Scientific Research Publishing

International Journal of Geosciences, 2023, 14, 877-894
https://www.scirp.org/journal/ijg
ISSN Online: 2156-8367
ISSN Print: 2156-8359

A Review on Clustering Methods for Climatology Analysis and Its Application over South America

Luana Albertani Pampuch[1*], Rogério Galante Negri[1], Paul C. Loikith[2], Cassiano Antonio Bortolozo[1,3]

[1]Environmental Engineering Department, Institute of Science and Technology, São Paulo State University, São José dos Campos, Brazil

[2]Department of Geography, Portland State University, Portland, Oregon

[3]Cemaden-National Center for Monitoring and Early Warning of Natural Disasters, General Coordination of Research and Development, São José dos Campos, Brazil

Email: *luana.pampuch@unesp.br

How to cite this paper: Pampuch, L.A., Negri, R.G., Loikith, P.C. and Bortolozo, C.A. (2023) A Review on Clustering Methods for Climatology Analysis and Its Application over South America. *International Journal of Geosciences*, **14**, 877-894.
https://doi.org/10.4236/ijg.2023.149047

Received: August 13, 2023
Accepted: September 24, 2023
Published: September 27, 2023

Abstract

South America's climatic diversity is a product of its vast geographical expanse, encompassing tropical to subtropical latitudes. The variations in precipitation and temperature across the region stem from the influence of distinct atmospheric systems. While some studies have characterized the prevailing systems over South America, they often lacked the utilization of statistical techniques for homogenization. On the other hand, other research has employed multivariate statistical methods to identify homogeneous regions regarding temperature and precipitation, but their focus has been limited to specific areas, such as the south, southeast, and northeast. Surprisingly, there is a lack of work that compares various multivariate statistical techniques to determine homogeneous regions across the entirety of South America concerning temperature and precipitation. This paper aims to address this gap by comparing three such techniques: Cluster Analysis (K-means and Ward) and Self Organizing Maps, using data from different sources for temperature (ERA5, ERA5-Land, and CRU) and precipitation (ERA5, ERA5-Land, and CPC). Spatial patterns and time series were generated for each region over the period 1981-2010. The results from this analysis of spatially homogeneous regions concerning temperature and precipitation have the potential to significantly benefit climate analysis and forecasts. Moreover, they can offer valuable insights for various climatological studies, guiding decision-making processes in diverse fields that rely on climate information, such as agriculture, disaster management, and water resources planning.

Keywords

Climatology, Clustering Methods, Clustering Regionalization, Reanalysis Data, South America

1. Introduction

In various climate and meteorological studies, it is often essential to categorize data (observations or variables) into distinct subgroups containing elements that share similar characteristics. For instance, this separation can be used to create spatially homogeneous observations, either by utilizing weather stations or gridded data, and considering different variables like temperature or precipitation. This allows for climate regionalization, aiding in understanding regional climate patterns. Additionally, data grouping based on different temporal scales, such as hours, days, or months, and various parameters, can be employed to identify patterns, such as synoptic types. This facilitates the analysis of meteorological events and their underlying dynamics. Furthermore, for forecast analysis and evaluation, data grouping plays a crucial role. It enables the grouping of ensemble members, which can help in assessing the uncertainty and performance of forecasting models [1] [2] [3].

Cluster Analysis is a versatile technique that facilitates various types of studies. Its roots can be traced back to Tryon's proposal in 1939 [4], where it was primarily applied in biological taxonomy. However, it gained significant attention in the 1960s with the advent of faster computers, and by the 1970s, it began finding applications in diverse fields such as biology, sociology, and medicine [1]. Over the years, with the advancements in high-speed computers and data science, Cluster Analysis has evolved and is now classified as an unsupervised learning method [2]. This means that it can identify patterns and structures within data without the need for pre-labeled or labeled examples, making it a valuable tool in modern data analysis and exploration.

Several studies have been conducted over South America to classify homogeneous regions based on climate variables, employing various clustering methods. These investigations are of significant importance, given the continent's vast territory and diverse climate patterns. The outcomes of such studies find valuable applications in multiple domains, including agriculture, natural disaster management, understanding climate impacts, water resources management, and climate and weather forecasting [5].

In the study [6] gauge stations were used over Brazil to classify precipitation in six homogenenous regions using Ward Method. In [7] was performed a revision of precipitation regimes over South America and classify in eight homogeneous regions using a subjective analysis (graphics climatology analysis of meteorological stations data). [5] used a multivariate technique based on fuzzy theory to identify nine climate profiles (Grade of Membership) over Brazil using precipitation, relative humidity and maximum and minimum temperature from 1980-2013. [8] using K-Means Clustering Method and monthly precipitation data from ECMWF-SEAS5 and CPC for the period 1993-2016 found eight homogeneous regions. In addition to general studies across South America, specific research focusing on particular regions has also been conducted. For instance, [9] utilized clustering analyses to investigate the distributions of anomalies of sea surface temperature (SST) and moisture sources in the South Atlantic Ocean during extreme dry events in southeastern Brazil throughout the austral autumn, winter, and spring. This targeted approach allows for a deeper understanding of the factors influencing such extreme events in the specified region, shedding light on the complex interactions between SST and moisture source patterns during these critical periods. The methodology of Cluster Analysis has also been employed at a regional scale in South America in numerous studies, primarily oriented towards the demarcation of homogeneous zones with respect to rainfall and temperature patterns [10] [11] [12]. Notably, investigations have extended to the utilization of such data in the context of grain production [13]. The versatility inherent to Cluster Analysis methodology has facilitated its application in diverse contexts. Notably, it has been employed for the temporal characterization of temperature variability [14], as well as for the delineation of large-scale meteorological patterns within the South American region [15].

But these studies used only one preview choiced method. A comparison between methods was performed from few studies for other regions. [1] highlight the importance of the intercomparison of different clustering techniques using geophysical data in comparison with synthetic data. The knowledge of the method skill it is only possible with the application on real data. [16] compared four hierarquical methods using tropical rainfall stations, showing that there is no significant difference between methods performance. [1] found that nonhierarquical methods outperformed hierarquical for central-eastern North America. [3] found that K-Means clustering method produced stable cluster boundaries compared to other methods for Ethiopia precipitation. [17] regionalize annual precipitation for Iran using K-Means and Self Organizing Maps methods and show that K-means has better performance (using Silhouette Coefficient, Dunn index and Davis Bouldin index). [18] found that K-Means presents better results (using Calinski-Harabasz and Davies-Bouldin measures) than Ward and Self Organizing Maps methods for clustering precipitation over southeastern Brazil.

Regarding the entire South America continent there was no founded studies using Self Organizing Maps (SOM) for clustering regionalization. Furthermore, there are no studies comparing different methodologies for clustering regionalization in climatology. Facing this lack of investigation, the objective of this study is to perform a review of different methodologies for clustering (Ward, K-means and SOM) and the metrics for its evaluation (Silhouette value, Calinski-Harabasz index, Davies-Bouldin Index, Elbow Method and Modified Elbow Method). This study also presents a comparison between different reanalysis temperature (ERA5, ERA5-Land and CRU) and precipitation data (ERA5, ERA5-Land and CPC). An application of cluster methods and the metrics for its evaluation for precipitation and temperature were performed over South America only for ERA5-Land data to create spatial homogeneous groups of these climate variables.

