Science of the Total Environment 946 (2024) 174099

Contents lists available at ScienceDirect

Science of the Total Environment

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

The pH paradox

Kagiso Samuel More [a,b], Christian Wolkersdorfer [a,*]

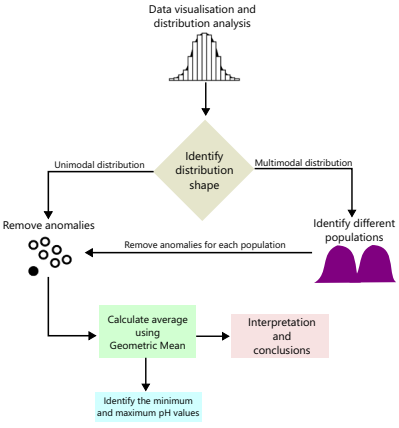
[a] SARChI Chair for Mine Water Management, Department of Environmental, Water and Earth Sciences, Tshwane University of Technology, Private Bag X680, Pretoria 0001, South Africa

[b] Institute for Nanotechnology and Water Sustainability (iNanoWS), College of Science, Engineering and Technology, University of South Africa, Private Bag X6, Science Campus, Florida, Johannesburg 1709, South Africa

HIGHLIGHTS

- Statistical rigor guides us on whether to use pH or $\{H^+\}$ for accurate calculations.
- pH and $\{H^+\}$ exhibit diverse distributions, influencing environmental data analysis.
- Geometric mean of pH in environmental data is suitable for accurate analyses.
- Reliable pH calculations require standardised protocols to ensure reproducibility.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: JV Cruz

Keywords:
Environmental parameters
Statistical analyses
Data distribution
Multimodal distribution
Geometric mean
k-Means clustering algorithm

ABSTRACT

This paper highlights the critical role of pH or proton activity measurements in environmental studies and emphasises the importance of applying proper statistical approaches when handling pH data. This allows for more informed decisions to effectively manage environmental data such as from mining influenced water. Both the pH and $\{H^+\}$ of the same system display different distributions, with pH mostly displaying a normal or bimodal distribution and $\{H^+\}$ showing a lognormal distribution. It is therefore a challenge of whether to use pH or $\{H^+\}$ to compute the mean or measures of central tendency for further environmental statistical analyses. In this study, different statistical techniques were applied to understand the distribution of pH and $\{H^+\}$ from four different mine sites, Metsämonttu in Finland, Felsendome Rabenstein in Germany, Eastrand and Westrand mine water treatment plants in South Africa. Based on the statistical results, the geometric mean can be used to calculate the average of pH if the distribution is unimodal. For a multimodal pH data distribution, peak identifying methods can be applied to extract the mean for each data population and use them for further statistical analyses.

* Corresponding author.
E-mail addresses: moreks@unisa.ac.za (K.S. More), christian@wolkersdorfer.info (C. Wolkersdorfer).

https://doi.org/10.1016/j.scitotenv.2024.174099
Received 13 December 2023; Received in revised form 27 May 2024; Accepted 16 June 2024
Available online 23 June 2024
0048-9697/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Measurement of pH or proton activity is of critical importance in the natural environment due to its role in understanding a wide range of chemical, physicochemical and biological processes. The measurement of $\{H^+\}$ or its negative decimal logarithm (pH) gives information about the acid within a given system (Boyd et al., 2011; Kuna-Broniowska and Smal, 2017; Patnaik, 2017). This parameter is extensively measured by researchers or samplers in various fields such as water science, hydrology, soil science, physiology and environmental science, to assess and monitor the chemical characteristics of water bodies, soil and other environmental components (e.g. Kissel et al., 2009; Nordstrom, 2011).

Accurate measurement of pH requires standardised protocols and precise instrumentation to ensure reliable and reproducible results. pH meters equipped with suitable electrodes are often used to measure the activity of H^+ ions. These measurements are performed across a wide range of pH values, from highly acid environments with low pH values (e.g. acid mine drainage; Schmiermund and Drozd, 1997) to alkaline environments with high pH values. The logarithmic nature of the pH allows for a convenient representation of these extreme $\{H^+\}$ variations. In mine water management, pH plays a crucial role as it directly influences the solubility, mobility and bioavailability of various metals and minerals, controlling their potential to pollute mine water and surrounding water bodies (Cravotta III, 2008; Stumm and Morgan, 1996). In addition, pH measurements help assess the effectiveness of remediation techniques and guide the implementation of appropriate treatment strategies to mitigate the potential effects of acid or alkaline mine water on the surrounding environment (Wolkersdorfer, 2022).

When performing statistical analyses, a precise and accurate value of pH or $\{H^+\}$ is required. This value is typically a single mean pH value of an individual population which provides essential information about the acid in the system. This average is therefore used to make critical statistical conclusions on the soil type, water or environmental components. It has often been mentioned that it is not possible to average the pH because "logarithms cannot be averaged until they have been converted back to real numbers" (Giesecke et al., 1978). This assumption does not hold, as we will show.

The mean is calculated alongside other measures of variability and central tendency such as standard deviation or median, to capture the full complexity of the data set. $\{H^+\}$ and pH use different measurement scales, meaning their statistical distributions are always different (Kuna-Broniowska and Smal, 2017). The $\{H^+\}$ quantifies the concentration of hydrogen ions in a solution: the higher the concentration of hydrogen ions, the more acid the solution. Its scale is expressed in terms of molarity (mol/L) or activity (a thermodynamic measure of ion concentration). The pH is operationally defined (Covington et al., 1985) as the negative logarithm of the hydrogen ion activity ($pH = -\log_{10}\{H^+\}$), has no unit (Cohen et al., 2008) and was introduced by Sørensen (1909), though his concept was widely criticised, as smaller numbers mean higher activities and a difference of 1 means a 10 times higher or lower H^+ activity. If the concept of Tillmans (1919), who recommended to divide the H^+ concentration by 10^{-7} and naming this number $h^\bullet$, would be in use today, this paper would not have been written, because the $h^\bullet$ for a liquid with pH 7 would have been $h^\bullet = 1$ and $h^\bullet$ would show the same distribution as all other environmental data. Yet, using the pH, the lowest ever measured pH in mining influenced water (MIW) is −3.6 (Nordstrom et al., 2000) and highest being 12.8 (Roadcap et al., 2005) – this would give us $h^\bullet_{-3.6} = 3.981 \times 10^{10}$ and $h^\bullet_{12.8} = 1.585 \times 10^{-6}$. Therefore, with these differences in pH and $\{H^+\}$, the question is whether to use pH or $\{H^+\}$ values to compute measures of variability and central tendency. This study aims to examine the statistical differences between the distribution of pH and $\{H^+\}$ values and highlights key points in statistically handling multimodal pH distributions.

