** and spatial distribution of the obtained clusters.

to mixing process of acid mine waters with alkaline waters (e.g. Colline Metallifere, Grosseto). The composition of cluster 3 tends towards the lower-left corner, showing a typical Ca(Mg)-HCO_3 composition (Fig. 7a). Generally, these waters are pristine groundwaters indicative of the first stages of water-rock interaction. In fact, based on the map, the majority of them is located in upland vegetated areas. Nevertheless, there are some exceptions represented by the outliers identified in Fig. 4. These last may refer to Ca-HCO_3 mineral-thermal waters circulating in deep aquifers with significantly higher salinity (up to 3000 μS/cm) than those flowing in shallow hydrological circuits (Minissale et al., 1997). Fig. 7b shows that samples of clusters 4 and 5, which tend to have a Na (K)-Cl dominance, are located mainly near the coast. This might indicate that they have been partially (n.4) or heavily (n.5) affected by seawater mixing, respectively. These results hereby confirm that several coastal aquifers in Tuscany are impacted by seawater intrusion (Nisi et al., 2016b; Franceschini and Signorini, 2016; Grassi and Netti, 2000). Some of the most severely affected samples by seawater mixing (EC above 10,000 μS/cm, see Fig. 4) are found along the Piombino and Orbetello headlands. Finally cluster 6 could indicate waters interacting with CO_2-rich gas discharge, such as those found in the area of Mt. Amiata (Minissale, 2004; Frondini et al., 2012). Nevertheless, since clusters are not clearly distinct from one another the interpretation of sample membership required cautions, especially for points located close to the borders of cluster 1 and near adjacent clusters.

A further geochemical improvement on the new LL diagram is related to the fact that its axes have a structure that remembers the Law of Mass Action and the proportionality that governs reactants (denominator) and products (numerator) in a chemical reaction:

$$\frac{(C)^c(D)^d}{(A)^a(B)^b} = k \tag{9}$$

with A, B, C and D chemical species and k equilibrium constant. The law states that the rate of any chemical reaction is proportional to the product of the masses of the reacting substances, with each mass raised to a power equal to the coefficient that occurs in the chemical equation. This law was formulated over the period 1864–79 by the Norwegian scientists Cato M. Guldberg and Peter Waage. Thus LL axes can represent the non-linearity that characterizes the entangled relationships among chemical species when participating in different reactions during water/rocks interaction processes. In this perspective, the inspection of the shape of the frequency distribution of the equations of the axes may be useful to understand if the relationships among the parts are homogeneously concentrated around a barycenter and if there are multimodalities, thus indicating the presence of more alternative stable states, presence of skewness, heavy tails and anomalous values. These features are very useful to understand how biogeochemical processes are working thus predicting their behavior in time and space as well as their resilience to change (Hirota et al., 2011; Scheffer et al., 2012, 2015).

For the data previously analyzed, Arezzo and Tuscany groundwaters, the density distributions of b_2 and b_5 balances are reported in Fig. 8. In both cases, they are sufficiently symmetrical with a limited number of anomalous values. The presence of significant alternative stable states (e.g. bimodality) is not clearly identified. Only a small bimodality can be noticed in b_2 for both datasets. This result is attributable to the dominant behavior of carbonates cycle (mainly negative values for b_2 and positive values for b_5) involving Ca^{2+}, HCO_3^- and, secondarily, Mg^{2+}, which clearly marks the nature of the waters in the investigated lithological context, generating Ca^{2+}-HCO_3^- geochemical facies. In addition, the result indicates that the system has reached a stationary "sink" or attractor state from which it cannot escape without important driving processes. This occurs independently from the local (Arezzo groundwater) or regional (Tuscany groundwater) scale, since the curves are more or less overlapped. In this perspective, balances, taking into account proportional (not-linear) relationships among variables, as occurs in water/rock interaction processes, could be able to detect important changes in groundwater geochemistry at different spatial and temporal scales. In 2019, the identification of tipping points in the water cycle has been listed as one of the top 23 Unsolved Problems in Hydrology (UPH) by the International Association of Hydrological Sciences (IAHS). A tipping point is the value of the critical threshold at which the future state of a system is altered by natural factors, e.g. climate change or anthropogenic pressure, thereby generating bi- or multi stability (Schellnhuber, 2009; Biggs et al., 2018). Hence, by using a simple diagram such as the LL revised from the compositional perspective, it is possible to obtain manageable indices to monitor the dynamic of groundwater exploring its stability/instability, fragmentation in groups, and presence of transient states as a first exploratory picture of a complex system. Furthermore, since the LL diagram is based on the interaction between two balances, the shape of the bivariate distribution b_2-b_5 can also be analyzed in order to get a more comprehensive understanding of the entangled relationships between chemical species in aqueous solutions. With this aim, the 2D kernel density estimation of the bivariate distributions for Arezzo and Tuscany groundwater are reported in Fig. 9.

