The Influence of Data Standardization on Cluster Analysis: A Case Study of Marajoara Ceramics

Amirah Muhammad
Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia.
amirahm7554@gmail.com

Correspondence should be addressed to Amirah Muhammad : amirahm7554@gmail.com.

Article Info
Journal of Computational Intelligence in Materials Science (https://anapub.co.ke/journals/jcims/jcims.html)
Doi: https://doi.org/10.53759/832X/JCIMS202402006.
Received 02 December 2023; Revised from 02 March 2024; Accepted 27 March 2024.
Available online 31 March 2024.
©2024 The Authors. Published by AnaPub Publications.
This is an open access article under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Abstract – The importance of data standardization from the viewpoint of various enterprises is explored in this study. In this article, we'll look at how data standards have evolved over time and how application programming interface (APIs) have become the de facto norm. Promoted system-to-system interoperability, less translation hurdles, and elimination of missing data issues are just a few of the benefits of data standardization highlighted in the paper. Multiple approaches to data normalization are investigated in the research as well. These include maximum scores, logarithms, and z-scores. This study compares three different standardization approaches to cluster analysis, looking at their effects and weighing the pros and downsides of each. Also included are the results of tests conducted with two databases and a study of data standardization as it pertains to Marajoara ceramics. The findings show that certain standardization methods work well when looking for correlations between different variables. Data standardization and its implications in many academic and corporate contexts are thoroughly examined in this work.

Keywords – Cluster Analysis, Data Standardization, Marajoara Ceramics, Multi-Channel Analyzer, Self-Organizing Map, Min-Max Standardization, Application Programming Interface.

I. INTRODUCTION

The term “data standardization” refers to the practice of creating guidelines for all aspects of data management. Data components that could be subject to standards include the nature of the data that will be collected, the language, framework, and structure of the dataset, the specifics of storage (like the location), and procedures for data portability. The first data standard was established after to the conclusion of World War II, as a result of the intricate logistical challenges encountered during the 1948 Berlin Airlift [1]. Unloading produced bottlenecks that impeded air travel, as ground staff were required to meticulously inspect extensive lists of commodities transported by each aircraft. In order to address this issue, a uniform system of codes was established, enabling the computerized submission of shipment notifications prior to the aircraft's arrival. APIs are the most often used data standards.

These are computer protocols that provide the rules for communication between software components. Application programming interface **(**APIs**)** facilitate the transfer of data by specifying the types of data that may be obtained, the method for obtaining it, and the format in which the data will be exchanged. In addition, they may include the accompanying metadata, which provide descriptions of the properties or semantics of the data and allow users to interpret the meaning and importance of various data points. However, although several businesses have reached a consensus over which APIs to use, many other industries lack agreement on how APIs should be determined and if open, standardized APIs are more desirable. Moreover, APIs do not inherently address issues related to data conversion and the absence of data.

Standardizing data has the ability to eliminate any barriers that may hinder others from using the data. Initially, it may reduce uncertainty related to metadata by mandating that data semantics adhere to specific standards and regulations. An instance of this is the standardized dot matrix typeface collection designed specifically for the purpose of exchanging information using Chinese ideograms [2]. The standard facilitates interoperability across multiple datasets by maintaining the uniformity of tens of thousands of Chinese characters used across all datasets. Data standards may also mitigate barriers to data transformation, such as by homogenizing the structure and arrangement of information. For instance, several public transportation companies have implemented a uniform structure for documenting public transit timetables and related geographical data.

This standard promotes data interoperability by presenting a modular collection of compatible data, together with standardized data models and schemas. The development of multi-modal synchronized travel planner software, such as

Moovit and Google Maps, has been made feasible. Ultimately, the process of data standardization may effectively mitigate the issue of missing data. The National Coordinator's office in the department of Health Information Technology in USA, for instance, establishes criteria for gathering data about patients' allergies [3]. Naturally, data standardization is but a single approach to tackle barriers in sharing or integrating data, and it may entail substantial expenses— factors that will be examined in detail.

The future progress of physicochemical approaches will lead to a substantial rise in the number of produced findings. To analyze the information, it is important to use advanced approaches like multivariate frameworks. Generally, multivariate statistical approaches enable the assessment of a group of samples based on the interrelationships among variables. These approaches operate under the assumption that every sample may be shown as a point in a multifaceted space. In this space, each axis represents a specific physicochemical constituent of the samples. One method to confirm the presence of comparable behaviors among the samples with respect to various factors is to conduct a clustering analysis. An issue that occurs in cluster analysis is the choice of whether to normalize the samples prior to computing the distance figures. There are a lot of normalization procedures, which makes this choice even more complicated.

The purpose of this work is to explore the influence of three standardization procedures on cluster analysis: log, enhanced min-max [4, 5]. A Self-Organizing Map (SOM) neural network is fed the data once data standardization has been implemented. This network's objective is to aggregate samples into groups with the aim of establishing internal group homogeneity and exterior group heterogeneity [6]. An SOM that can learn without human supervision is the SOM network. Fundamental to SOM competitive learning technology is the idea that neurons compete with one another in response to a sample. Next, the top-performing neuron modifies its weights to improve its responsiveness to inputs from the network. Neurons also work in tandem with their neighboring neurons, which also change, via a process. A certain part of the system associated with a given category will be activated by the sample characteristics.