When dealing with unimodal distributions, i.e. a statistical distribution that has a single peak, calculating central measures of tendency and variability, such as the mean and standard deviation, is relatively easy (McCluskey and Lalkhen, 2007). These measures provide essential insights into the central location and spread of the data. The pH data often gives multimodal distributions which is characterised by multiple peaks and has not only been shown by the data of this paper, but in the Pennsylvanian anthracite and bituminous coalfields (Cravotta III et al., 1999), and also for a regional dataset reported by Kirby and Cravotta III (2005). These different peaks represent different populations within the data set that has varying characteristics (Selvamuthu and Das, 2018). It is important to note that these populations may have different underlying causes, behaviours or properties, which becomes a challenge as to how to calculate the measures of central tendency and variability. The term "population" in this context refers to a subset of data points within a multimodal distribution that share similar characteristics. In a multimodal distribution, calculating parameters like the mean and standard deviation for the entire data set as if it were unimodal would not accurately capture the underlying patterns and may lead to misleading conclusions. By applying techniques such as the *k*-means clustering algorithm, which is a data clustering method that groups data points into clusters based on their similarity, one can effectively separate the different populations within the multimodal distribution (Ikotun et al., 2023; Lloyd, 1982; MacQueen, 1967). This separation allows for the calculation of central measures of tendency (e.g. mean) and measures of variability (e.g. standard deviations) that are specific to each population. Additionally, this separation may also be possible using probability plots whereby different subpopulations will result in break-in slopes (e.g. Rose et al., 1979).

When data are knowingly taken from a single population, but the number of individual samples is too small to plot or calculate meaningful histogram data, it may still be possible to calculate a mean. This follows from the Central Limit Theorem, as outlined by Davis (2002), who shows that meaningful statistics can be calculated even when a population does not have a normal distribution. If the number of measurements for a population increases, the means of individual measurement campaigns tend to follow a normal distribution and so does the data (Andersson, 2021).

Yet, it is important to note that even if the pH data displays a unimodal distribution, calculating the pH average is not a straightforward process, which has been discussed by various researchers for a long time already (e.g. Boutilier and Shelton, 1980). Thus, this paper explored the four different possibilities of calculating the average, i.e. arithmetic mean, median, geometric mean and harmonic mean. Some of the aforementioned average calculating methods are suitable when the data are normally distributed, and others give accurate and precise results even when the data are not normally distributed and contains anomalies.

It is important to emphasise that the pH resulting from the mixing of two solutions with distinct pH values, pH 1 and pH 2, is not simply the arithmetic average of the two pH values. When solutions are mixed, chemical reactions can occur, leading to changes in the concentrations of ions present in the solution. These reactions can influence the resulting pH, which may deviate from a simple average of the initial pH values. Factors such as the nature of the substances being mixed, their concentrations and the presence of buffering agents can all influence the final pH of the mixture. Therefore, determining the pH of a mixture requires consideration of these chemical reactions and their effects on proton concentration (Nordstrom, 2020).

This study focuses solely on estimating the average pH from one population of MIW samples, rather than addressing the resulting pH from mixing of two or more samples, as discussed in the beforementioned paper by Nordstrom (2020). While the calculation of pH for mixtures is indeed relevant in environmental contexts, this paper aims to explore statistical methods for estimating central tendency measures for pH data of one population. By analysing the distribution of pH values in these populations and identifying appropriate statistical approaches, this study aims to provide insights into effectively summarising and interpreting pH data in environmental studies. This distinction

highlights the importance of contextualising pH computations within the scope of the research objectives and dataset characteristics.

2. Environmental data distribution with a bird's eye view on pH

Environmental data such as major ions or trace elements, TDS, EC and pH, to name a few, are measurements which are commonly used to characterise the physical and chemical properties of natural systems. These data can help in understanding the state of water quality and other environmental conditions. When performing statistical analyses, researchers often assume that these types of environmental data display a normal distribution where data points cluster around the mean value, with fewer points deviating from the mean as they move away from it. This is often done by computing statistical tests which are based on the assumption that the data follows a normal or Gaussian distribution (e.g. correlation, regression, *t*-tests; Patel and Read, 1996; Thomopoulos, 2017). However, this assumption is wrong because environmental data too often tend to have a lognormal distribution (Gaddum, 1945; Reimann and Filzmoser, 2000).

While pH data might visually resemble a normal or multimodal distribution, the distribution of the underlying proton activity is not normal, it is more accurately described by a lognormal distribution (Fig. 1). This log-normal distribution of environmental data is characteristic for many scientific fields, though the reasons are only partly understood (Andersson, 2021), and require careful consideration to avoid biased or even faulty interpretations (Reimann and Filzmoser, 2000). Unlike most environmental data, pH often has a normal-like distribution, hence the title of this paper, which statistically explains the paradox arising from the common assumption that pH data resembles the properties of any typical environmental parameter data. This assumption ignores that pH is derived from $\{H^+\}$ by a log transformation and when the $\{H^+\}$, like most environmental data, is lognormal like distributed, the pH distribution becomes normal like. Therefore, the paradox comes from a misunderstanding of the relationship between pH, $\{H^+\}$ and environmental data in general not normally distributed (Blythe and Merhaut, 2010; Giesecke, 1979; Grogono, 1980; Kuna-Broniowska and Smal, 2017; Pace et al., 1979).

Understanding this paradox is important for proper statistical analysis and interpretation of environmental data, which may result in crucial conclusions such as the type of treatment technique to use in the case of analysing water quality at the treatment plant. Boyd et al. (2011) gives an example where using the wrong pH average could result in fish mortality for warmwater fish that are intolerant to pH values below 4. Their example is also of relevance for MIW treatment plants, where using wrong pH averages could cause fish death. The example also shows the relevance of reporting not only the central tendencies, but the range of the measured pH values.

When analysing pH data, it is important to account for the logarithmic transformation and the characteristic lognormal distribution of $\{H^+\}$. Failing to recognise this distinction could lead to incorrect assumptions about the data's distribution and result in flawed analyses, as has already been noticed by Gaddum (1945).