2. Clustering Methods

According to [19], clusters are regions in the attribute space that contain a high density of patterns and are separated from each other by regions of low density. The central idea of clustering methods is to divide a data set into groups (clusters) such that similar elements are found into each group at the same time these elements show a distinct behavior concerning elements found into other groups. The literature on cluster analysis is very extensive, and its applications span from signal processing to psychology, archaeology, and linguistics [20].

We can express a clustering method as function g: $I \to G$, where $I \subseteq X$ is a set of m observed examples/objects $x = [x_1, \cdots, x_n]$ defined on the attribute space X, and $G = \{G_1, \cdots, G_c\}$ is a partition of I into c subsets. It is of utmost importance to emphasize that there is no prior knowledge about the labels of the examples contained in I. According to a general overview presented by [21], clustering methods can be categorized into: hierarchical methods, based on cost function optimization (non-hierarchical), and others, including neural network-based methods. In the following sections are discussed remarkable clustering methods according these categories.

2.1. Hierarchical Methods

Hierarchical methods are commonly used to synthesize the organizational structure of how the elements are related to each other. A representation based on a dendrogram, exemplified in **Figure 1**, supports the mentioned structure understanding. A dendrogram is a diagram that shows the hierarchical relationship between objects, with its main use is to work out the best way to allocate objects to clusters [2]. Through dissimilarity values, the existence of subdivisions becomes evident with respect to a given threshold τ_f. These subdivisions naturally determine the configuration of the clusters. In the dendrogram in **Figure 1**, the height of the dendrogram indicates the order in which the clusters were joined. In **Figure 1**, we can see that x_4 and x_5 are most similar, as the height of the link that joins them together is the smallest. The next two most similar objects are x_3 and x_6.

The different algorithms proposed in the literature lead to the construction of a hierarchical relationship among the data. Agglomerative hierarchical approaches derive this relationship through consecutive clustering steps on the dataset until a single cluster is obtained at the end. Conversely, a divisive hierarchical algorithm starts with a single cluster composed of all the data involved in the problem and undergoes successive subdivisions until clusters composed of a single example are obtained.

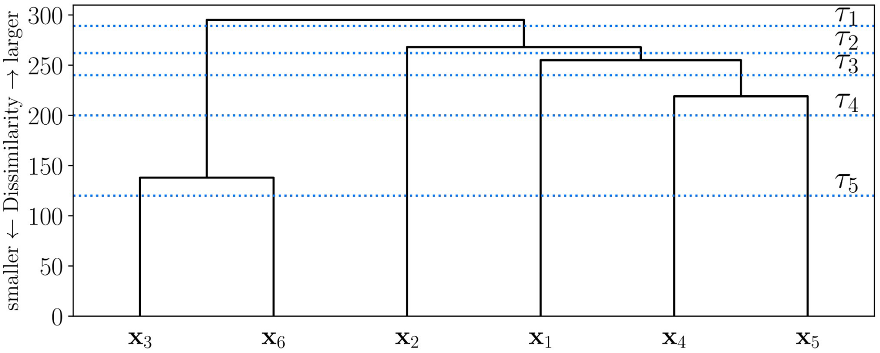

Figure 1. Example of dendrogram. The hierarchical structure depicted in terms of dissimilarity shows how the data is clustered.

Conveniently and in a generic way, we denote by $d(x_i, x_k)$ the dissimilarity between the objects x_i and x_k, where $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+^*$. According to this measure, $d(x_i, x_k) \to \infty$ indicates a higher dissimilarity between input patterns, and conversely, greater similarity is observed as $d(x_i, x_k) \to 0$. Furthermore, we can denote $D(G_j, G_\ell)$ as the observed dissimilarity between clusters G_j and G_ℓ.

Based on the above-presented concepts of dissimilarity, the hierarchical method of Ward clusters the data ensuring the minimum internal variability within the clusters by adopting the dissimilarity measure defined in Equation (1). According to this measure, the dissimilarity between G_ℓ and a given cluster, resulting from grouping G_j and G_k, is recursively computed. Initially, each element in the dataset defines a cluster and, in this case, the dissimilarity $D(G_i, G_j) = d(\mathbf{x}_i, \mathbf{x}_j)$ stands for the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

$$\begin{aligned} D(G_j \cup G_k, G_\ell) = & \frac{\#G_j + \#G_\ell}{\#G_j + \#G_k + \#G_\ell} D(G_j, G_\ell) \\ & + \frac{\#G_k + \#G_\ell}{\#G_j + \#G_k + \#G_\ell} D(G_k, G_\ell) \\ & - \frac{\#G_\ell}{\#G_j + \#G_k + \#G_\ell} D(G_j, G_k) \end{aligned} \tag{1}$$

2.2. K-Means

Clustering methods based on "function optimization" persist in defining a partition for the dataset such that the internal variability of the clusters is minimized while the separation between clusters is maximized. The K-Means algorithm is an extensively known algorithm that is based on such concept [22]. Aiming to achieve the objective of partition a given dataset I into k clusters, the following objective function should be minimized:

$$\min_{\mu_1, j=1,\cdots,k} \frac{1}{m} \sum_{j=1}^{k} \sum_{x_i \in \mathcal{G}_j} \left\| x_i - \mu_j \right\|^2 \tag{2}$$

Two main and straightforward steps characterize this algorithm: 1) assigning elements to clusters based on the smallest dissimilarity, expressed in terms of

Euclidean distance, between a given pattern and the mean vector of the cluster, represented by the cluster's centroids μ_j, $j = 1, \cdots, k$; 2) update the centroid vector that represents each cluster according to the average vector computed through the elements assigned in the previous step.

These two described steps are performed iteratively until convergence is reached. Commonly adopted convergence criteria include the absence of changes in element assignments between clusters and/or no alteration in the internal variability of the clusters.

2.3. Self-Organizing Maps

Self-Organizing Maps (SOM) comprises a neural network-based model for data clustering. In this model, neurons are represented by topologically organized maps where the location/coordinate of these neurons expresses a specific feature of the input data [23].

Conveniently, a map of neurons is represented by a matrix $\mathbf{M}$ of size $L_1 \times L_2 \times n$, where L_1 and L_2 defines the neuron map dimensions and n stands for the dimension of the attribute space $\mathcal{X}$. For a given neuron inserted in this map, at coordinates (u, v), it is denoted by $w_{uv} = [w_{uv1}, \cdots, w_{uvn}]$ as the associated weight vector. Thus, for a given object $\mathbf{x} = [x_1, \cdots, x_n]$, it is possible to assess its similarity to each neuron in the network and make adjustments to their respective associated weights when necessary. This relationship is summarized in **Figure 2**, where the attributes of an object are compared to each neuron in this network through a weight associated with the neuron.