Data standardization is crucial in many industries and contexts, and this research explains why. The benefits of data standardization, as highlighted in the text, include facilitating interoperability, lowering barriers to the transformation of data, and fixing issues with missing data. Data standardization methods are also covered in the paper. These methods include leveraging APIs, the z-score algorithm, logarithmic transformations, and min-max improved standardization. The study also evaluates several standardization processes in the context of Marajoara ceramics and investigates how to use data standardization to cluster analysis. This effort aims to improve comprehending of data standardization and how it affects data analysis and interpretation.

The remaining sections of the article are structured as follows: In Section II, we survey the literature on cluster analysis and data standardization. Section III presents a discussion of the materials and methodology regarding standardization methods, and validation indeces. Section IV presents a critical discussion of the findings in this paper. Lastly, a summary of the findings in this paper is provided in Section V.

II. RELATED WORKS

According to Chua et al. [7], the obstacles to data standardization would be comprehensible if the only benefit was improved systemic stability at a later stage. This section demonstrates the advantages of well-crafted and well executed data standardization, which extend beyond times of crisis. By doing this, it demonstrates why implementing data standards seems to be a simpler task compared to post-crisis reforms that only incur expenses during non-crisis times or that prioritize the public's benefit above regulated institutions. Standards may further aid in decreasing expenses associated with regulatory compliance. As the comprehension and use of standards increase, the need for redundant information in slightly varied formats should decrease in response to regulatory requirements. Standardization facilitates effective communication between companies and authorities by guaranteeing exact definition of the measurements and their underlying rationale.

According to Schaffer and Green [8], there are many different methods available for standardizing variables. Hazra and Gogtay [9] focused on the scenario that involves numerical variables (NV). Only classification variables, or a combination of category and NV, are excluded from consideration. Researchers with social science backgrounds often make the assumption that normalized variable has been adjusted to have a variance of one and an average of zero, as determined by the commonly used "z-score" calculation. Nevertheless, other suggestions for the scaling or variables standardization are also in the categorization literature, and this section provides a comprehensive overview of them. In this article, the word standardization will be used generically for the sake of convenience. The first method of normalization involves using the z-score algorithm to convert normal variables into standard score format:

$$Z_1 = (X - \bar{X})/s, \tag{1}$$

Let X represent the original information value, and let X and s be the standard deviation (SD) and sample mean. The variable after transformation will have a variance value of 1.00 and a mean value of 0.0. The suggested transfiguration has been advocated by many writers, such as [10], [11], and [12]. Sp[th] cautions that the performance of Z_1 may be compromised if there are significant disparities in the within-cluster standard deviations.

According to Milligan and Cooper [13], it is crucial to emphasize that the standardization Z1 should be implemented globally, including all items inside the variable, rather than within specific clusters. In order to comprehend this limitation, let us examine a scenario where the data contains three distinct clusters that are widely spaced apart. Furthermore, suppose a sample point is located at each of the three cluster centroids. Implementing standardization inside clusters would result in

scores for vector for each of the three centroid locations, with all entries being assigned a value of zero. Any clustering algorithm used to the standardized data will typically group the three centroid locations together in a single cluster.

The same result occurs, according to Ichikawa and Morishita [14], when the three points are positioned in respect to each of the three cluster centroids at the same standardized coordinate vector. For instance, in a two-dimensional space, there will exist three distinct places with coordinate values of (1.0, -1.0). Data points located in close proximity to these three distinct places would have minimal interpoint distances and are likely to form clusters. Therefore, these findings would result in an inaccurate and very deceptive clustering solution. Therefore, it is imperative to avoid using Z1 when calculating standardization inside a cluster. The subsequent normalization is analogous to Z1 and is calculated as:

$$Z_2 = X/s \tag{2}$$

Applying Formula Z_2 will provide a converted variable with a transformed mean equal to X and a variance of 1.00 divided by s. Since the scores have not been normalized by removing the mean, the data on the relative position of the scores is still preserved. Therefore, Z_2 does not encounter the issue of losing information on the cluster centroids, which was previously mentioned as a problem for Z_l. The use of Z_2 has been suggested by Fu and Kane [15]. It is important to observe that Z1 and Z_2 are directly proportional to each other. Therefore, when world variances and means are used, the Euclidean distances calculated using the two formulae result in equal dissimilarity values. The third approach involves standardization by using the maximum score on the variable:

$$Z_3 = X/Max(X) \tag{3}$$

If all values are non-negative, then the converted variable functions as a ratio measure, with all scores ranging from 0.0 to 1.0. If there are negative values among X's, it is possible to add a sufficiently big positive constant to all values in order to achieve the proportionality condition. The SD and mean after transformation are X divided by the maximum value of X, and s divided by the highest value of X, respectively.

Hageman [16] introduced a mathematical conversion known as "Rankits." A rankit is calculated as the mean deviation of the r^{th} biggest value in a sample of n inspections randomly taken from a standard normal distribution. The adaptation of this upgrading might be used in a clustering environment. However, the approach is necessary only in circumstances when the analysis is sensitive to deviations of information from a normal distribution. In a similar vein, Henderson [17] introduced a specific technique for adjusting measurements in the presence of a reference standard. Typically, reference standards needed for Stoddard's approach are only available in laboratory research carried out in the fields of chemistry or biology.