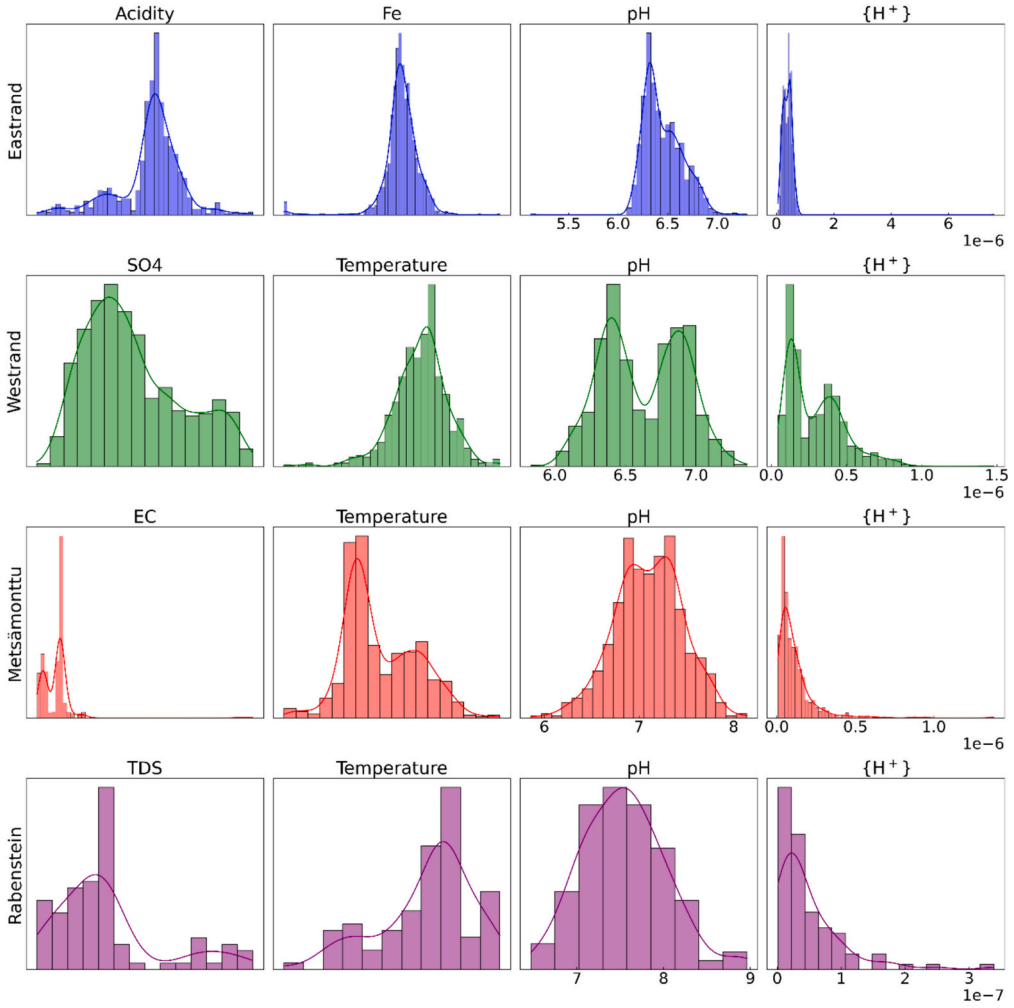


Fig. 1. Histograms for different environmental data sets for the Eastrand, Westrand, Metsämonttu and Felsendome Rabenstein abandoned mine sites; EC: electrical conductivity, TDS: total dissolved solids.

3. pH data distribution

3.1. Descriptive statistics

This study used data from four abandoned mines, i.e. the Metsämonttu mine site in Finland (Wolkersdorfer, 2017) with data ranging from 2016 to 2020, the Felsendome Rabenstein limestone mine in Germany (Wolkersdorfer et al., 2013) with historical data from 2002 to 2004, as well as the Westrand and Eastrand mine sites in South Africa (More and Wolkersdorfer, 2023a; More and Wolkersdorfer, 2023b) with data measured from 2016 to 2021 for both sites (Table 1). In all cases, as expected, the minimum value for pH is the maximum pH value calculated from the $\{H^+\}$ and vice versa. This means that the distribution for both pH and $\{H^+\}$ will always be different. Furthermore, the mean pH derived from $\{H^+\}$ is always different from the calculated pH. For example, from the descriptive statistics results obtained, using Sørensen's equation to calculate pH ($-\log_{10}\{H^+\}$), the values are 6.95, 6.54, 6.41 and 7.31, which are different from the calculated pH averages (7.10, 6.63, 6.45 and 7.54). Therefore, this shows that there is a need for careful consideration in using the values of central tendency for pH for further statistical calculations.

3.2. Visual analysis

The distribution of pH data in the abandoned Felsendome Rabenstein limestone mine shows a normal distribution which indicates that the data points are centered around the mean with a symmetric spread, while its $\{H^+\}$ data shows a lognormal distribution with a right skew. This means most of the $\{H^+\}$ data is concentrated towards the left side of the distribution, while a few values extend towards the right. A bimodal distribution can be seen on the Eastrand, Westrand and Metsämonttu pH data, showing two distinct peaks. This implies that there may be two different environmental conditions contributing to the observations. Additionally, the associated $\{H^+\}$ shows lognormal distributions with right skew (Fig. 2).

Graphical visual representations provide a more comprehensive understanding of the distributions. Comparing the pH and $\{H^+\}$ data sets, it can be seen that the distributions are different, which makes it challenging to select a data set to use for computing measures of central tendency. In most cases, the average of pH is used because of the assumption that environmental data ($\{H^+\}$ in this context) have a normal distribution (Howe and Howe, 1981; Pace et al., 1979; Verma, 2020). However, this is not always the case as can be seen in this study where $\{H^+\}$ data, as for most other environmental data, displays a lognormal distribution. Thus, it is crucial to know if the pH or $\{H^+\}$ data set should be used and how to correctly determine the mean of the data.

Table 1
pH and $\{H^+\}$ summary statistics for the Metsämonttu, Westrand, Eastrand and Felsendome Rabenstein abandoned mines; n: number of measurements, $\bar{x}$: arithmetic mean, σ: standard deviation.

	Metsämonttu		Westrand		Eastrand		Felsendome Rabenstein	
	pH	$\{H^+\}$ (pH)	pH	$\{H^+\}$ (pH)	pH	$\{H^+\}$ (pH)	pH	$\{H^+\}$ (pH)
n	642		1122		1380		96	
$\bar{x}$	7.10	1.12×10^{-7} (6.95)	6.63	2.91×10^{-7} (6.54)	6.45	3.87×10^{-7} (6.41)	7.54	4.86×10^{-8} (7.31)
σ	0.36	1.19×10^{-7}	0.29	1.86×10^{-7}	0.19	2.44×10^{-7}	0.48	5.49×10^{-8}
Min.	5.86	7.24×10^{-9}	5.83	4.37×10^{-8}	5.12	5.01×10^{-8}	6.47	1.10×10^{-9}
25 %	6.86	4.47×10^{-8}	6.39	1.32×10^{-7}	6.31	2.63×10^{-7}	7.20	1.35×10^{-8}
50 %	7.12	7.68×10^{-8}	6.59	2.57×10^{-7}	6.41	3.89×10^{-7}	7.53	2.99×10^{-8}
75 %	7.35	1.38×10^{-7}	6.88	4.07×10^{-7}	6.58	4.90×10^{-7}	7.87	6.28×10^{-8}
Max.	8.14	1.38×10^{-6}	7.36	1.48×10^{-6}	7.30	7.59×10^{-6}	8.96	3.39×10^{-7}

3.3. Anomaly detection

From the descriptive statistics and visual analysis methods conducted (Table 1; Fig. 2), it can be seen that the pH data contains anomalies. Therefore, to continue performing further statistical analysis such as accurately determining central tendency calculations, anomalies in the data sets must be removed. There are several advanced statistical analysis methods to detect anomalies in environmental data (e.g. More and Wolkersdorfer, 2023b); however, this paper only explored the isolation forest technique. The isolation forest is an unsupervised learning algorithm designed to efficiently isolate and identify anomalies within data sets. It is particularly effective in situations where anomalies are rare and distinct from normal data points, such as in the environmental data sets. Isolation forest detects anomalies by creating an ensemble of decision trees that collectively work to isolate anomalies more rapidly than normal data points (Xu et al., 2017).