During the execution of such a neural network, three main processes are involved: competition, cooperation, and adaptation. The competitive process consists of determining the neuron in the network that has the minimum dissimilarity to the presented object. The neuron selection according to the minimum object-neuron dissimilarity is expressed by:

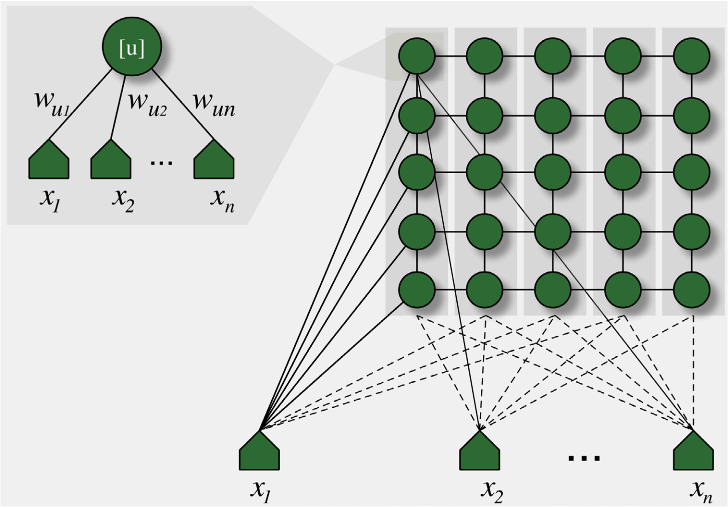

Figure 2. The SOM architecture. Each input data is compared to each neuron according the respective components and weights.

$$[uv] = \arg \min_{\substack{u=1,\cdots,L_1 \\ v=1,\cdots,L_2}} \left\| x_i - \mu_j \right\| \tag{3}$$

Once the neuron at coordinates (*u*, *v*) demonstrates the highest similarity (*i.e.*, lowest dissimilarity) to the pattern *x*, corrections must be made to all other neurons in the network based on the configuration of the identified neuron and the presented pattern. Such corrections are conducted to benefit the neurons located in the neighborhood of (*u*, *v*), thus characterizing a cooperative process.

For this purpose, "topological neighborhood functions" are used. Among different proposals in the literature, the Gaussian function $V(\mathbf{a},\mathbf{b};\sigma) = \mathrm{e}^{-\frac{\|\mathbf{a}-\mathbf{b}\|^2}{2\sigma^2}}$ is widely used for this purpose, where $\mathbf{a},\mathbf{b} \in \mathbb{R}^2$ represent a spatial coordinate pair, and $\sigma \in \mathbb{R}_+^*$ controls the range of the neighborhood.

Lastly, the adaptive process is responsible for adjusting the weights of the neurons as patterns are presented to the neural network. Assuming that the neuron at coordinates (*u*, *v*) was identified during the competitive process, and $(i,j) \in L_1 \times L_2$ represents the coordinates of the neurons in $\mathbf{M}$, the adaptive process is defined as:

$$\mathbf{w}_{ij} := \mathbf{w}_{ij} + \eta V\left((u,v),(i,j),\sigma\right) \cdot \left(\mathbf{x} - \mathbf{w}_{ij}\right) \tag{4}$$

where $\eta \in \mathbb{R}_+$ represents a learning rate.

In general, the training process is iteratively executed until convergence is achieved in the weight adjustment process. Once convergence is detected, the final configuration of the neurons, with their adjusted weights, provides a flat representation of the analyzed object features and, consequently, allows groups the data assigned to specific regions of the neuron map.

3. Cluster Evaluation and Number of Clusters

The cluster evaluation can be performed using some index that allows the comparison between the methods for different number of groups. Silhouette Value [24], Calinski-Harabasz [25] and Davies-Bouldin [26] are some of these methods for clustering assessment. For Silhouette Values (SL) each cluster is represented by a silhouette, and is a comparison of its tightness and separation (how similar an element is to other in the same cluster, compared to points of other clusters) [24]. The Silhouette values are calculated by:

$$SL_i(k) = 1 - \frac{d_{ik}}{\delta_{i,-k}} \tag{5}$$

where d_{ik} is the average intra-cluster distance between station *i* and all other stations associated with medoid *k* and $\delta_{i,-k}$ is the smallest average distance between station *i* and all other stations associated with a medoid different from *k*. The values of $SL_i(k)$ are in interval $[-1,1]$, and $s_i(k) \approx 1$ indicates better results (intra-cluster distance is much smaller than the inter-cluster distance).

Calinski-Harabasz index (*CH*) also called as Variance Ratio Criterion is calculated by:

$$CH_k = \frac{SS_B}{SS_w} \times \frac{N-k}{k-1} \tag{6}$$

where:

$$SS_B = \sum_{i=1}^{k} n_i \left\| m_i - m \right\|^2 \tag{7}$$

is the overall between-cluster variance. And the overall within-cluster variance is:

$$SS_w = \sum_{i=1}^{k} \sum_{x \in c_i} \left\| x - m_i \right\|^2 \tag{8}$$

with n_i being the number of observations in cluster *i*, m_i the centroid of cluster *i*, *m* the overall mean of the sample data, *x* a data point, c_i the *i*th cluster, *k* the number of clusters, and *N* the number of observations. The values of CH_k are in interval $[0,\infty)$, and highest CH_k indicates better data partition, large between-cluster variance SS_B and a small within-cluster variance SS_w [25].

The Davies-Bouldin criterion (*DB*) is based on a ratio of within-cluster and between-cluster distances, defined by:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left\{ D_{i,j} \right\} \tag{9}$$

where $D_{i,j}$ is the within-to-between cluster distance ratio for the *i*th and *j*th clusters. The values of *DB* are in interval $[0,\infty)$, and lowest *DB* indicates optimal clustering solution [26].

The choice of the number of clusters is a key point on cluster analysis and although all analysis performed on methods been objectives, the determination of the number of groups presents some subjectivity [2] [3].

For hierarquical clustering methods, the choice of *k* can be performed with a traditional subjective approach inspecting the plot of the distances between merged clusters as a function of the stage of the analysis. The stage where the difference between the distances is bigger (a big jump between points occurs) indicates that these elements are not so closed, and the process can be stopped just before these distances become large [2].

Another approach to determine the number of clusters is by utilizing objective methods. The evaluation techniques introduced for assessing cluster methods can also be applied to make the selection of "*k*" (as presented in **Table 1**).

Elbow Method can also be used to the choice of the number of clusters for nonhierarquical methods. It consists in a graph analysis of within-cluster sum of

Table 1. Objective methods to choose the number of cluster and evaluation.

METHOD	ABBREVIATION	INTERVAL VALUES	BEST VALUES FOR *k* CHOICE	REFERENCE
Silhouette	SL values	[−1, 1]	Higher values	[24]
Calinski-Harabasz	CH index	[0, ∞)	Higher values	[25]
Davies-Bouldin	DB index	[0, ∞)	Lowest values	[26]

square errors (WSS) for different k values (searching for an elbow), defined by [3]:

$$WSS = \sum_{j}^{k} \sum_{g \in j} \left(t_g - \overline{t}_j \right)^2 \tag{10}$$

where WSS is the sum of the squared errors between the time series in each grid cell g (t_g) in cluster j ($g \in j$) and the average time series in cluster j ($\overline{t}_j$ is the centroid) and then summed over all k clusters.

Although, Elbow method may be problematic mainly when there are a large number of elements to be clustered in a small number of groups (graph is smoothed and an elbow is not clear), as the case of a large number of grid points to found climate patterns [3]. In this sense, [3] proposed the analysis of a modified Elbow Method based on the analysis of the differences between $WSS_{(k-1)}$ and $WSS_{(k)}$, that present more apparent elbow in these cases.

4. Cluster Analysis Application South America

We utilized three distinct sources of monthly reanalysis data for precipitation and temperature: ERA5 [27] and ERA5-Land [28] for both temperature and precipitation, CRU [29] exclusively for temperature, and CPC [30] solely for precipitation. Detailed descriptions of each data source are in **Table 2**. To facilitate intercomparison, all the data were interpolated onto a common grid with a resolution of $0.5^\circ \times 0.5^\circ$.