Woodall et al. [18] introduced the concept of scaling and advocated for the use of Mahalanobis or generalized distance as a comprehensive solution to several issues, instead of relying on basic Euclidean distance. Nevertheless, Martino et al. [19] has contested the use of Mahalanobis distance in clustering scenarios due to the challenge of determining a suitable variance-covariance matrix. In addition, we disagree with the notion that, in the majority of circumstances, the variables chosen for a cluster analysis constitute a random test of the variables accessible to the researcher. According to the findings examined by Steinley and Brusco [20], the process of choosing variables for a cluster analysis requires careful consideration.

Charalampidis [21] have examined k-means clustering from several angles. Several influential data elements that may significantly impact the effectiveness of the k-means algorithm have been found and resolved. One of the factors is data standardization. In marketing research, it is customary to normalize the columns of a database using the variables data matrix. This involves adjusting the values to have a mean of zero and a standard deviation of one. After this standardization is completed, the entities that match the rows of the matrix are clustered. When variables are quantified using numerous scales, standardization becomes even more important. For example, combining a variable measuring age with a variable measuring wealth (typically expressed in thousands of dollars) is recommended in a cluster analysis. The rationale for this is that the resultant clusters will be more affected by the magnitude of the income component.

To make sure all the variables have the same amount of variance, normalizing them is important before attempting k-means clustering, according to Ikotun et al. [22]. The goal is to avoid groupings where the most variable-intensive variables have a disproportionate amount of weight. When variables are quantified using numerous scales, standardization becomes even more important. Cluster analysis may not be the best place to combine two variables that reflect age and wealth, the former of which may be measured in thousands of dollars. The rationale for this is that the resultant clusters will be more affected by the magnitude of the income component.

Data standardization and its many uses are the main topics of the paper. This article takes a look at the origins and relevance of data standards, focusing on APIs and data's ability to collaborate. Furthermore, it delves into the possibility of data standardization in lowering regulatory compliance costs, resolving issues with missing data, and removing barriers to data usage. Several methods and techniques for normalizing data are discussed in the article. These include min-max scaling, logarithms, and z-scores. Cluster analysis and self-organizing maps are also covered in terms of data standardization's potential applications. Along with discussing the need of standardization and its results, the article offers a case study of ceramic data standardization. The page provides a comprehensive overview of the main academic studies on data standardization, which includes all the different applications and methods.

III. MATERIALS AND METHODOLOGY

Standardization methods

In several actual measurements, the raw data, and applications of cluster analysis, are not directly used unless there is a probabilistic framework for pattern development [23]. Hence, it is necessary to preprocess the cluster analysis data by applying a transformation that aims to standardize the data. This study examines three standardizing techniques: logarithm, min-max enhanced [24], and Generalized-Log Spectral Mean [25]. In the enhanced min-max standardization (MMS) method [26], a collection of R_K is formed for each column, consisting of values that are repeated more than once. R_{KA} is obtained by adding the SD and mean of R_K. Finally, the enhanced MMS is implemented using the given phrase:

$$F(x) = \begin{cases} \dfrac{x_k - min(x_k)}{2(R_{KA} - \min(x_k))}, x_k \le R_{KA} \\ 0.5 + \dfrac{x_k - R_{KA}}{max(x_k) - R_{KA}}, x_k > R_{KA} \end{cases} \tag{4}$$

$R_{KA} = R_K\text{avg} + R_K\text{std}$, where $R_K$avg= mean (R_{KA}) and $R_K$std= standard deviation (R_{KA}). One of the foundational functions of log-generalized mean spectral normalization n is the q-logarithm function, also known as the generalized logarithmic function:

$$\bar{x}^k = exp_q\left(\frac{log_q(x^k) - \frac{1}{N}\sum_{i=1}^{N-1} log_q x^i}{1 +_q \frac{1}{N}\sum_{i=1}^{N-1} log_q x^i}\right) \tag{5}$$

Where

$$log_q(x) = \begin{cases} \dfrac{x^q - 1}{q}, q \ne 0 \\ log(x), q = 0 \end{cases} \tag{6}$$

$$exp_q(x) = \begin{cases} (1 + qx)^{\frac{1}{q}}, q \ne 0 \\ exp(x), q = 0 \end{cases} \tag{7}$$

Validation indeces

By contrasting the output of the SOM neural network with previously established data, the indices detailed below assess the clustering algorithm's efficacy. There are four main parts to the process of self-organization formation:

Activation: The connection weights are configured with randomly generated modest values.

Competition: Each input pattern is processed by the neurons, which calculate the discriminant function (DF) values that serves as the foundation for further computation.

The neuron with the lowest value of the DF is designated as the winner. For example, input with m dimensions:

$$\bar{X} = [x_1 x_2 - - - x_n]^T \tag{8}$$

$$W_j = [w_{j1} w_{j2} - - - w_{jm}]^T \tag{9}$$

Let m represent the number of input neurons and l represent the total number of output neurons in the network. The winner will be determined by the best $\vec{X}$ and $\overrightarrow{W_j}$:J match, with j being the winning index and the winning weight vector being the appropriate weight. One definition of the discriminant function is the squared distance between W_j for each neuron j, and the input vectors x as measured in terms of the Euclidean distance:

$$D(j) = \sum (W_{ij} - x_{ij})^2, i = 1 to\ n\ \&\ j = 1\ to\ m \tag{10}$$

An easy way to transform the continuous input space into the discrete output space of neurons is via a fundamental phenomenon called competition.