The process of isolation forest algorithm works by randomly selecting a feature and a random value within that feature's range. This initial selection serves as the basis for splitting data points into two subsets. This random subsampling and splitting continues iteratively, with each subset undergoing recursive splitting using randomly chosen features and values. Therefore, this recursive division gives rise to a tree-like structure, where data points ultimately reach terminal nodes. The depth of the tree where a data point reaches a terminal node is recorded as its path length. Shorter path lengths indicate that a data point required fewer splits to be isolated. Anomalies are expected to have shorter path lengths on average than normal data points because they are often positioned far from the majority of data points, allowing them to be separated with fewer splits. Hence, a decision threshold is determined based on the average path length values. Data points with shorter average path lengths are then concluded to be anomalies (Lesouple et al., 2021; Xu et al., 2017).

For this study, the models for all the pH data sets were configured using the contamination parameter (proportion of anomalies in the data set) of 0.05. It should be noted that the contamination parameter can be adjusted to suit the data set. Finally, the isolation forest models detected 4.93 % (68 anomalies), 4.81 % (54 anomalies), 5.14 % (33 anomalies) and 5.21 % (5 anomalies) anomalies in the Eastrand (n: 1380), Westrand (n: 1122), Metsämonttu (n: 642) and Felsendome Rabenstein (n: 96) pH data sets, respectively, and they were removed which resulted in clean data (Fig. 3). This study only focused on examining the pH data, thus the unsupervised anomaly detection algorithm (isolation forest) only used a single parameter (pH) data set to build a model. The statistically precise and accurate way to detect anomalies is to use a multivariate method which takes into account the relationship between different parameters (e.g. More and Wolkersdorfer, 2023b).

3.4. pH averaging

Averaging of environmental data is often done using the arithmetic mean, which from a strict statistical point of view is the wrong way to do it because environmental data is usually not normally distributed. Arithmetic mean is the sum of all the values in a data set divided by the number of values. This approach is a good measure of central tendency for normal distributions, but it can be misleading for skewed, bimodal or multimodal distributions. For example, the pH data distribution for the Westrand mine water pool (Fig. 2) shows two clear peaks, meaning there are two distinct populations, each with its own mean. Therefore, the arithmetic mean will be somewhere between the two means, but it will not be representative of either population as it will be pulled towards the larger population. Thus it is crucial to explore different ways of calculating the mean and examining the accurate one. This study compared

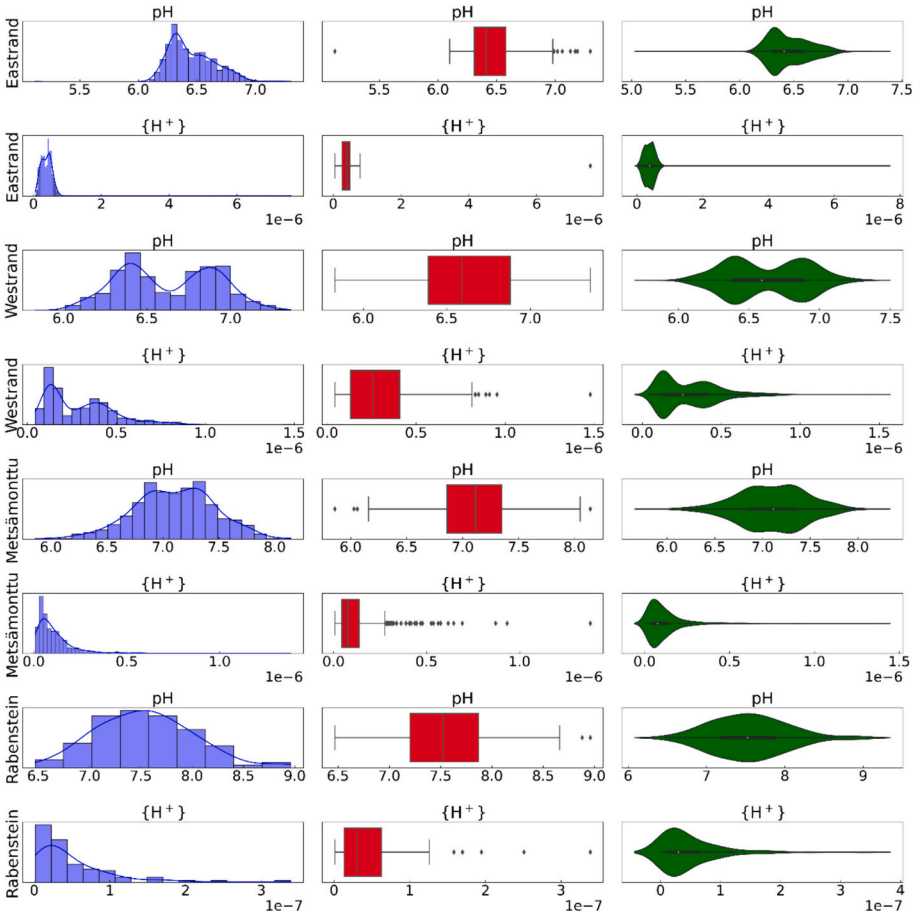


Fig. 2. Distribution plots (histograms with kernel density estimate line on the left side, box plots in the middle and violin density box plots on the right side) for both the pH and $\{H^+\}$ data for the Eastrand (n = 1380), Westrand (n = 1122), Metsämonttu (n = 642) and Felsendome Rabenstein (n = 96) abandoned mine sites.

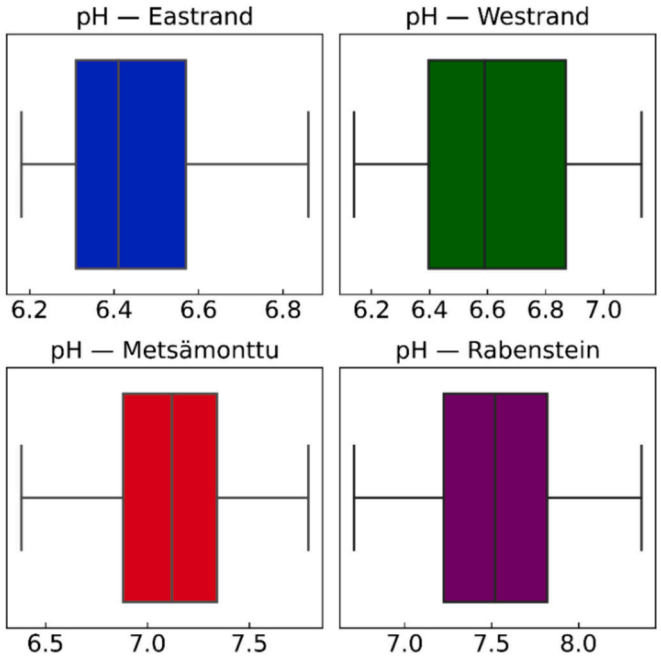


Fig. 3. Box plots of the Eastrand (n = 1312), Westrand (n = 1068), Metsämonttu (n = 609) and Felsendome Rabenstein (n = 91) clean pH data sets with removed anomalies.