The seasonal temperature patterns in South America are well captured by the three reanalyses. ERA5 generally exhibits higher temperatures compared to ERA5-Land across most of South America, except for specific areas like northeastern Argentina, central Brazil, and the far north of the continent. These differences are generally within 2˚C (both positive and negative). Similarly, when comparing CRU with ERA5 and ERA5-Land, they display similar temperature patterns. CRU tends to be warmer in the majority of the continent, except

Table 2. Reanalysis data used to comparison for temperature and precipitation for South America.

DATA	ESPACIAL RESOLUTION	CENTER	REFERENCE
ERA5 (precipitation and temperature)	$0.25^\circ \times 0.25^\circ$	European Center for Medium-Range Weather Forecast (ECMWF)	[27]
ERA5 – Land (precipitation and temperature)	$0.1^\circ \times 0.1^\circ$	European Center for Medium-Range Weather Forecast (ECMWF)	[28]
CRU (temperature)	$0.5^\circ \times 0.5^\circ$	National Center for Atmospheric Research (NCAR)	[29]
CPC (precipitation)	$0.5^\circ \times 0.5^\circ$	National Oceanic and Atmospheric Administration (NOAA)	[30]

for regions like Southern Argentina and the Andes. In winter, ERA5 and ERA5-Land appear warmer in the center and northeast of Brazil, with differences usually within 4˚C (both positive and negative).

Concerning precipitation patterns, the three reanalyses provide a reliable representation, reflecting the seasonality that is largely influenced by atmospheric systems in the region and the impact of sea surface temperature anomalies in the Atlantic and Pacific Oceans [7]. Both ERA5 and ERA5-Land exhibit a comparable precipitation pattern, with only minor punctual discrepancies possibly attributed to spatial resolution and differences in surface representation between the two datasets. On the other hand, when comparing CPC with ERA5 and ERA5-Land, substantial differences arise for some regions (up to 100 mm/month), particularly accentuated during the summer and in the northern part of the continent.

Given the superior representation of ERA5-Land (regridded) over the continent, it was chosen for conducting the cluster analysis in South America. A summary of all the steps involved in the cluster analysis and evaluation is depicted in **Figure 3**. Notably, as there is a high number of grid points (6148), dendrograms for hierarchical methods will not be presented.

5. Results

Within this section, we present the outcomes acquired through the methodologies delineated earlier, and proceed to analyze the diverse models thus identified.

5.1. Temperature Clustering Over South America

Figure 4 depicts the Cluster analysis applied to temperature patterns across the South American region. This analysis incorporates the employment of SL Values, CH Index, and DB Index, encompassing clustering scenarios ranging from k = 1 to 50. The dashed lines within the figure represent the methodological averages of Ward (depicted in blue), K-means (represented in red), and SOM (illustrated in green). It is noteworthy that the most favorable outcomes are observed in the case of K-means clustering, wherein higher SL values and CH index averages are evident, coupled with the attainment of the lowest DB index average.

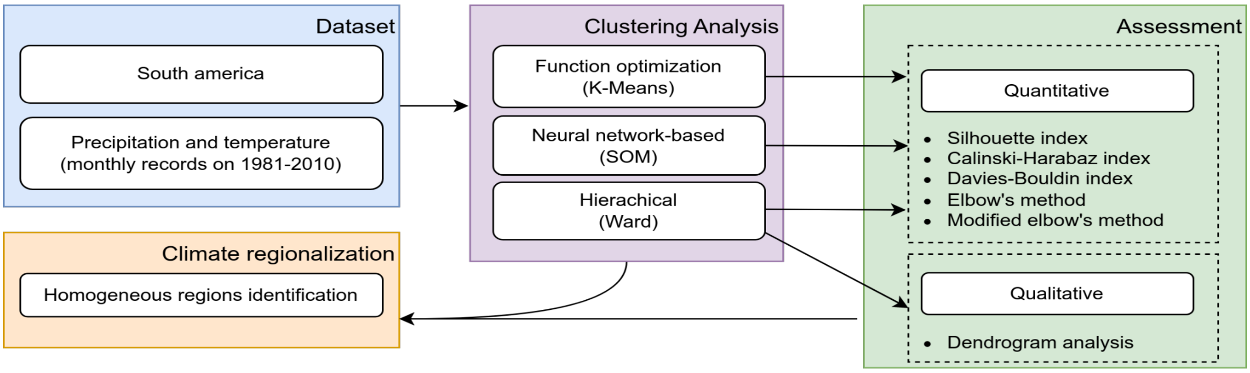

Figure 3. Summary of all steps for cluster analysis and evaluation.

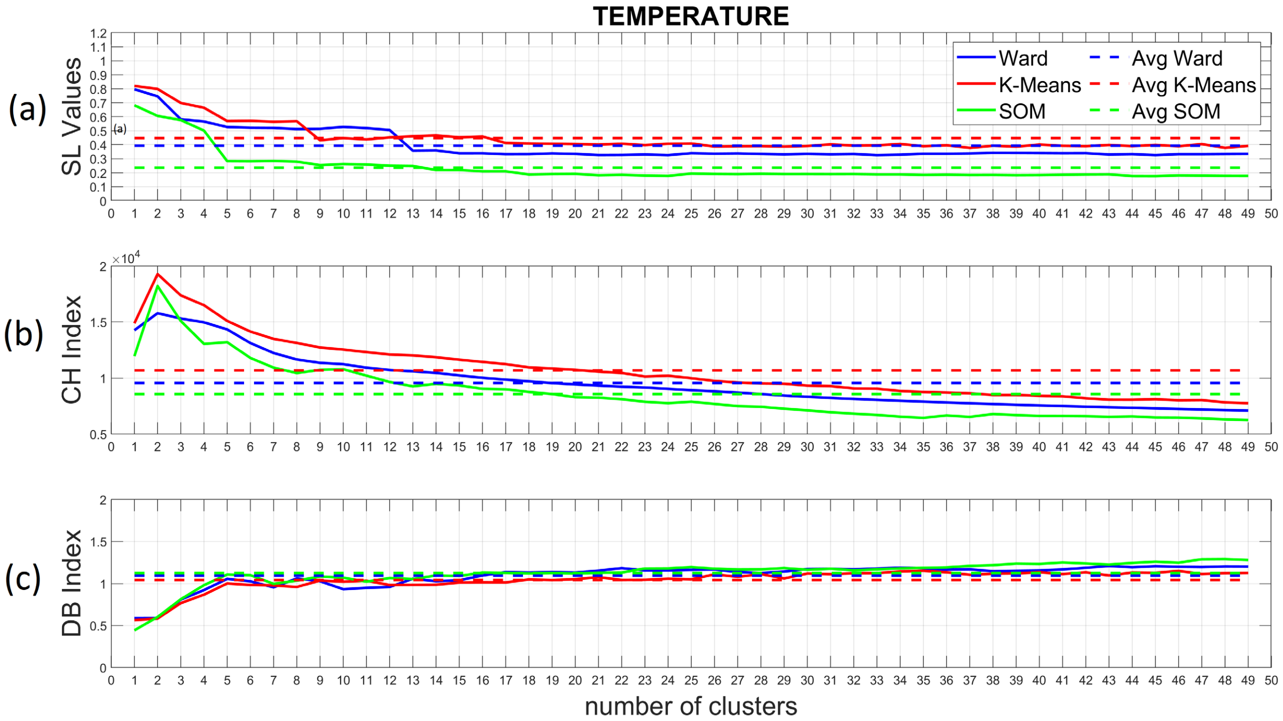

Figure 4. Clustering assessment for temperature over South America based on (a) Silhouette Value; (b) CH Index and (c) DB Index and number of clusters.