In both cases, the bivariate plots confirmed the presence of a stationary attraction region. Only a small isolated fragment, separated from the main sink, is visible in Tuscany groundwater. This group represents the barycenter of cluster 6 which characterizes waters interacting with CO_2. However, in the upper portion of the density estimation of Arezzo groundwater, a deformation of the contour lines is also apparent, similarly related to the influence of CO_2-rich waters. In the event of

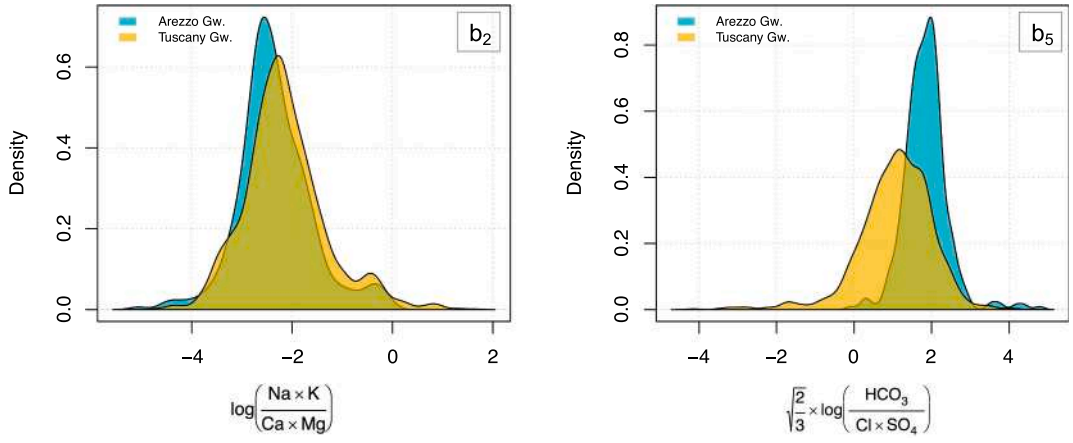


Fig. 8. Comparison of the density distributions of b_2 (left) and b_5 (right) log-ratios (i.e. axes of the compositional Langerlier-Ludwing diagram) for Arezzo and Tuscany groundwaters (Gw.).