Cooperation: The victorious neuron will establish the geographical position of the topological vicinity of stimulated neurons, thereby forming the foundation for collaboration among surrounding neurons. The topological neighborhood of neurons may be defined by the physical distance between neurons i and j on the grid of neurons:

$$T_{j\ I\ (x)} = expexp\left[-\frac{S^2_{j,I(x)}}{2\sigma^2}\right] \tag{11}$$

The index of the winning neuron is denoted by $I(x)$. These qualities include the maximality of the value at the winning neuron, the symmetry around that neuron, the monotonically decreasing nature of the value to zero as distance approaches infinity, and the independence of the value from the position of the winning neuron.

Spatial Adaptation: The stimulated neuron reduces its DF values in response to the input pattern by adjusting the CW. This adjustment improves the winning neuron response to similar input patterns. In a topographic neighborhood, the weight of the winning neuron is updated, along with the weights of its neighboring neurons. The weight update equation is defined as follows:

$$W_{j\,i,} = \propto (t).T_{j\ ,I(x)}(t).(x_i - W_{ji,} \tag{12}$$

Where t represents a certain moment in the period and $\propto$ denotes the learning rate.

$$\propto (t) = \propto_0 \exp exp\left[\frac{-t}{n_\propto}\right] \tag{13}$$

and the modifications are performed to every training instance x throughout many epochs.

It is believed that there are two partitions, one generated by the SOM NN and the other including further data about the base. Consider A and B as two divisions. The Jaccard index (JI) is a well-recognized metric for quantifying the similarity of groups based on the absence or presence of samples. It is commonly used in automatic classification. The algorithm calculates the count of sample pairs that belong to the same category in both B and A partitioners, as well as the count of sample pairs that belong to the same group in at least one of the partitions. The JI, often known as the coefficient of similarity, is calculated as follows:

$$R = \frac{a}{a+b+c} \tag{14}$$

How many sample pairs in both A and B belong to the same cluster is shown by the variable 'a'. The variable 'b' represents the number of sample pairs that belong to separate categories in A, but the same category in B. The variable 'c' represents the samples number belong to the same category in A, but distinct categories in B. The Rand index is an analytical metric that quantifies the ratio of sample pairs that belong to the same or distinct categories in two secessions. It is formally described by Brouwer and Heeger [27]:

$$R = \frac{a+d}{a+b+c+d} \tag{15}$$

where constants c, b, and a are unchanged from the preceding index, and d is the samples number that belong to separate categories in B and A. The Fowlkes-Mallows index [28] calculates the geometric average of the percentage of the sample pair that belong to the same category in both divisions. Consider two segmentations of A and B, which have an equal amount of samples. Consider the matrix $m = [m_{ij}]$, where i and j range from 1 to k. Each m_{ij} represents the samples number that are common to the j^{th} cluster (B) and the i^{th} cluster (A). The resemblance measurement suggested by Campello [29].

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}} \tag{16}$$

Where

$$P_k = \sum_{i=1}^{k} m_i^2 - n, Q_k = \sum_{j=1}^{k} m_j^2 - n, T_k = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}^2 - n, m_i = \sum_{j=1}^{k} m_{ij,}\ m_j = \sum_{i=1}^{k} m_{ij,}\ n = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij.} \tag{17}$$

For the experiments, two information sites were used, with one including 298 ceramic pieces originating from the Marajó island. The Marajó island has an area of 40,552 square kilometers and is situated inside an archipelago located at the entrance of the Amazon River. The Amazon area has a moist winter characterized by alternating periods of floods and drought, which posed challenges for the pre-Columbian people. Approximately 15 centuries before to the colonization of the land, emerged one of the most fascinating native civilizations in America. The civilization was distinguished by the creation of massive earthen mounds, reaching heights of up to 12 units and covering an area of 3 hectares. Additionally, they produced intricate pottery vessels and other ceremonial things.

The study of Marajoaras ceramics has received significant attention since the 19th century, particularly about the vessel's purpose, manufacturing techniques, and artistic style. The Marajoara or Marajó culture thrived during the pre-Columbian period on Marajó island, located near the mouth of the Amazon River in northern Brazil. The civilization seems to thrive between 800 AD to 1400 AD, as shown by Oliver [30]. Morales et al. [31] have recorded evidence indicating that human activity occurred at these locations as early as 1000 BC. The culture seems to have endured throughout the colonial period.

In the latter part of the 20th century, Lobao and Meyer [32] conducted a concentrated study on society, specifically examining social structure, and subsistence patterns. Radiocarbon dating indicates that the Marajoara civilization saw its most significant era of development and expansion among the 5th and 14th centuries. Marajoara pottery is part of the polychrome style, which is known for its intricate ceremonial objects that have elaborate shapes and decorations. Decorative methods include the application of paint, the act of cutting into a surface, the removal of material, and the act of scraping.

The pottery is infused with pulverized ash derived from the bark of a tree often referred to as caraiapé. This substance, which is related to other ceramics in the Polychrome style, was used in the Amazon Basin toward the conclusion of 1st millennium.

IV. RESULTS AND DISCUSSION

The exterior surface of the ceramic power samples was cleaned, and then the drill was used to collect the samples. The drill had a flexible shaft and a tungsten carbide rotary file embedded to its tip. then, the materials were subjected to drying in an oven at a temperature of 105 °C for a duration of 24 hours, and then kept in a desiccator. The standards utilized in the analysis were the Elements Constituting in NIST-SRM-1633b, Coal Fly Ash, and the samples were checked using the IAEA Soil-7, Trace Elements in Soil. The materials underwent a drying process in an oven at a temperature of 105 °C for a duration of 4 hours. Approximately 100 mg of several ceramic samples (CS), namely IAEA-Soil-7, and NIST-SRM-1633b were measured and placed in plastic bags. The bags were then sealed with aluminum foil.