the arithmetic mean, median, geometric mean and harmonic mean as the measures of central tendency for pH and the associated $\{H^+\}$ (Table 2). Due to the common but incorrect assumption of environmental data being normally distributed, several authors have resorted to using the arithmetic mean in getting the pH average (e.g. Al-Qallaf and Alali, 2022; Boutilier and Shelton, 1980; Stjernman Forsberg et al., 2008):

$$\overline{\mathrm{pH}} = \frac{\mathrm{pH}_1 + \mathrm{pH}_2 + \mathrm{pH}_3 + \ldots + \mathrm{pH}_n}{n} \tag{1}$$

Another approach used is the median which is less sensitive to anomalies than the arithmetic mean since it does not consider the magnitude of the values, only their order:

$$\mathrm{pH} = \begin{cases} \mathrm{pH}_{\mathrm{ordered}}\left[\dfrac{n+1}{2}\right] \text{ if n is odd} \\ \mathrm{pH}_{\mathrm{ordered}}\left[\dfrac{n}{2}\right] + \mathrm{pH}_{\mathrm{ordered}}\left[\dfrac{n}{2}+1\right] \text{ if n is even} \end{cases} \tag{2}$$

Alternatively, the geometric mean, introduced by Cauchy in 1821, may be an alternative for skewed distributions (Parkhurst, 1998; Vogel, 2022). This approach is less affected by anomalies or fluctuations in the sample compared to the arithmetic mean, because the geometric mean takes into account the product of all the values in the data set, rather than just the sum of the values. As a result, the geometric mean is less likely to be affected by a few very high or low values:

$$\overline{\mathrm{pH}_g} = \sqrt[n]{\mathrm{pH}_1 \bullet \mathrm{pH}_2 \bullet \mathrm{pH}_3 \bullet \ldots \bullet \mathrm{pH}_n} = \sqrt[n]{\prod_{i=1}^{n} \mathrm{pH}_i} \tag{3}$$

Table 2
pH and $\{H^+\}$ using different mathematical and positional averages; data used with removed anomalies.

Types of averages	Metsämonttu		Westrand		Eastrand		Felsendome Rabenstein	
	pH	$\{H^+\}$ (pH)	pH	$\{H^+\}$ (pH)	pH	$\{H^+\}$ (pH)	pH	$\{H^+\}$ (pH)
n	609		1068		1312		91	
Arithmetic mean	7.11	9.85×10^{-8} (7.01)	6.63	2.81×10^{-7} (6.55)	6.45	3.81×10^{-7} (6.42)	7.52	4.48×10^{-8} (7.35)
Median	7.12	7.59×10^{-8} (7.12)	6.59	2.57×10^{-7} (6.59)	6.41	3.89×10^{-7} (6.41)	7.52	3.02×10^{-8} (7.52)
Geometric mean	7.11	7.68×10^{-8} (7.11)	6.62	2.35×10^{-7} (6.63)	6.45	3.55×10^{-7} (6.45)	7.51	3.00×10^{-8} (7.52)
Harmonic mean	7.10	6.00×10^{-8} (7.22)	6.62	1.97×10^{-7} (6.71)	6.45	3.26×10^{-7} (6.49)	7.50	1.97×10^{-8} (7.71)

Another alternative to the arithmetic mean is the harmonic mean which is also less sensitive to anomalies. The harmonic mean is calculated by dividing the number of data points by the sum of the reciprocals of the values in the data set:

$$\overline{\mathrm{pH}_{\mathrm{harm}}} = \frac{n}{\left(\frac{1}{\mathrm{pH}_1} + \frac{1}{\mathrm{pH}_2} + \frac{1}{\mathrm{pH}_3} + \ldots + \frac{1}{\mathrm{pH}_n}\right)} = \frac{n}{\sum_{i=1}^{n} \frac{1}{pH_i}} \quad (4)$$

The calculated arithmetic mean for the pH data for all the abandoned mine sites is different from the arithmetic mean pH derived from the $\{H^+\}$ average of the same mine sites. This is similar to the harmonic mean which gives larger mean values for the pH derived from the $\{H^+\}$ average and relatively smaller mean values for the pH data. The median and geometric mean show accurate and precise results with both the mean for pH data and the mean for pH derived from $\{H^+\}$ average being similar in most cases. However, it might not be appropriate to use the median as the preferred type of average due to some of the pH data distribution displaying two or more populations. Therefore, based on the calculations and properties of different averages, the arithmetic mean is more appropriate for calculating the average for pH when the data is outlier free and the geometric mean when it includes outliers. This conclusion, based on evidence from four data sets of MIW data, is in line with Gaddum (1945), who wrote: "One logical consequence of the adoption of this method would be that the mean of the logarithms, or the geometric mean of the observations, would be taken as the most likely value, instead of the arithmetic mean".

3.5. Multimodal distribution average

pH data can display a mono-, bi- or multimodal distribution, where several distinct peaks indicate that the data consists of different populations (Fig. 4). Calculating the average for such distributions comes with several challenges and might not provide a meaningful description of the data (Braga et al., 2018). One of the main issues with using the mean in a multimodal distribution is that it can be highly influenced by features present in each population. Since the mean considers all data points equally, it may not accurately and precisely represent the central tendency of either population, leading to a misleading average. In addition to the challenges posed by multimodal distributions, it is important to consider the practical implications of misinterpreting average values. When dealing with data that exhibits multimodality, relying solely on traditional mean calculations can lead to incorrect conclusions about the underlying patterns. For instance, if the data represents Fe concentration at the mine water treatment plant, calculating a simple average (e.g. arithmetic mean) might not accurately reflect the quality of the water at the treatment plant. This could potentially result in inappropriate conclusions on the treatment technique and quantity of treatment reagent to use to precipitate the dissolved metals.

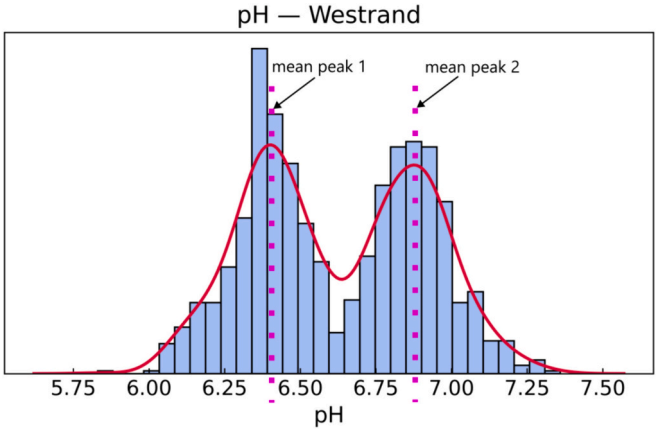


Fig. 4. Extracting the mean from each population of the bimodal distribution for the pH data set in the Westrand mine water pool data with anomalies not removed; $n = 1122$.