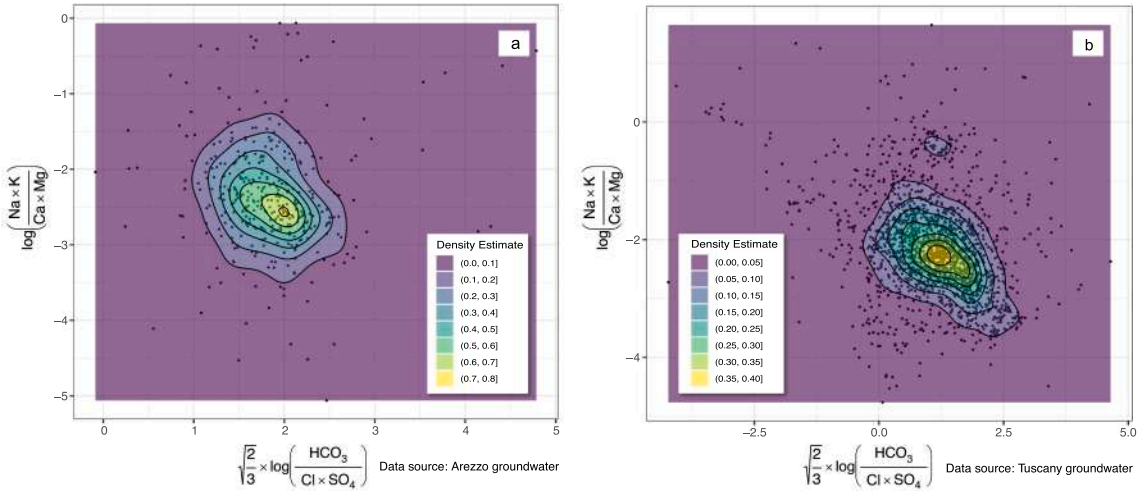


Fig. 9. 2D kernel density estimation of the bivariate distribution b_2-b_5 for a) Arezzo and b) Tuscany groundwater.

environmental modifications, this protrusion might turn into a completely separate state or could be adsorbed by the main sink. Based on the picture of the system provided by the analyzed data at the given time of monitoring, CO_2-rich gas discharge is likely to be one of the major factors potentially contributing to the bi-stabilization of the stationary behavior of Tuscany and Arezzo groundwaters. CO_2-rich gas discharge, in turn, is mainly controlled by the structural and geological setting of the area (Section 2.1). Nevertheless, other environmental factors such as topography, meteorological, hydrogeological and climatic conditions could also condition its discharge rate (e.g., Minissale, 2004).

Nevertheless, by analyzing the system on a broader time scale, the effects of other potential perturbing agents may be detected such as those related to anthropogenic activities or climate change. The first ones tend to increase the water salinity through agricultural practices, industrial activities and municipal and industrial sewage (Nisi et al., 2008). The second ones determine an accelerated hydrological cycle with more frequent floods and droughts (Gleick, 2010) and numerous other effects (e.g., increased evaporation, changes in precipitation intensity and duration and runoff magnitude and timing, variations in volume and distribution of groundwater recharge). These variations have strong consequences on the concentration of solutes, nutrients and pollutants in surface and ground waters (e.g. Rice and Bricker, 1995; Green et al., 2011), thereby affecting water quality and composition.

5. Conclusions

The Langelier-Ludwig diagram is one of the most commonly used diagrams in geochemistry to identify hydrogeochemical facies of water samples. Its interpretation, however, could be dangerous, since the increase of one element would automatically decrease the other, due to the "closed" proportionality of geochemical data. In this paper, a compositional version of this diagram is presented with axes defined by isometric coordinates. In this way, a real space is created where the identification of linear or nonlinear patterns follows the Euclidean geometry. The new diagram was further implemented with: i) the computation of robust tolerance ellipses, ii) clustering algorithms with the option to choose whether all input variables should be included in the calculation or only those of the Langelier-Ludwig diagram and iii) different plotting options for enhanced exploration and visualization of the results. The benefit of the revised diagram is that all information is contained in the log-ratios which describe the entangled relationship between the chemical species in aqueous solutions. Geochemical interpretation of the diagram is based on the principle of relative dominance of major ions and distance from robust data barycenter. This perspective enables to go beyond the concept of an absolute classification of water types towards a more efficient tool for monitoring the system stability and presence of transient states. This improvement can facilitate a better understanding of water-environment interactions and the detection of potential tipping points in the water cycle.

5.1. Software implementation

The presented methods are implemented in the R package `robCompositions` (Templ et al., 2011; Filzmoser et al., 2018) as of version 2.3.3, available through the function `LLdiagram`. It also allows interactive diagrams through `plotly` supporting zooming and mouse overlays showing additional information about observations/points in the diagram.