Eight samples, together with one reference material, were assembled and exposed to irradiation in the investigation reactor segment, IEA-R1, at IPEN CNEN/SP. The irradiation was conducted at a tepid baryon flux of about 5 × 1012 cm-2 × s-1 for a duration of 8 hours. The experiment included doing two measurement series using a Ge (hyper pure) detector, namely the GX 1925 model from Canberra. At the 60Co gamma peak at 1332.49 keV, the detector's resolution was 1.90 keV. The measurements were performed using the S-100 MCA (Multi-Channel Analyzer) from Canberra, which had 8192 channels. The study of concentrations determination and gamma ray spectra were performed utilizing the Genie-2000 baryon Activation study manufacturing process developed by Canberra. A comprehensive account of the technique, the specimens, the established protocol for preparing the standard samples, and the archaeological locations were previously documented by Cartwright [33]. The **Table 1** and **2** provide the SD and means of the B2 and B1 datasets.

Table 1. The SD, mean, and range of CS from the B1 dataset, measured in mgg-1. The sample size for this dataset is 298, unless stated otherwise

Element	**Mean ± sd**	**Range**
Th	18.17±1.40	14.10-23.10
Hf	7.34±1.53	3.60-13.75
Tb	1.07±0.29	0.01-2.10
Eu	1.80±0.26	0.90-2.57
Cs	9.17±4.97	4.36-91.40
Fe(%)	5.42±1.15	1.76-12.21
Cr	94.50±49.96	57.80-933.34
Sc	18.74±4.30	12.30-88.13
U	4.06±0.65	2.60-6.90
Lu	0.59±0.06	0.25-0.77
Yb	3.83±0.47	0.50-5.00
La	54.68±7.70	4.01-73.30
K(%)	1.96±0.70	0.01-4.20
Na(%)	0.45±0.14	0.04-1.20

Table 2. The ceramic samples from the B2 datasets analyzed for their means and standard deviations. The measurements were expressed in milligrams per gram (mg/g^{-1})

Elements	**Rezende (n=31)**	**Agua Limpa (n=81)**	**Prado (n=34)**
U	1.37±0.23	1.37±0.29	4.24±0.87
Th	6.40±0.77	12.78±1.91	17.47±0.96
Sm	10.48±1.61	9.66±1.40	7.45±0.63
Sc	43.99±3.06	15.61±2.34	29.66±2.02
Nd	52.45±9.06	58.50±10.72	38.23±7.59
Na	158.70±40.43	1948.22±576.09	565.08±107.71
La	37.72±6.57	71.50±10.73	33.23±3.97
Hf	11.49±0.74	8.36±1.02	8.87±0.69
Fe	10821.61±2375.87	33461.73±7753.766	28535.29±5639.03
Eu	3.20± 0.45	2.50±0.38	1.40±0.16
Cr	218.34±27.97	160.73±30.48	13820±20.61
Ce	85.21±34.71	122.68±20.93	115±9.92
As	1.86±0.49	2.23±1.01	1.57±0.38

The experiments were conducted using two databases: B1, which consists of 298 samples representing Marajoara pottery, and B2, which has 146 samples representing ceramics from three different locations. The elements identified in the B1 datasets were sodium (Na), potassium (K), lanthanum (La), ytterbium (Yb), lutetium (Lu), uranium (U), scandium (Sc),

chromium (Cr), iron (Fe), cesium (Cs), europium (Eu), terbium (Tb), and hafnium (Hf). The identified elements for the B2 datasets include U,As, Th, Ce, Sm, Cr, Sc, Eu, Nd, Fe, La, Hf, and Na. The mass fraction (MF) of each sample was determined using INAA. Tables 1 and 2 provide the MF of the 146 and 298 samples. **Fig 1** and **2** show the graphs of the principal component analysis (PCA), specifically the relationship between PCA1 and PCA2, for B2 and B1. PCA is a method that linearly transforms a collection of p variants into a tiny set of k non-correlated variants. This transformation aims to capture a significant percentage of the covariance in the data.

Principal Component Analysis (PCA) [34] involves using eigenvector techniques to alter the dataset, enabling the identification of the direction and extent of the highest variance. PCA starts by selecting the p variables that are correlated, and then converts them into a collection of p new variables that are uncorrelated. Principal Component Analysis (PCA) offers a method to decrease the number of dimensions in a dataset while preserving the maximum amount of information. The primary components are the converted variables obtained from the original variables. The PCs are arranged in a manner so that the first element elucidates the most significant proportion of the variance, followed by the second element explaining the second most substantial proportion, and so on. The plot depicting the first two main elements produced by PCA for the B1 and B2 datasets is shown in **Fig 1**. The base graph of B1 exhibits a significant convergence of sample groups, each representing distinct chemical compositions. Nevertheless, as seen in **Fig 1**, the graph displays three distinct groups that are noticeably well-separated. Each cluster exhibits distinct characteristics that highlight the variation in the basic materials used for manufacturing the samples.