The data sets for the Westrand, Eastrand and Metsämonttu abandoned mines have bimodal distributions, which makes it difficult to calculate the average of the data. Therefore, clustering algorithm (*k*-means clustering) was applied to separate and analyse the two different populations of each data set (Table 3). The objective of a clustering algorithm is to group data points based on their similarity, where points within the same cluster are more similar to each other than to those in other clusters. This method can be applied to examine hidden patterns in a data set such as the presence of multiple populations within a distribution. The "*k*" in the *k*-means clustering algorithm represents the number of clusters (Ikotun et al., 2023; Lloyd, 1982; MacQueen, 1967). From the statistical analyses conducted, the data sets display bimodal distributions, so the *k* was set to the number 2 to separate the data into two clusters, each representing one of the underlying populations. The algorithm assigns each data point to one of these clusters based on its similarity to others. As a result, the cluster assignments are obtained, which indicate which population each data point belongs to. With the data separated into two populations, statistical measures for each population can then be calculated individually, such as the mean and standard deviation as they provide insights into the central tendencies and variability of each distinct population within the bimodal distribution.

4. Calculating pH average for environmental analysis — a step-by-step guide

The pH central tendency calculations provides information about the

Table 3
Central tendency and variability calculations for the bimodal distributions of the Westrand, Eastrand and Metsämonttu abandoned mine data sets using *k*-means clustering algorithm; data used with anomalies not removed; Popul.: Population. Also refer to the geometric mean in Table 2 for comparison.

	Westrand		Eastrand		Metsämonttu	
	Popul. 1	Popul. 2	Popul. 1	Popul. 2	Popul. 1	Popul. 2
n	580	542	852	528	320	322
Geometric mean	6.38	6.89	6.33	6.66	6.81	7.39
Standard deviation	0.13	0.13	0.09	0.13	0.22	0.20

acid within a system and is crucial for assessing water quality and other environmental parameters. This study was conducted to provide a step-by-step process of calculating pH averages to ensure reliable results across different scientific fields such as mine water management (Fig. 5). The process consists of six steps which have been broken down as follows:

Step 1: Data visualisation and distribution analysis — to begin, it is important to visualise the pH data using distribution graphs such as histograms, box plots or violin plots (e.g. Fig. 2). These graphs help in understanding the distribution of the data sets and identifying possible anomalies or multiple populations (e.g. Westrand pH data distribution). The purpose of this step is to determine whether the data represents one population or if there are distinct peaks indicating different pH populations.

Step 2: Identifying distribution shape — Once the data has been plotted, it is necessary to assess whether the distribution is normally distributed or not. A normal distribution has a bell shaped curve entered around the mean, with symmetrical tails on either side. If the pH data follows a normal distribution, anomalies can be removed for further analysis.

Step 3: Handling multimodal distributions — In cases where the pH data does not follow a normal distribution, it is important to identify the peaks corresponding to different populations within the data set. Once the peaks are identified, anomalies must be removed for analysis.

Step 4: Removing anomalies — Removing anomalies is a crucial step in ensuring that the calculated pH central tendency values represent the system's characteristics. Anomalies have great potential to distort the final pH central tendency calculations. By removing anomalies, the analysis becomes more robust and reliable.

Step 5: Calculating the average using the geometric mean — After data processing, the next step is to calculate the average pH value (and other measures of central tendency). This study compared the different types of averages and concluded that the geometric mean gives more reliable results compared to other types of averages (arithmetic mean, median and harmonic mean). The minimum and maximum values of the data set need to be identified so they can be related to the calculated average.

Step 6: Interpretation and conclusions — Finally, the calculated average pH value including the range of the data (minimum, maximum) is used alongside other measures of central tendency and variability,

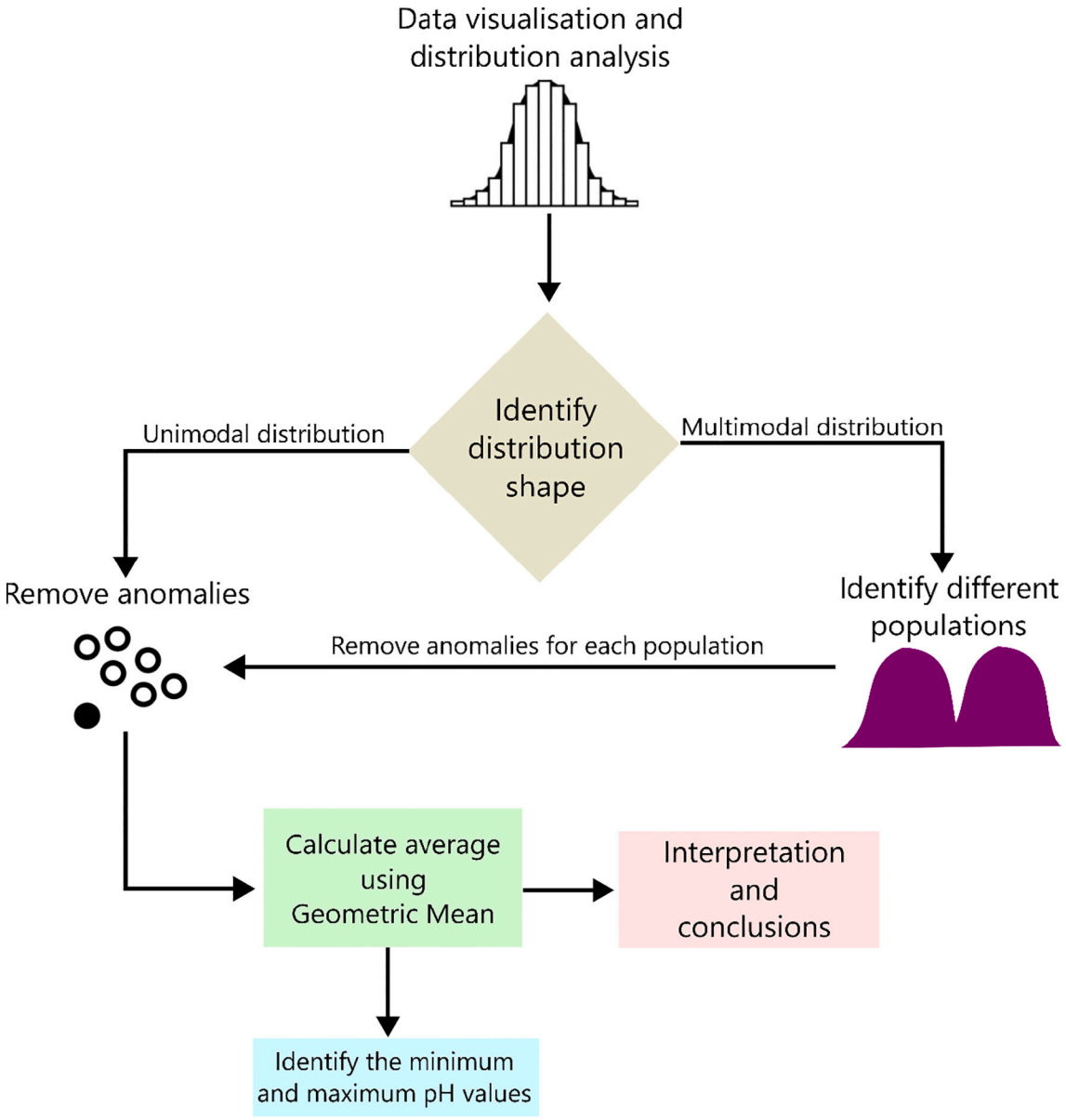


Fig. 5. A step-by-step process for calculating the pH average.

such as standard deviation, to effectively describe the data set's characteristics.