CRediT authorship contribution statement

Matthias Templ: Conceptualization, Methodology, Formal analysis, Software implementation, Writing - Original Draft, Writing - Review & Editing

Caterina Gozzi: Conceptualization, Methodology, Visualization, Writing - Original Draft, Writing - Review & Editing

Antonella Buccianti: Conceptualization, Resources, Investigation, Writing - Original Draft, Writing - Review & Editing

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

The University of Florence is acknowledged for the financial assistance to the present research through University funds (A.B). The International Association for Mathematical Geosciences (IAMG) is thanked for the support provided by the Computers & Geosciences Research Scholarship 2020 (C.G.).

Funding

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

References

Aitchison, J., 1982. The statistical analysis of compositional data (with discussion). J. R. Stat. Soc. Ser. B 44 (2), 139–177.

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd, London.

Bhat, N., Bhat, A., Nath, S., Singh, B., Guha, D., 2016. Assessment of drinking and irrigation water quality of surface water resources of South-West Kashmir, India. J. Civ. Environ. Eng. 6 https://doi.org/10.4172/2165-784X.1000222.

Biggs, R., Peterson, G.D., Rocha, J.C., 2018. The regime shifts database: a framework for analyzing regime shifts in social-ecological systems. Ecol. Soc. 23 https://doi.org/10.5751/ES-10264-230309.

Billheimer, D., Guttorp, P., Fagan, W., 2001. Statistical interpretation of species composition. J. Am. Stat. Assoc. 96, 1205–1214. https://doi.org/10.1198/016214501753381850.

Buccianti, A., 2015. The FOREGS repository: modelling variability in stream water on a continental scale revising classical diagrams from CoDA (compositional data analysis) perspective. J. Geochem. Explor. 154, 94–104. https://doi.org/10.1016/j.gexplo.2014.12.003.

Buccianti, A., Magli, R., 2011. Metric concepts and implications in describing compositional changes for world river's water chemistry. Comput. Geosci. 37, 670–676. https://doi.org/10.1016/j.cageo.2010.04.017.

Buccianti, A., Zuo, R., 2016. Weathering reactions and isometric log-ratio coordinates: do they speak to each other? Appl. Geochem. 75, 189–199. https://doi.org/10.1016/j.apgeochem.2016.08.007.

Buccianti, A., Nisi, B., Martín-Fernández, J., Palarea-Albaladejo, J., 2014. Methods to investigate the geochemistry of groundwaters with values for nitrogen compounds below the detection limit. J. Geochem. Explor. 141, 78–88. https://doi.org/10.1016/j.gexplo.2014.01.014.

Cangemi, M., Madonia, P., Albano, L., Bonfardeci, A., Di Figlia, M.G., Di Martino, R.M.R., Nicolosi, M., Favara, R., 2019. Heavy metal concentrations in the groundwater of the Barcellona-Milazzo Plain (Italy): contributions from geogenic and anthropogenic sources. Int. J. Environ. Res. Public Health 16. https://doi.org/10.3390/ijerph16020285.

Chayes, F., 1960. On correlation between variables of constant sum. J. Geophys. Res. 65, 4185–4193. https://doi.org/10.1029/JZ065i012p04185.

Chen, J., Hron, K., Templ, M., Li, S., 2018. Regression imputation with q-mode clustering for rounded zero replacement in high-dimensional compositional data. J. Appl. Stat. 45 (11), 2067–2080.

Collins, W., 1923. Graphic representation of water analyses. J. Ind. Eng. Chem. 15, 394. https://doi.org/10.1021/ie50160a030.

Cortecci, G., Dinelli, E., Bencini, A., Adorni-Braccesi, A., La Ruffa, G., 2002. Natural and anthropogenic SO4 sources in the Arno river catchment, northern Tuscany, Italy: a chemical and isotopic reconnaissance. Appl. Geochem. 17 (2), 79–92. https://doi.org/10.1016/S0883-2927(01)00100-7.