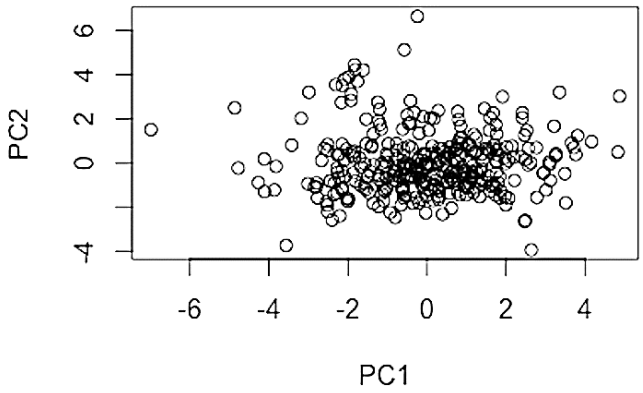

Fig 1. Scatter plot (SP) displaying the distribution of base B1 data after projecting it using two main factors

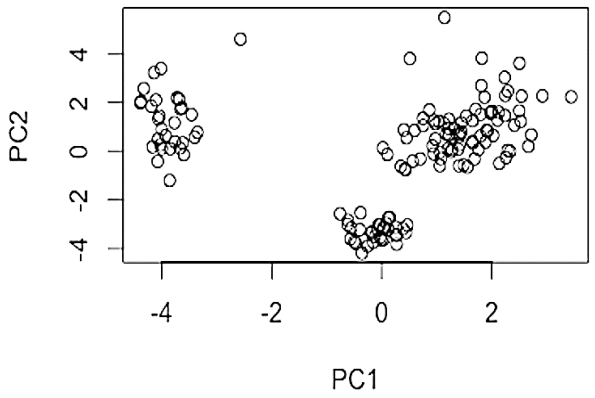

Fig 2. SP displaying the distribution of base B2 data after it has been projected using two main components

The assessment of the influence of standardization on congregating analysis using independent maps was conducted by using the authentication indices of Jaccard [35], Fowlkes-Mallows [36], and Rand [37]. According to Vesanto and Alhoniemi [38], a higher index corresponds to a more favorable outcome achieved by the SOM NN. The outcomes derived from the cluster analysis of the converted data for B2 and B2 databases are shown in **Fig 3** and **4**. According to **Fig 3**, the tests conducted with B1 indicate that the enhanced min-max standardization approach exhibited superior performance. The Rand indices, Fowlkes Mallows, and Jaccard yielded values of 0.77, 0.79, and 0.64. The figures acquired by log normalization were 0.57, 0.71, and 0.54, whereas the figures obtained using log generalized standardization (LGS) were 0.62, 0.50, and 0.33. The enhanced min-max method outperformed the log-generalized and arithmetic methods due to the inclusion of the SD of the captivations in the indexes. Additionally, the two clusters had a very comparable chemical makeup.

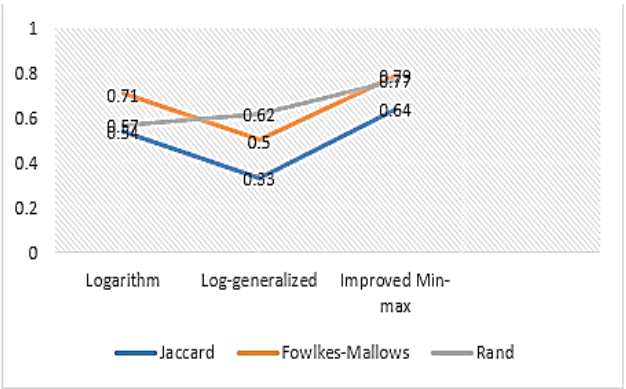

Fig 3. Displays the indices acquired after the use of normalization procedures in the B1

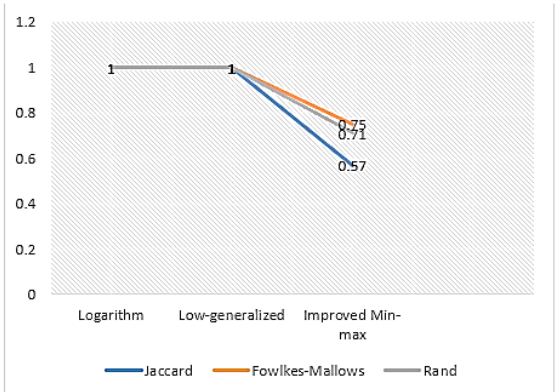

Fig 4. Indices derived from the use of standardizing processes in the B2 database

Fig 2 demonstrates that the specimen inside each cluster constitute a tightly-knit chemically homogenous group, indicating a significant level of chemical similarity among them. This is evident from the PCA1 vs. PCA2 plot. From an archaeological perspective, the findings indicated that the clay found in pottery pieces from the three sites came from three separate raw materials. Nevertheless, the enhanced MMS exhibited the poor performance when applied to dataset B2, as seen in **Fig 4**, with Jaccard, Fowlkes-Mallows, and Rand validation index values of 0.57, 0.75, and 0.71, respectively. **Fig 2** demonstrates that the samples inside each cluster are distinctly segregated, forming a tightly-knit chemically homogenous group. This indicates a significant level of chemical resemblance among the samples. However, both log and LGS yielded validation indices with a value of 1.