5. Conclusions and recommendations

This study highlighted the importance of calculating pH or $\{H^+\}$ and how crucial it is for understanding and assessing various environmental processes. pH measurements play an important role in fields such as water science, hydrology, soil science and environmental science, helping in evaluating the chemical characteristics of water bodies, soil and other environmental components. Good pH calculations requires standardised protocols, proper instrumentation and calibration to ensure reliable and reproducible results. The logarithmic nature of the pH scale enables the convenient representation of extreme variations in proton activity, making it applicable across a wide range of environments.

In statistical analyses, a precise value of pH or $\{H^+\}$ is commonly required to draw meaningful conclusions about the water quality or environmental components. However, it is essential to note that pH and $\{H^+\}$ values use different measurement scales, resulting in distinct distributions as elaborated in this study. Consequently, the decision of whether to use pH values or $\{H^+\}$ concentrations for calculating measures of variability and central tendency must be made with statistical considerations in mind. In this study, mine water data was used and various statistical techniques were applied to understand the distributions of pH and $\{H^+\}$. Therefore, based on the statistical results, this study recommends that the pH geometric mean be used to compute further statistical analyses, and the proposed six steps must be followed thoroughly for accurate calculations of pH central tendency and measures of variability.

Funding

This work is funded and supported by the National Research Foundation (NRF Grant UID 86948) South Africa under the SARChI Chair for Mine Water Management, the Tshwane University of Technology (TUT).

CRediT authorship contribution statement

Kagiso Samuel More: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Christian Wolkersdorfer:** Writing – review & editing, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no competing interests.

Data availability

The data supporting the findings of this study are available upon request.

Acknowledgements

Thanks to the National Research Foundation (NRF Grant UID 86948) South Africa under the SARChI Chair for Mine Water Management and the Tshwane University of Technology (TUT) for funding this project and supporting this research. We also thank our colleagues Charles Cravotta, Broder Merkel, Sylke Hilberg, Thomas Rüde, Traugott Scheytt for answering our questions, but specifically Kirk Nordstrom for a lengthy discussion. Thanking these colleges does not imply that they agree to our findings. We also thank two anonymous reviewers for their critical input.

References

Al-Qallaf, M., Alali, W.Q., 2022. Bromate concentrations and pH values in bottled drinking water in Kuwait. Environ. Monit. Assess. 194 (3), 144. https://doi.org/10.1007/s10661-022-09783-2.

Andersson, A., 2021. Mechanisms for log normal concentration distributions in the environment. Sci. Rep. 11 (1), 16418. https://doi.org/10.1038/s41598-021-96010-6.

Blythe, E.K., Merhaut, D.J., 2010. Calculating average pH in substrate research: should pH or $[H^+]$ data values be used? HortScience 45 (8), S286–S287.

Boutilier, R., Shelton, G., 1980. The statistical treatment of hydrogen ion concentration and pH. J. Exp. Biol. 84 (1), 335–340.

Boyd, C.E., Tucker, C.S., Viriyatum, R., 2011. Interpretation of pH, acidity, and alkalinity in aquaculture and fisheries. N. Am. J. Aquacult. 73 (4), 403–408. https://doi.org/10.1080/15222055.2011.620861.

Braga, A.d.S., Cordeiro, G.M., Ortega, E.M.M., 2018. A new skew-bimodal distribution with applications. Commun. Stat. Theory Methods 47 (12), 2950–2968. https://doi.org/10.1080/03610926.2017.1343851.

Cohen, E., Cvitas, T., Frey, J., Holmström, B., Kuchitsu, K., Marquardt, R., Mills, I., Pavese, F., Quack, M., Stohner, J., Strauss, H., Takami, M., Thor, A., 2008. Quantities, Units and Symbols in Physical Chemistry, 3rd edn. International Union of Pure and Applied Chemistry, Cambridge. https://doi.org/10.1039/9781847557889.

Covington, A.K., Bates, R., Durst, R., 1985. Definition of pH scales, standard reference values, measurement of pH and related terminology (recommendations 1984). Pure Appl. Chem. 57 (3), 531–542. https://doi.org/10.1351/pac198557030531.

Cravotta III, C.A., 2008. Dissolved metals and associated constituents in abandoned coal-mine discharges, Pennsylvania, USA. Part 2: geochemical controls on constituent concentrations. Appl. Geochem. 23 (2), 203–226. https://doi.org/10.1016/j.apgeochem.2007.10.003.

Cravotta III, C.A., Brady, K.B., Rose, A.W., Douds, J.B., 1999. Frequency distribution of the pH of coal-mine drainage in Pennsylvania, Report N⍰ 99-4018A. In: US Geological Survey Water Resources Investigations, South Carolina. https://doi.org/10.3133/wri994018A, 313–324 p.

Davis, J.C., 2002. Statistics and Data Analysis in Geology, 3rd edn. Wiley, New York.

Gaddum, J.H., 1945. Lognormal distributions. Nature 156 (3964), 463–466. https://doi.org/10.1038/156463a0.

Giesecke, A.H., 1979. Averaging pH vs. H^+ values. Anesthesiology 51 (5), 482. https://doi.org/10.1097/00000542-197911000-00037.

Giesecke, A.H.J., Beyer, C.W., Kallus, F.T., 1978. More on interpolation of pH data. Anesth. Analg. 57 (3), 379–380.

Grogono, A.W., 1980. Averaging pH vs. H^+ values, an irrelevant debate. Anesthesiology 53 (1), 85–86. https://doi.org/10.1097/00000542-198007000-00029.

Howe, R.H.L., Howe, R.C., 1981. Avoid errors in averaging pH values. Chem. Eng. 88 (7), 109.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J., 2023. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. Inform. Sci. 622, 178–210. https://doi.org/10.1016/j.ins.2022.11.139.

Kirby, C.S., Cravotta III, C.A., 2005. Net alkalinity and net acidity 2: practical considerations. Appl. Geochem. 20 (10), 1941–1964. https://doi.org/10.1016/j.apgeochem.2005.07.003.