Daunis-I-Estadella, J., Barceló-Vidal, C., Buccianti, A., 2006. Exploratory compositional data analysis. Geol. Soc. Lond., Spec. Publ. 264, 161–174. https://doi.org/10.1144/GSL.SP.2006.264.01.12.

Durov, S., 1948. Natural waters and graphic representation of their composition. Dokl. Akad. Nauk SSSR 59, 87–90.

Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geosci. 37, 795–828. https://doi.org/10.1007/s11004-005-7381-9.

Engle, M.A., Rowan, E.L., 2014. Geochemical evolution of produced waters from hydraulic fracturing of the Marcellus Shale, northern Appalachian Basin: a multivariate compositional data analysis approach. Int. J. Coal Geol. 126, 45–56. https://doi.org/10.1016/j.coal.2013.11.010.

Filzmoser, P., Hron, K., 2008a. Correlation analysis for compositional data. Math. Geosci. 41, 905. https://doi.org/10.1007/s11004-008-9196-y.

Filzmoser, P., Hron, K., 2008b. Outlier detection for compositional data using robust methods. Math. Geosci. 40, 233–248. https://doi.org/10.1007/s11004-007-9141-5.

Filzmoser, P., Hron, K., Templ, M., 2018. Applied Compositional Data Analysis: With Worked Examples in R. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-96422-5, 280pp.

Fraley, C., Raftery, A., 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput. J. 41, 578–588. https://doi.org/10.1093/comjnl/41.8.578.

Franceschini, F., Signorini, R., 2016. Seawater intrusion via surface water vs. deep shoreline salt-wedge: a case history from the Pisa coastal plain (Italy). Groundw. Sustain. Dev. 2–3, 73–84. https://doi.org/10.1016/j.gsd.2016.05.003.

Frondini, F., Cardellini, C., Caliro, S., Chiodini, G.G., Morgantini, N., 2012. Regional groundwater flow and interactions with deep fluids in western Apennine: the case of Narni-Amelia chain (Central italy). Geofluids 12, 182–196. https://doi.org/10.1111/j.1468-8123.2011.00356.x.

Gibbs, R., 1970. Mechanisms controlling world water chemistry. Sciences 170, 1088–1090. https://doi.org/10.1126/science.170.3962.1088.

Gleick, P., 2010. The World's Water 2008–2009. The Biennial Report on Freshwater Resources. Island press, Washington DC, 400pp.

Gozzi, C., Buccianti, A., 2022. Assessing indices tracking changes in River geochemistry: implications for monitoring. Nat. Resour. Res. 31, 1061–1079. https://doi.org/10.1007/s11053-022-10014-1.

Gozzi, C., Dakos, V., Buccianti, A., Vaselli, O., 2021. Are geochemical regime shifts identifiable in river waters? Exploring the compositional dynamics of the Tiber River (Italy). Sci. Total Environ. 785, 147268 https://doi.org/10.1016/j.scitotenv.2021.147268.

Grassi, S., Netti, R., 2000. Sea water intrusion and mercury pollution of some coastal aquifers in the province of Grosseto (southern tuscany — italy). J. Hydrol. 237, 198–211. https://doi.org/10.1016/S0022-1694(00)00307-3.

Green, T.R., Taniguchi, M., Kooi, H., Gurdak, J.J., Allen, D.M., Hiscock, K.M., Treidel, H., Aureli, A., 2011. Beneath the surface of global change: Impacts of climate change on groundwater. J. Hydrol. 405, 532–560. https://doi.org/10.1016/j.jhydrol.2011.05.002.

Hill, R., 1940. Geochemical patterns in the Coachella Valley, California. Trans. Am. Geophys. Union 21 (21), 46–49. https://doi.org/10.1029/TR021i001p00046.

Hirota, M., Holmgren, M., Van Nes, E.H., Scheffer, M., 2011. Global resilience of tropical forest and savanna to critical transitions. Science 334, 232–235. https://doi.org/10.1126/science.1210657.