V. CONCLUSION

Data standardization is an essential procedure in data management that sets standards and regulations for organizing, communicating, and transferring data. It has the capacity to remove obstacles to data use, promote compatibility across different systems, address problems related to missing data, and decrease costs connected with regulatory compliance. API standardization and the demand for open, standardized APIs are themes that provoke disagreement and controversy. The primary focus of future data standardization research should be to prioritize resolving issues and disparities in API standardization across various businesses. One way to accomplish this goal is to provide guidelines or frameworks for developing and implementing acceptable and mutually beneficial API standards Future research could investigate the effects of different data standardization methods on other data types, including statistical and distributional variables, and their combinations to determine the most effective standardization methods for different data conditions with this approach this is used. To improve the accuracy and efficiency of the clustering algorithms, future research in cluster analysis should investigate the effectiveness of other standardized methods, such as log-normalized, logarithmic, minimum-maximum-augmented methods and this provides valuable insights into the best standardization techniques for various clustering applications. Future research could explore the applicability of data standards in non-traditional disciplines such as archaeology. Analysis of archaeological data as a case study using Marajoara pottery revealed the use of standardized working methods. The more we study in different archaeological contexts, the more we can learn about the relationships between objects and archaeological sites.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding was received to assist with the preparation of this manuscript.

Competing Interests

There are no competing interests.

References

[1]. R. G. Miller, “To Save a City: The Berlin Airlift, 1948-1949,” *The SHAFR Guide Online*. Oct. 02, 2017. doi: 10.1163/2468-1733_shafr_sim140160374.

[2]. D. A. Jabs, R. B. Nussenblatt, and J. T. Rosenbaum, “Standardization of UVeitis Nomenclature for Reporting Clinical Data. Results of the first international Workshop,” *American Journal of Ophthalmology*, vol. 140, no. 3, pp. 509–516, Sep. 2005, doi: 10.1016/j.ajo.2005.03.057.

[3]. D. Blumenthal, C. M. DesRoches, K. Donelan, S. Rosenbaum, and T. G. Ferris, “Health Information Technology in the United States: The Information Base for Progress,” *HEALTH POLICY AND MANAGEMENT FACULTY PUBLICATIONS*, Jan. 2006, [Online]. Available: https://hsrc.himmelfarb.gwu.edu/cgi/viewcontent.cgi?article=1473&context=sphhs_policy_facpubs

[4]. O. M. Elzeki, M. Z. Reshad, and M. A. ElSoud, “Improved Max-Min algorithm in cloud computing,” *International Journal of Computer Applications*, vol. 50, no. 12, pp. 22–27, Jul. 2012, doi: 10.5120/7823-1009.

[5]. J. N. Darroch and D. Ratcliff, “Generalized iterative scaling for Log-Linear models,” *Annals of Mathematical Statistics*, vol. 43, no. 5, pp. 1470–1480, Oct. 1972, doi: 10.1214/aoms/1177692379.

[6]. L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, and K. Hoagwood, “Purposeful sampling for qualitative data collection and analysis in mixed method implementation research,” *Administration and Policy in Mental Health and Mental Health Services Research*, vol. 42, no. 5, pp. 533–544, Nov. 2013, doi: 10.1007/s10488-013-0528-y.

[7]. I. S. Chua *et al.*, “Artificial intelligence in oncology: Path to implementation,” *Cancer Medicine*, vol. 10, no. 12, pp. 4138–4149, May 2021, doi: 10.1002/cam4.3935.

[8]. C. M. Schaffer and P. E. Green, “An Empirical comparison of variable standardization methods in cluster analysis,” *Multivariate Behavioral Research*, vol. 31, no. 2, pp. 149–167, Apr. 1996, doi: 10.1207/s15327906mbr3102_1.

[9]. A. Hazra and N. J. Gogtay, “Biostatistics series module 3: Comparing groups: Numerical variables,” *Indian Journal of Dermatology*, vol. 61, no. 3, p. 251, Jan. 2016, doi: 10.4103/0019-5154.182416.

[10]. T. Cole, “Fitting smoothed centile curves to reference data,” *Journal of the Royal Statistical Society*, vol. 151, no. 3, p. 385, Jan. 1988, doi: 10.2307/2982992.

[11]. J. T. Holladay, J. R. Moran, and G. M. Kezirian, “Analysis of aggregate surgically induced refractive change, prediction error, and intraocular astigmatism,” *Journal of Cataract and Refractive Surgery*, vol. 27, no. 1, pp. 61–79, Jan. 2001, doi: 10.1016/s0886-3350(00)00796-3.

[12]. A. I. Fleishman, “A method for simulating non-normal distributions,” *Psychometrika*, vol. 43, no. 4, pp. 521–532, Dec. 1978, doi: 10.1007/bf02293811.

[13]. G. W. Milligan and M. Cooper, "A study of standardization of variables in cluster analysis," *Journal of Classification*, vol. 5, no. 2, pp. 181–204, Sep. 1988, doi: 10.1007/bf01897163.

[14]. K. Ichikawa and S. Morishita, "A Simple but Powerful Heuristic Method for Accelerating <formula formulatype="inline"><tex Notation="TeX">k</tex></formula> -Means Clustering of Large-Scale Data in Life Science," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 4, pp. 681–692, Jul. 2014, doi: 10.1109/tcbb.2014.2306200.

[15]. L. Fu and C. L. Kane, "Time reversal polarization and aZ2adiabatic spin pump," *Physical Review B*, vol. 74, no. 19, Nov. 2006, doi: 10.1103/physrevb.74.195312.