Kissel, D.E., Sonon, L., Vendrell, P.F., Isaac, R.A., 2009. Salt concentration and measurement of soil pH. Commun. Soil Sci. Plant Anal. 40 (1–6), 179–187. https://doi.org/10.1080/00103620802625377.

Kuna-Broniowska, I., Smal, H., 2017. Statistical measures of the central tendency for H^+ activity and pH. Soil Sci. Ann. 68 (4), 174. https://doi.org/10.1515/ssa-2017-0022.

Lesouple, J., Baudoin, C., Spigai, M., Tourneret, J.-Y., 2021. Generalized isolation forest for anomaly detection. Pattern Recogn. Lett. 149, 109–119. https://doi.org/10.1016/j.patrec.2021.05.022.

Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28 (2), 129–137. https://doi.org/10.1109/TIT.1982.1056489.

MacQueen, J., 1967. Some Methods for Classification and Analysis of Multivariate Observations. Paper presented at the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, 1, pp. 281–297.

McCluskey, A., Lalkhen, A.G., 2007. Statistics II: central tendency and spread of data. Anaesth. Crit. Care Pain Med. 7 (4), 127–130. https://doi.org/10.1093/bjaceaccp/mkm020.

More, K.S., Wolkersdorfer, C., 2023a. Application of machine learning algorithms for nonlinear system forecasting through analytics — a case study with mining influenced water data. Water Resour. Ind. 29, 100209. https://doi.org/10.1016/j.wri.2023.100209.

More, K.S., Wolkersdorfer, C., 2023b. Exploring advanced statistical data analysis techniques for interpolating missing observations and detecting anomalies in mining influenced water data. ACS ES&T Water. https://doi.org/10.1021/acsestwater.3c00163.

Nordstrom, D.K., 2011. Mine waters: acidic to circumneutral. Elements 7 (6), 393–398. https://doi.org/10.2113/gselements.7.6.393.

Nordstrom, D.K., 2020. Geochemical modeling of iron and aluminum precipitation during mixing and neutralization of acid mine drainage. Minerals 10 (6), 547. https://doi.org/10.3390/min10060547.

Nordstrom, D.K., Alpers, C.N., Ptacek, C.J., Blowes, D.W., 2000. Negative pH and extremely acidic mine waters from Iron Mountain, California. Environ. Sci. Technol. 34 (2), 254–258. https://doi.org/10.1021/es990646v.

Pace, N.L., Ohmura, A., Mashimo, T., 1979. Averaging pH vs. H^+ values. Anesthesiology 51 (5), 482. https://doi.org/10.1097/00000542-197911000-00036.

Parkhurst, D.F., 1998. Peer reviewed: arithmetic versus geometric means for environmental concentration data. Environ. Sci. Technol. 32 (3), 92A–98A. https://doi.org/10.1021/es9834069.

Patel, J.K., Read, C.B., 1996. Handbook of the Normal Distribution. In: Statistics: A Series of Textbooks and Monographs, 2nd edn, vol 150. CRC Press, New York.

Patnaik, P., 2017. Handbook of Environmental Analysis: Chemical Pollutants in Air, Water, Soil, and Solid Wastes, 3rd edn. CRC Press, London.

Reimann, C., Filzmoser, P., 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environ. Geol. 39 (9), 1001–1014. https://doi.org/10.1007/s002549900081.

Roadcap, G.S., Kelly, W.R., Bethke, C.M., 2005. Geochemistry of extremely alkaline (pH>12) ground water in slag-fill aquifers. Ground Water 43 (6), 806–816. https://doi.org/10.1111/j.1745-6584.2005.00060.x.

Rose, A.W., Hawkes, H.E., Webb, J.S., 1979. Geochemistry in Mineral Exploration, 2nd edn. Academic Press, Cambridge, Massachusetts.

Schmiermund, R.L., Drozd, M.A., 1997. Acid mine drainage and other mining-influenced waters (MIW). In: Marcus, J.J. (Ed.), Mining Environmental Handbook: Effects of Mining on the Environment and American Environmental Controls on Mining. Imperial College Press, London, pp. 599–617.

Selvamuthu, D., Das, D., 2018. Introduction to statistical methods. In: Design of Experiments and Statistical Quality Control. Springer, Singapore. https://doi.org/10.1007/978-981-13-1736-1.

Sørensen, S.P.L., 1909. Über die Messung und die Bedeutung der Wasserstoffionenkonzentration bei enzymatischen Prozessen (On the measurement and significance of hydrogen ion concentration in enzymatic processes). Biochem. Z. 21, 131–200.

Stjernman Forsberg, L., Gustafsson, J.-P., Berggren Kleja, D., Ledin, S., 2008. Leaching of metals from oxidising sulphide mine tailings with and without sewage sludge application. Water Air Soil Pollut. 194 (1), 331–341. https://doi.org/10.1007/s11270-008-9720-1.

Stumm, W., Morgan, J.J., 1996. Aquatic Chemistry — Chemical Equilibria and Rates in Natural Waters, 3rd edn. Wiley, New York.

Thomopoulos, N.T., 2017. Statistical Distributions: Applications and Parameter Estimates. Springer, Cham. https://doi.org/10.1007/978-3-319-65112-5.

Tillmans, J., 1919. Über die quantitative Bestimmung der Reaktion in natürlichen Wässern (On the quantitative determination of the Reaction in natural waters). Zeitschrift für Untersuchung der Nahrungs-und Genußmittel, sowie der Gebrauchsgegenstände 38 (1/2), 1–16.

Verma, S.P., 2020. Basic concepts of statistics. In: Verma, S.P. (Ed.), Road from Geochemistry to Geochemometrics. Springer, Singapore, pp. 227–246. https://doi.org/10.1007/978-981-13-9278-8_4.

Vogel, R.M., 2022. The geometric mean? Commun. Stat. Theory Methods 51 (1), 82–94. https://doi.org/10.1080/03610926.2020.1743313.

Wolkersdorfer, C., 2017. Mine Water Hydrodynamics, Stratification and Geochemistry for Mine Closure–The Metsämonttu Zn-Cu-Pb-Au-Ag-Mine, Finland. Paper presented at the International Mine Water Association — Mine Water & Circular Economy, Lappeenranta, vol. I, pp. 132–139.

Wolkersdorfer, C., 2022. Mine Water Treatment – Active and Passive Methods. Springer, Berlin. https://doi.org/10.1007/978-3-662-65770-6.

Wolkersdorfer, C., Komischke, H., Pester, S., Hasche-Berger, A., 2013. Hydrodynamics in a Flooded Underground Limestone Mine. Paper presented at the International Mine Water Association — Reliable Mine Water Technology, Denver, vol. II, pp. 1165–1172.

Xu, D., Wang, Y., Meng, Y., Zhang, Z., 2017. An Improved Data Anomaly Detection Method Based on Isolation Forest. Paper presented at the 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, 2, pp. 287–291. https://doi.org/10.1109/ISCID41719.2017.