ISPRA Ambiente, 2017. Geoportale Ispra Ambiente. URL: http://geoportale.isprambiente.it/sfoglia-il-catalogo/?lang=en. (Accessed 24 September 2019).

Langelier, W., Ludwig, H., 1942. Graphical methods for indicating the mineral character of natural waters. Am. Water Works Assoc. J. 34, 335–352. https://doi.org/10.1002/j.1551-8833.1942.tb19682.x.

Lavecchia, G., 1990. The Tyrrehenian-Apennine system: structural setting and seismotectogenesis. Tectonophysics 47, 263–296. https://doi.org/10.1016/0040-1951(88)90190-4.

Maronna, R., Martin, R., Yohai, V., 2006. Robust Statistics: Theory and Methods. John Wiley & Sons, New York. ISBN 978-0-470-01092-1.

Martin-Fernandez, J., Hron, K., Filzmoser, P., Palarea-Albaladejo, J., 2015. Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat. Model. 37 (7), 134–158.

Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J., 2011. The principle fo working in coordinates. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd, Chichester, pp. 34–42. https://doi.org/10.1002/9781119976462.ch3.

Merino, L., Aguilera, H., Gonzales-Jimenez, M., Diaz-Losada, E., 1944. D-Piper, a modified piper diagram to represent big sets of hydrochemical analyses. Environ. Model. Softw. 138, 104979 https://doi.org/10.1016/j.envsoft.2021.104979.

Minissale, A., 2004. Origin, transport and discharge of in Central Italy. Earth Sci. Rev. 66, 89–141. https://doi.org/10.1016/j.earscirev.2003.09.001.

Minissale, A., Evans, W., Magro, G., Vaselli, O., 1997. Multiple source components in gas manifestations from north-Central Italy. Chem. Geol. 142, 175–192. https://doi.org/10.1016/S0009-2541(97)00081-8.

Minissale, A., Magro, G., Martinelli, G., Vaselli, O., Tassi, F., 2000. Fluid geochemical transect in the northern apennines (central-Northern italy): fluid genesis and migration and tectonic implications. Tectonophysics 319, 199–222. https://doi.org/10.1016/S0040-1951(00)00031-7.

Nisi, B., Buccianti, A., Vaselli, O., Perini, G., Tassi, F., Minissale, A., Montegrossi, G., 2008. Hydrogeochemistry and strontium isotopes in the Arno River Basin (Tuscany, Italy): constraints on natural controls by statistical modeling. J. Hydrol. 360 (1–4), 166–183.

Nisi, B., Buccianti, A., Raco, B., Battaglini, R., 2016a. Analysis of complex regional databases and their support in the identification of background/baseline compositional facies in groundwater investigation: developments and application examples. J. Geochem. Explor. 164, 3–17.

Nisi, B., Buccianti, A., Raco, B., Battaglini, R., 2016b. Analysis of complex regional databases and their support in the identification of background/baseline compositional facies in groundwater investigation: developments and application examples. J. Geochem. Explor. 164, 3–17. https://doi.org/10.1016/j.gexplo.2015.06.019.

Pawlowsky-Glahn, V., Egozcue, J., 2001. Geometric approach to statistical analysis on the simplex. Stoch. Env. Res. Risk A. 15, 384–398. https://doi.org/10.1007/s004770100077.

Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, J., 2007. Lecture notes on compositional data analysis. available online. http://www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDa.pdf. (Accessed 31 May 2022).

Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. Statistics in Practice. John Wiley & Sons Ltd, 272 pp.

Pearson, K., 1897. Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc. R. Soc. Lond. 60, 489–502. https://doi.org/10.1098/rspl.1896.0076.

Piper, A., 1944. A graphic procedure in the geochemical interpretation of water-analyses. EOS Trans. Am. Geophys. Union 25, 914–928. https://doi.org/10.1029/TR025i006p00914.