[16]. S. J. Hageman, "Alternative methods for dealing by nonnormality and heteroscedasticity in paleontological data," *Journal of Paleontology*, vol. 66, no. 6, pp. 857–867, Nov. 1992, doi: 10.1017/s0022336000020989.

[17]. T. J. Henderson, "Quantitative NMR spectroscopy using coaxial inserts containing a reference standard: purity determinations for military nerve agents," *Analytical Chemistry*, vol. 74, no. 1, pp. 191–198, Dec. 2001, doi: 10.1021/ac010809.

[18]. W. H. Woodall, R. Koudelik, K. L. Tsui, S. B. Kim, Z. G. Stoumbos, and C. P. Carvounis, "A review and analysis of the Mahalanobis—Taguchi system," *Technometrics*, vol. 45, no. 1, pp. 1–15, Feb. 2003, doi: 10.1198/004017002188618626.

[19]. A. Martino, A. Ghiglietti, F. Ieva, and A. M. Paganoni, "A k-means procedure based on a Mahalanobis type distance for clustering multivariate functional data," *Statistical Methods & Applications*, vol. 28, no. 2, pp. 301–322, Nov. 2018, doi: 10.1007/s10260-018-00446-6.

[20]. D. Steinley and M. J. Brusco, "Selection of variables in cluster analysis: An empirical comparison of eight procedures," *Psychometrika*, vol. 73, no. 1, pp. 125–144, Aug. 2007, doi: 10.1007/s11336-007-9019-y.

[21]. D. Charalampidis, "A modified k-means algorithm for circular invariant clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1856–1865, Dec. 2005, doi: 10.1109/tpami.2005.230.

[22]. A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and H. Jia, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[23]. D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, Nov. 2004, doi: 10.1109/tkde.2004.68.

[24]. I. S. Aref, J. Kadum, and A. Kadum, "Optimization of Max-Min and Min-Min Task Scheduling Algorithms Using G.A in Cloud Computing," *2022 5th International Conference on Engineering Technology and Its Applications (IICETA)*, May 2022, doi: 10.1109/iiceta54559.2022.9888542.

[25]. H. F. Pardede and K. Sairyo, "Generalized-log spectral mean normalization for speech recognition," *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, Aug. 2011, doi: 10.21437/interspeech.2011-209.

[26]. M. Mazziotta and A. Pareto, "Normalization methods for spatio-temporal analysis of environmental performance: Revisiting the Min–Max method," *Environmetrics*, vol. 33, no. 5, May 2022, doi: 10.1002/env.2730.

[27]. G. J. Brouwer and D. J. Heeger, "Categorical clustering of the neural representation of color," *The Journal of Neuroscience*, vol. 33, no. 39, pp. 15454–15465, Sep. 2013, doi: 10.1523/jneurosci.2472-13.2013.

[28]. D. Chicco and G. Jurman, "A statistical comparison between Matthews correlation coefficient (MCC), prevalence threshold, and Fowlkes–Mallows index," *Journal of Biomedical Informatics*, vol. 144, p. 104426, Aug. 2023, doi: 10.1016/j.jbi.2023.104426.

[29]. R. J. G. B. Campello, "A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment," *Pattern Recognition Letters*, vol. 28, no. 7, pp. 833–841, May 2007, doi: 10.1016/j.patrec.2006.11.010.

[30]. J. R. Oliver, "The Archaeology of Agriculture in Ancient Amazonia," in *Springer eBooks*, 2008, pp. 185–216. doi: 10.1007/978-0-387-74907-5_12.

[31]. J. Morales, A. R. Rodríguez, V. Alberto, C. Machado, and C. C. Hernández, "The impact of human activities on the natural environment of the Canary Islands (Spain) during the pre-Hispanic stage (3rd–2nd Century BC to 15th Century AD): an overview," *Environmental Archaeology*, vol. 14, no. 1, pp. 27–36, Apr. 2009, doi: 10.1179/174963109x400655.

[32]. L. Lobao and K. Meyer, "The Great Agricultural Transition: Crisis, change, and social consequences of twentieth century US farming," *Annual Review of Sociology*, vol. 27, no. 1, pp. 103–124, Aug. 2001, doi: 10.1146/annurev.soc.27.1.103.

[33]. C. Cartwright, "The principles, procedures and pitfalls in identifying archaeological and historical wood samples," *Annals of Botany*, vol. 116, no. 1, pp. 1–13, May 2015, doi: 10.1093/aob/mcv056.

[34]. A. L. Price, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, Jul. 2006, doi: 10.1038/ng1847.

[35]. D. M. Nemeskey, "Natural language processing methods for language modeling," 2023. doi: 10.15476/elte.2020.066.

[36]. A. F. L. Nemec and R. O. Brinkhurst, "The Fowlkes–Mallows statistic and the comparison of two independently determined dendrograms," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 45, no. 6, pp. 971–975, Jun. 1988, doi: 10.1139/f88-119.

[37]. R. Hansen *et al.*, "Adaptations to the current ECCO/ESPGHAN guidelines on the management of paediatric acute severe colitis in the context of the COVID-19 pandemic: a RAND appropriateness panel," *Gut*, vol. 70, no. 6, pp. 1044–1052, Sep. 2020, doi: 10.1136/gutjnl-2020-322449.

[38]. J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586–600, May 2000, doi: 10.1109/72.846731.