The application of these metrics also lends itself to the determination of the optimal cluster count. In the case of the Ward Method, an optimal selection for the number of temperature clusters emerges at $k = 10$. At this point, the SL values and CH index ascend to their zenith prior to encountering a decline (occurring at $k = 11$), concomitantly with a decline in the DB index, which attains its nadir. Meanwhile, in the context of K-means clustering, an appropriate choice for the number of temperature clusters is discerned at $k = 8$. Here, the SL values and CH index exhibit a peak before undergoing a precipitous descent (commencing at $k = 9$), while the DB index experiences a descent followed by an ascent (at $k = 9$). Conversely, in the case of SOM, the metrics do not converge on a unanimous optimal value for k. Specifically, the SL values indicate $k = 8$, the CH index suggests $k = 10$ (with both metrics showcasing an ascending pattern before declining at this point), and the DB index demonstrates its lowest value at $k = 7$. In light of these considerations, a prudent selection could be made at $k = 8$.

Regarding the non-hierarchical K-means method, supplementary approaches such as the Elbow Method and the Changed Elbow method can be employed. Nevertheless, discerning the optimal number of clusters remains a formidable task within this framework. Figure 5 elucidates an inflection point observable at $k = 5$ and $k = 6$. However, given the dimension of the South American region, these cluster counts prove to be relatively modest, inadequately addressing the multifaceted nature of temperature variations across the region. Emphasizing the imperative consideration that the mathematical outcome must align with

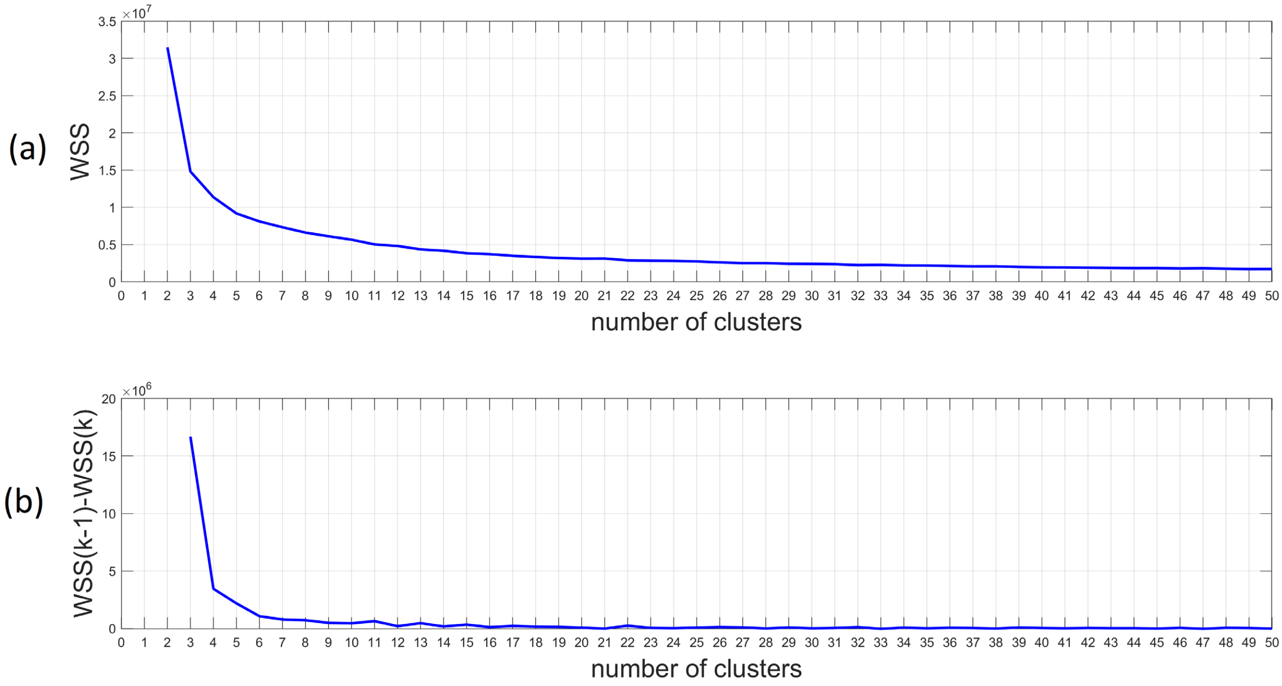

Figure 5. Metrics to the choice of the number of clusters for temperature over South America with K-Means method (a) Traditional Elbow method (WSS) and (b) Changed Elbow Method ($WSS_{(k-1)} - WSS_{(k)}$).

physical significance, it is crucial to underscore that the determination of the cluster count, even when facilitated by metrics, entails a degree of subjectivity.

Considering the best choice for each method spatial maps were made to verify the temperature spatial distribution of the groups (Figure 6). It can be observed that the methods yield models characterized by relatively distinct group distributions, a result that aligns with anticipated expectations. However, certain clusters exhibit a degree of proximity across all three methods. Noteworthy examples include the clusters encompassing the Amazon basin and a portion of northeastern Brazil. Another instance pertains to the northern coastal region of Argentina, extending inland to approximate proximity with the Chilean border and southward until nearly reaching a latitude of −45˚. Furthermore, the Chilean coastline demonstrates a congruent cluster distribution across all three cases. Regrettably, no extant studies pertaining to temperature clustering analysis have been identified that could serve as comparative benchmarks against the findings delineated within this article.

5.2. Precipitation Clustering Over South America

The evaluation of clustering with regard to precipitation patterns across the South American expanse is depicted in Figure 7, leveraging the utilization of SL Values, CH Index, and DB Index for cluster counts spanning from $k = 1$ to 50. The dashed lines within the figure correspond to methodological averages associated with Ward (represented in blue), K-means (illustrated in red), and SOM (depicted in green). The most promising outcomes manifest within the domain

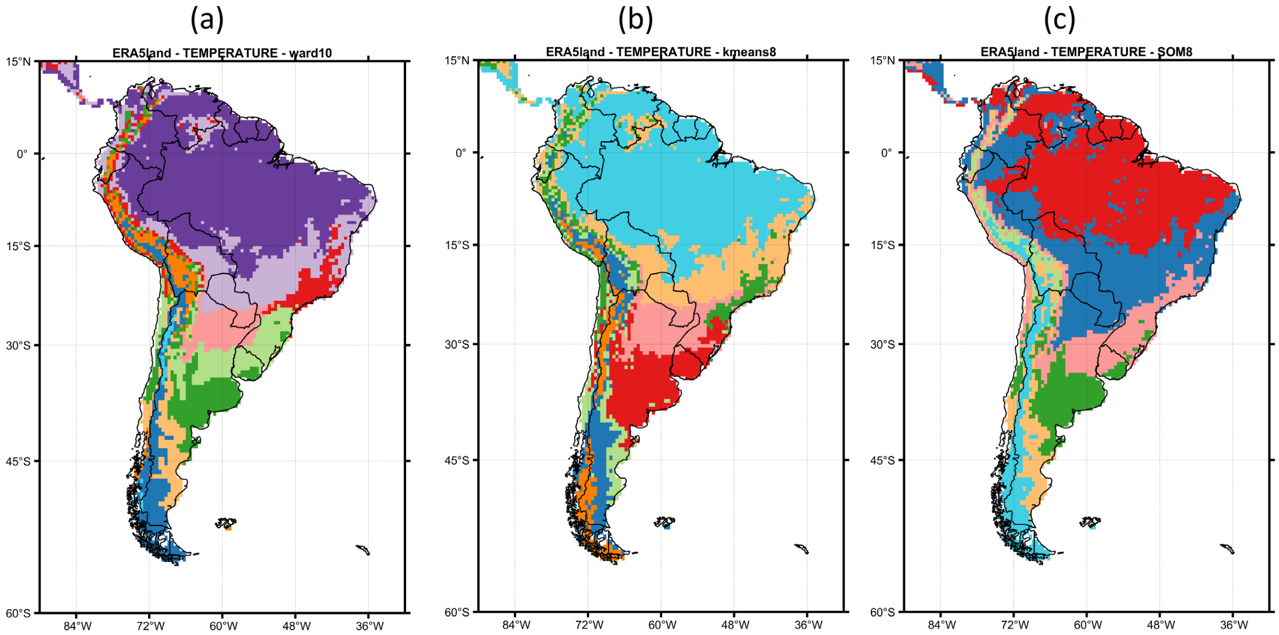
(a)
ERA5land - TEMPERATURE - ward10
(b)
ERA5land - TEMPERATURE - kmeans8
(c)
ERA5land - TEMPERATURE - SOM8
15°N
0°
15°S
30°S
45°S
60°S
84°W
72°W
60°W
48°W
36°W