R Development Core Team, 2022. R: a language and environment for statistical computing. Version 4.2.0. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Railsback, L., 2003. An earth scientist's periodic table of the elements and their ions. Geology 31, 737–740. https://doi.org/10.1093/oso/9780190668532.001.0001.

Rice, K., Bricker, O., 1995. Seasonal cycles of dissolved constituents in streamwater in two forested catchments in the mid-Atlantic region of the eastern Usa. J. Hydrol. 170, 137–158.

Rock, N.M.S., 1988. Summary statistics in geochemistry: a study of the performance of robust estimates. Math. Geol. 20, 243–275. https://doi.org/10.1007/BF00890256.

Safari, M., Hezarkhani, A., Mashhadi, S.R., 2020. Hydrogeochemical characteristics and water quality of Aji-Chay river, eastern catchment of Lake Urmia, Iran. J. Earth Syst. Sci. 129, 199. https://doi.org/10.1007/s12040-020-01469-y.

Scheffer, M., Carpenter, S., Lenton, T., Bascompte, J., Brock, W., Dakos, V., van de Koppel, J., van de Leemput, I., Levin, S., van Nes, E., Pascual, M., Vandermeer, J., 2012. Anticipating critical transitions. Science 338, 344–348. https://doi.org/10.1126/science.1225244.

Scheffer, M., Carpenter, S.R., Dakos, V., van Nes, E.H., 2015. Generic indicators of ecological resilience: Inferring the chance of a critical transition. Annu. Rev. Ecol. Evol. Syst. 46, 145–167. https://doi.org/10.1146/annurev-ecolsys-112414-054242.

Schellnhuber, H.J., 2009. Tipping elements in the earth system. Proc. Natl. Acad. Sci. 106, 20561–20563. https://doi.org/10.1073/pnas.0911106106.

Shelton, J., Engle, M., Buccianti, A., Blonde, M., 2018. The isometric log-ratio (ilr)-ion plot: a proposed alternative to Piper diagram. J. Geochem. Explor. 190 (190), 130–141. https://doi.org/10.1016/j.gexplo.2018.03.003.

Sievert, C., 2020. Interactive web-based data visualization with R, `plotly`, and shiny. Chapman and Hall/CRC. https://plotly-r.com.

Stiff, H., 1970. The interpretation of chemical water analysis by means of patterns. J. Petrol. 3 (10), 15–17. https://doi.org/10.2118/951376-G, 3.

Templ, M., Templ, B., 2020. Analysis of chemical compounds in beverages- guidance for establishing a compositional analysis. Food Chem. 325, 1–7. https://doi.org/10.1016/j.foodchem.2020.126755.

Templ, M., Templ, B., 2021. Statistical analysis of chemical element compositions in food science: problems and possibilities. Molecules 26, 5752. https://doi.org/10.3390/molecules26195752.

Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data: problems and possibilities. Appl. Geochem. 23, 2198–2213. https://doi.org/10.1016/j.apgeochem.2008.03.004.

Templ, M., Hron, K., Filzmoser, P., 2011. In: `robCompositions`: An R-package for Robust Statistical Analysis of Compositional Data. John Wiley & Sons, Ltd, pp. 341–355. https://doi.org/10.1002/9781119976462.ch25 chapter 25.

Templ, M., Hron, K., Filzmoser, P., Gardio, A., 2016. Imputation of rounded zeros for high-dimensional compositional data. Chemom. Intell. Lab. Syst. 155, 183–190.

Vaselli, O., 2011. Atlante geochimico delle acque sotterranee e di scorrimento superficiale del Comune di Arezzo: gli isotopi stabili nelle acque e nei gas ed i metalli pesanti come traccianti di inquinamento naturale ed antropico. Comune di Arezzo.

Vaselli, O., Buccianti, A., Romizi, A., Nisi, B., Cantucci, B., Tassi, F., Minissale, A., Montegrossi, G., 2005. A geochemical atlas of the ground-and running waters of Arezzo (Tuscany, Italy). In: GEOITALIA 2005.