Figure 6. Spatial results for cluster analysis over South America for temperature using (a) Ward ($k = 10$); (b) K-means ($k = 8$); (c) SOM ($k = 8$) methods. Where k is the number of groups.

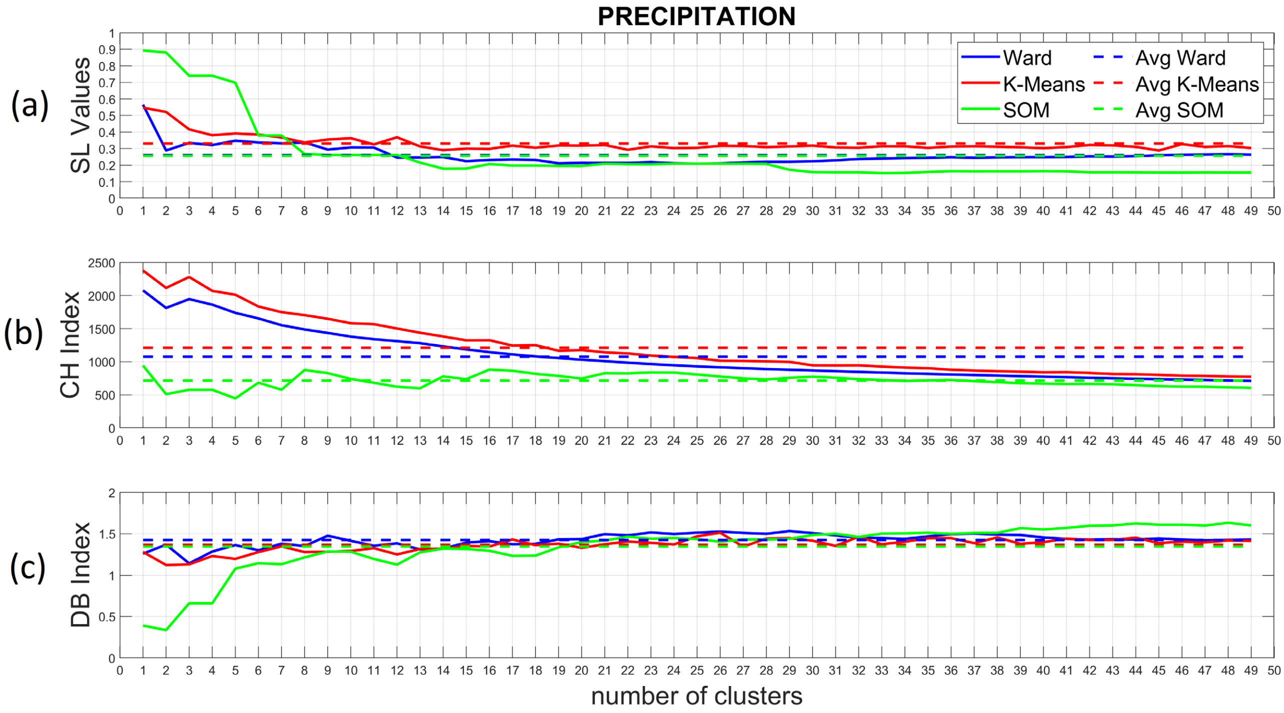

Figure 7. Clustering assessment for precipitation over South America based on (a) Silhouette Value; (b) CH Index and (c) DB Index and number of clusters.

of K-means clustering, as evident from the utilization of SL values and CH index, wherein elevated average values are attained. Conversely, the DB index reveals that SOM offers the most favorable performance, substantiated by the observation of lower average values, signifying enhanced effectiveness.

These metrics also serve a pivotal role in guiding the determination of an optimal cluster count. Within the framework of the Ward Method, an apt selection for the number of precipitation clusters emerges at $k = 11$, where both SL values and the CH index culminate in their peak values before undergoing a decline (manifesting at $k = 12$), concomitant with the DB index exhibiting a decline and attaining its nadir. In the context of K-means clustering, a judicious choice for the number of precipitation clusters is discerned at $k = 12$ for SL values and $k = 11$ for the CH index. In both cases, these values are characterized by an ascent to higher values prior to a subsequent decline. Furthermore, the DB index indicates an optimal selection at $k = 12$, corresponding to its lowest value. When considering SOM, congruence between SL values and DB index is observed at $k = 12$, while the CH index advocates for $k = 14$ as an optimal choice. Taking these factors into account, a judicious selection could indeed be made at $k = 12$.

In the context of the non-hierarchical K-means method, additional approaches such as the Elbow Method and the Changed Elbow method can be employed. However, in this instance, the task of determining the optimal cluster count is notably intricate due to the absence of a distinctly discernible inflection point (**Figure 8**).

The determination of the most suitable option for each methodology prompted the creation of spatial maps aimed at scrutinizing the geographical distribution of precipitation within the identified groups. Evidently, within the SOM framework, a pronounced large-scale cluster emerges (**Figure 9(c)**), accompanied by several smaller clusters located in the northern expanse of the continent. In the

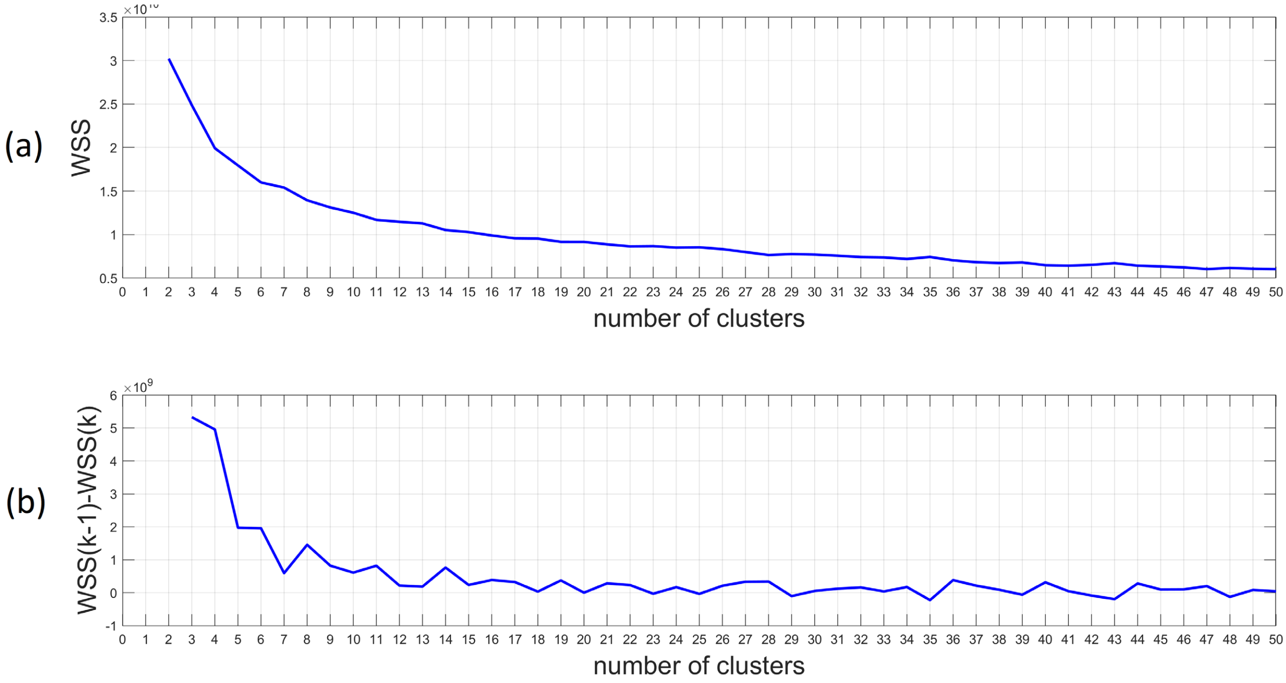

Figure 8. Metrics to the choice of the number of clusters for precipitation over South America with K-Means method (a) Traditional Elbow method (WSS) and (b) Changed Elbow Method ($WSS_{(k-1)} - WSS_{(k)}$).

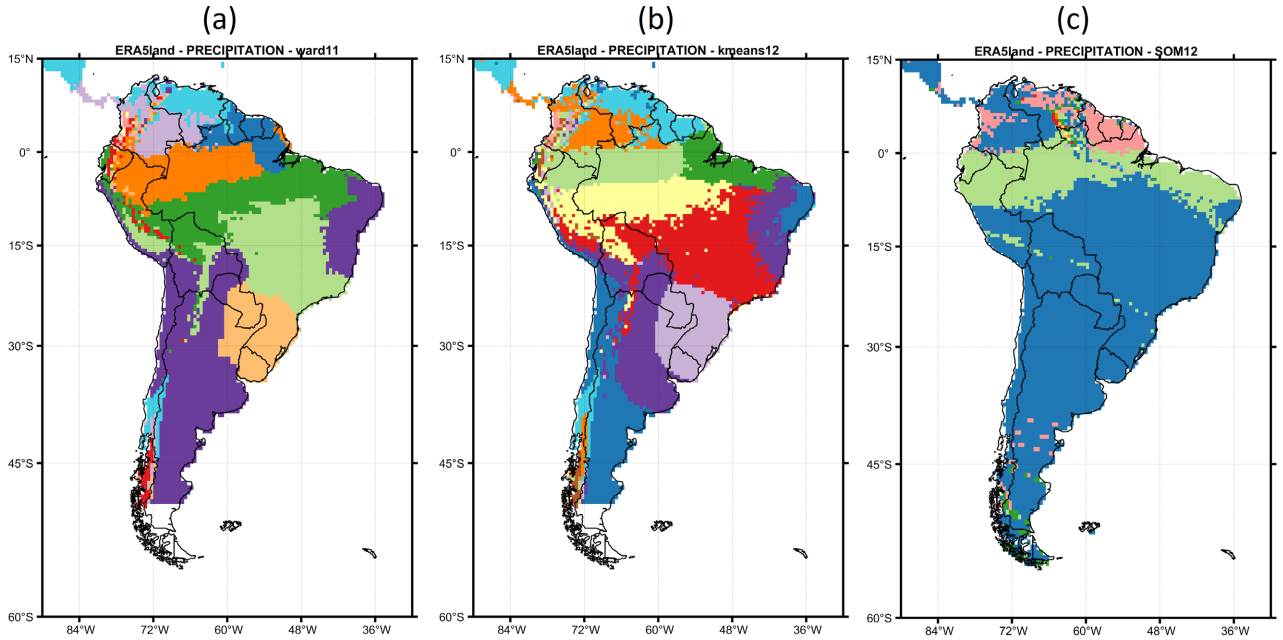
(a)
ERA5land - PRECIPITATION - ward11
(b)
ERA5land - PRECIPITATION - kmeans12
(c)
ERA5land - PRECIPITATION - SOM12
15°N
0°
15°S
30°S
45°S
60°S
84°W
72°W
60°W
48°W
36°W

Figure 9. Spatial results for cluster analysis over South America for precipitation using (a) Ward ($k = 11$); (b) K-means ($k = 12$); (c) SOM ($k = 12$) methods. Where k is the number of groups.

context of the Ward method (**Figure 9(a)**) and K-means approach (**Figure 9(b)**), disparities surface, primarily attributable to the variation in the number of clusters within the southwestern, northern, and northeastern sectors of the continent. Subsequent to attaining the optimum outcomes as deduced by the K-means approach, a thorough examination of the spatial distribution of precipitation patterns across South America was conducted. In light of the indices signifying the efficacy of K-means and SOM, it becomes evident that the spatial distribution evinced by the latter does not align with physical reality. Conversely, with regard to precipitation patterns, a notable inclination is observed in favor of the K-means methodology. Upon juxtaposing the findings of K-means ($k = 12$) with the study conducted by [8], which also employed the K-means technique albeit with a distinct dataset ($k = 8$), semblances in the distribution patterns across certain regions become apparent despite disparities in the number of clusters, an aspect that contributes to segmenting the groups. The salient distinctions stem from the omission of data pertaining to the western geographic sector of the continent.

6. Conclusions

The analysis of clustering holds a pivotal role in climatic studies, facilitating the discernment of intricate structures within climatological datasets. A notable observation pertains to the scarcity of comparative endeavors across distinct methodologies within the context of South America's climatic investigation. Addressing this gap, this study has introduced three distinct clustering methodologies—Ward, K-means, and Self-Organizing Maps (SOM)—while elucidating the processes encompassing method assessment, selection, and cluster count determination.

Silhouette Value, CH Index, and DB Index have emerged as indispensable tools for cluster validation and the judicious determination of cluster counts. It is imperative to acknowledge that conventional approaches, such as the Elbow Method and the Changed Elbow Method employed within non-hierarchical frameworks for cluster count selection, yield challenges that compound the decision-making process.

In relation to temperature patterns, K-means has showcased superior performance, leading to the formation of 8 distinct clusters across the South American expanse. Conversely, for precipitation, the most favorable outcomes have been achieved through K-means clustering, resulting in the identification of 12 distinct clusters across the same region. Consequently, K-means emerges as a robust method for the climatic regionalization of both temperature and precipitation patterns across the South American landscape.

In conclusion, this study not only serves as a valuable reference for the exploration of climatic clustering methodologies but also lays the foundation for future investigations focused on the continent's climatic intricacies.

Acknowledgements

Luana Albertani Pampuch thanks to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the Research Financial Support (Universal Project grant 426530/2018-7). Cassiano Bortolozo also thanks to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the Postdoctoral Scholarship (grant 152269/2022-3) and also for the Research Financial Support (Universal Project grant 433481/2018-8). Rogério Galante Negri thanks to the São Paulo Research Foundation (FAPESP, grant 2021/01305-6), and CNPq (grant 305220/2022-5).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Gong, X. and Richman, M.B. (1995) On the Application of Cluster Analysis to Growing Season Precipitation Data in North America East of the Rockies. *Journal of Climate*, **8**, 897-931. https://doi.org/10.1175/1520-0442(1995)008<0897:OTAOCA>2.0.CO;2

[2] Wilks, D.S. (2020) Statistical Methods in the Atmospheric Sciences. International Geophysics Series, 4th Edition, Elsevier, Amsterdam.

[3] Zhang, Z. and Li, J. (2020) Big Data Mining for Climate Change. Elsevier, Amsterdam.

[4] Tryon, R.C. (1939) Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers,

Ann Arbor.

[5] Marinho, K.F.S., Andrade, L.M.B., Spyrides, M.H.C., Silva, C.M.S., Oliveira, C.P., Bezerra, N.G. and Mutti, P.R. (2020) Climate Profiles in Brazilian Microregions. *Atmosphere*, **11**, Article No. 1217. https://doi.org/10.3390/atmos11111217

[6] Keller-Filho, T., Assad, E.D. and Lima, P.R.S.R. (2005) Regiões pluviométricas homogêneas no Brasil. *Pesquisa Agropecuaria Brasileira*, **40**, 311-322. https://doi.org/10.1590/S0100-204X2005000400001

[7] Reboita, M.S., Gan, M.A., da Rocha, R.P. and Ambrizzi, T. (2010) Regimes de precipitação na América do Sul: Uma revisão bibliográfica. *Revista Brasileira de Meteorologia*, **25**, 185-204. https://doi.org/10.1590/S0102-77862010000200004

[8] Ferreira, G.W.S. and Reboita, M.S. (2022) A New Look into the South America Precipitation Regimes: Observation and Forecast. *Atmosphere*, **13**, Article No. 873. https://doi.org/10.3390/atmos13060873

[9] Pampuch, L.A., Drumond, A., Gimeno, L. and Ambrizzi, T. (2016) Anomalous Patterns of SST and Moisture Sources in the South Atlantic Ocean Associated with Dry Events in Southeastern Brazil. *International Journal of Climatology*, **36**, 4913-4928. https://doi.org/10.1002/joc.4679

[10] Dourado, C.D.S., Oliveira, S.R.D.M. and Avila, A.M.H.D. (2013) Análise de zonas homogêneas em séries temporais de precipitação no Estado da Bahia. *Bragantia*, **72**, 192-198. https://doi.org/10.1590/S0006-87052013000200012

[11] Lyra, G.B., Oliveira-Júnior, J.F. and Zeri, M. (2014) Cluster Analysis Applied to the Spatial and Temporal Variability of Monthly Rainfall in Alagoas State, Northeast of Brazil. *International Journal of Climatology*, **34**, 3546-3558. https://doi.org/10.1002/joc.3926

[12] Souza, A.D., Abreu, M.C., de Oliveira-Júnior, J.F., Aristone, F., Fernandes, W.A., Aviv-Sharon, E. and Graf, R. (2022) Climate Regionalization in Mato Grosso do Sul: A Combination of Hierarchical and Non-Hierarchical Clustering Analyses Based on Precipitation and Temperature. *Brazilian Archives of Biology and Technology*, **65**, e22210331. https://doi.org/10.1590/1678-4324-2022210331

[13] Lopes, A.R., Marcolin, J., Johann, J.A., Boas, M.A.V. and Schuelter, A.R. (2019) Identification of Homogeneous Rainfall Zones during Grain Crops in Paraná, Brazil. *Engenharia Agrícola*, **39**, 707-714. https://doi.org/10.1590/1809-4430-eng.agric.v39n6p707-714/2019

[14] Detzer, J., Loikith, P.C., Pampuch, L.A., Mechoso, C.R., Barkhordarian, A. and Lee, H. (2019) Characterizing Monthly Temperature Variability States and Associated Meteorology across Southern South America. *International Journal of Climatology*, **40**, 492-508. https://doi.org/10.1002/joc.6224

[15] Loikith, P.C., Pampuch, L.A., Slinskey, E., Detzer, J., Mechoso, C.R. and Barkhordarian, A. (2019) A Climatology of Daily Synoptic Circulation Patterns and Associated Surface Meteorology over Southern South America. *Climate Dynamics*, **53**, 4019-4035. https://doi.org/10.1007/s00382-019-04768-3

[16] Jackson, I.J. and Weinand, H. (1995) Classification of Tropical Rainfall Stations: A Comparison of Clustering Techniques. *International Journal of Climatology*, **15**, 985-994. https://doi.org/10.1002/joc.3370150905

[17] Roushangar, K. and Alizadeh, F. (2018) A Multiscale Spatio-Temporal Framework to Regionalize Annual Precipitation Using k-Means and Self-Organizing Map Technique. *Journal of Mountain Science*, **15**, 1481-1497.

[18] Miranda, B.G., Negri, R.G. and Pampuch, L.A. (2023) Using Clustering Algorithms and GPM Data to Identify Spatial Precipitation Patterns over Southeastern Brazil.

Atmósfera, **37**, 365-381. https://doi.org/10.20937/ATM.53155

[19] Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011) Cluster Analysis. 5th Edition, Wiley Series in Probability and Statistics, Wiley, Hoboken. https://doi.org/10.1002/9780470977811

[20] Webb, A.R. and Copsey, K.D. (2011) Statistical Pattern Recognition. 3rd Edition, John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781119952954

[21] Theodoridis, S. and Koutroumbas, K. (2008) Pattern Recognition. 4th Edition, Academic Press, Cambridge.

[22] Lloyd, S. (1982) Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, **28**, 129-137. https://doi.org/10.1109/TIT.1982.1056489

[23] Kohonen, T. (2001) Self-Organizing Maps. Springer, Heidelberg. https://doi.org/10.1007/978-3-642-56927-2

[24] Rousseeuw, P.J. (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

[25] Calinski, T. and Harabasz, J. (1974) A Dendrite Method for Cluster Analysis. *Communications in Statistics*, **3**, 1-27. https://doi.org/10.1080/03610927408827101

[26] Davies, D.L. and Bouldin, D.W. (1979) A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-1**, 224-227. https://doi.org/10.1109/TPAMI.1979.4766909

[27] Hersbach, H. and Dee, D. (2016) ERA5 Reanalysis Is in Production. ECMWF Newsletter No. 147, 7.

[28] Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D.G., Piles, M., Rodríguez-Fernández, N.J., Zsoter, E., Buontempo, C. and Thépaut, J.-N. (2021) ERA5-Land: A State-of-the-Art Global Reanalysis Dataset for Land Applications. *Earth System Science Data*, **13**, 4349-4383. https://doi.org/10.5194/essd-13-4349-2021

[29] Harris, I., Osborn, T.J., Jones, P., *et al.* (2020) Version 4 of the CRU TS Monthly High-Resolution Gridded Multivariate Climate Dataset. *Scientific Data*, **7**, Article No. 109. https://doi.org/10.1038/s41597-020-0453-3

[30] Xie, P., Chen, M., Yang, S., Yatagai, A., Hayasaka, T., Fukushima, Y. and Liu, C. (2007) A Gauge-Based Analysis of Daily Precipitation over East Asia. *Journal of Hydrometeorology*, **8**, 607-626. https://doi.org/10.1175/JHM583.1