UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA HIDRÁULICA E AMBIENTAL
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA CIVIL

THAÍS ANTERO DE OLIVEIRA

APPLICATION OF CLUSTERING METHODS FOR HYDROLOGICAL REGIONALIZATION USING THE CAMELS-BR DATABASE

FORTALEZA
2023

THAÍS ANTERO DE OLIVEIRA

APPLICATION OF CLUSTERING METHODS FOR HYDROLOGICAL REGIONALIZATION USING THE CAMELS-BR DATABASE

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Civil da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Mestre em Engenharia Civil. Área de concentração: Recursos Hídricos.

Advisor: Prof. Francisco de Assis de Souza Filho, PhD.

FORTALEZA
2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O52a Oliveira, Thaís Antero de.
Application of Clustering Methods for Hydrological Regionalization using the CAMELS BR Database / Thaís Antero de Oliveira. – 2023.78 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Civil: Recursos Hídricos, Fortaleza, 2023.
Orientação: Prof. Dr. Francisco de Assis de Souza Filho.

1. Clusterização. 2. CAMELS. 3. classificação hidrológica. 4. multi-método. I. Título.
CDD 627

THAÍS ANTERO DE OLIVEIRA

APPLICATION OF CLUSTERING METHODS FOR HYDROLOGICAL REGIONALIZATION USING THE CAMELS-BR DATABASE

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Civil da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Mestre em Engenharia Civil. Área de concentração: Recursos Hídricos.

Approved in:

BANCA EXAMINADORA

____

Prof. Dr. Francisco de Assis de Souza Filho (Advisor)
Universidade Federal do Ceará (UFC)

____

Profa. Dra. Ticiana Marinho de Carvalho Studart
Universidade Federal do Ceará (UFC)

____

Dr. Renan Vieira Rocha
Fundação Cearense de Meteorologia e Recursos Hídrico (FUNCEME)

Aos meus pais, Ana e Elves.

AGRADECIMENTOS

Agradeço aos meus pais, Ana e Elves, por todo apoio e torcida nos caminhos que escolhi seguir. Também, ao Matheus por todo apoio e esforço em me fazer acreditar em mim mesma.

Agradeço ao professor Assis, por me fazer me sentir capaz ao acreditar tanto em mim, mais do que, muitas vezes, eu própria. Também, ao Victor, por toda parceria e ajuda que foram indispensáveis para fazer esta dissertação.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

“O Tejo é mais belo que o rio que corre pela minha aldeia,
Mas o Tejo não é mais belo que o rio que corre pela minha aldeia
Porque o Tejo não é o rio que corre pela minha aldeia.” (Fernando Pessoa).

RESUMO

A regionalização de parâmetros e de variáveis de bacias hidrográficas são cruciais para a previsão de vazão em bacias não monitoradas, parametrização de modelos e desenvolvimento e gestão de bacias. Para superar a limitação de quantidade reduzida de dados hidrológicos, foi produzido e disponibilizado publicamente o conjunto de dados Catchment Attributes and MEteorology for Large-sample Studies – Brazil (CAMELS-BR). A aplicação limitada de métodos de clusterização na análise de bacias hidrográficas no Brasil, especialmente utilizando o conjunto de dados CAMELS-BR, destaca uma lacuna na pesquisa científica. Este estudo apresenta uma metodologia robusta de clusterização de bacias hidrográficas que incorpora múltiplos métodos de clusterização e aborda suas divergências, utilizando os dados do CAMELS-BR. A metodologia introduzida neste estudo envolve uma abordagem de clusterização multi-método que combina as técnicas K-means, Partitioning Around Medoids (PAM) e Fuzzy C-means (FCM). A literatura não explorou o estabelecimento de um consenso entre os métodos de clusterização para classificação, ao contrário da metodologia proposta neste estudo, que enfatiza a obtenção de uma classificação baseada no acordo coletivo entre múltiplos métodos, em vez de depender exclusivamente de métricas de desempenho individuais. A clusterização hidrológica realizada neste estudo apresenta um baixo nível de concordância com as regiões hidrográficas definidas pela ANA.

Palavras-chave: Clusterização; CAMELS; classificação hidrológica; multi-método.

ABSTRACT

The catchments parameters regionalization is crucial for streamflow prediction in ungauged basins, model parameterization, and watershed development and management. To overcome the limitation of reduced amount of hydrological data, the Catchment Attributes and MEteorology for Large-sample Studies – Brazil (CAMELS – BR) was produced and made publicly available. Limited application of clustering methods in catchment analysis in Brazil, particularly using the CAMELS-BR dataset, highlights a research gap in the literature. This study presents a robust catchment clustering methodology that incorporates multiple clustering methods and addresses their divergences, applied to the CAMELS-BR dataset. The methodology introduced in this study involves a multi-method clustering approach that combines the K-means, Partitioning Around Medoids (PAM), and Fuzzy C-means (FCM) techniques. The literature has not explored the establishment of a consensus among clustering methods for classification, unlike the methodology proposed in this study, which emphasizes deriving a classification based on collective agreement among multiple methods rather than relying solely on individual performance metrics. The hydrological clustering conducted in this study shows a low level of agreement with the hydrographic regions defined by ANA.

Keywords: Clustering; CAMELS; hydrological classification; multi-method.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

When dealing with the challenge of making predictions in ungauged basins (PUB) due to limited resources and trained perso

nnel, it becomes crucial to rely on procedures that transfer information from gauged catchment areas to ungauged ones. It is important to note that contiguous regions do not necessarily align with political or administrative boundaries, as well as recognized geographic or hydrologic divisions. Therefore, greater emphasis should be placed on these transfer procedures to generalize our understanding of hydrology and account for potential environmental impacts. Clustering and classification methods can be utilized for this purpose.

Classification serves as the initial and fundamental step in the scientific analysis and synthesis process. It enables hydrologists to organize and group the vast spatial, temporal, and process variability observed in hydrological phenomena. By classifying similar systems, the variability within classes is effectively constrained (McDonnell and Woods, 2004). Since the 1980s, a significant majority of regionalization studies have focused on objective catchment clustering using various methods. Streamflow, which serves as an integrator of climatic and morphologic conditions within a specific watershed, has been predicted using clustering techniques, particularly for flood frequency analysis (Burn, 1989; Srinivas et al., 2008).

Following the clustering process, catchments within the same cluster can be considered to exhibit similar characteristics. This has several practical implications. For example:

I) A water resource management plan designed for a particular watershed can be applied (with some modifications) to other catchments within the same cluster that have similar hydro-climatic conditions, particularly in studies on Regional Flood Frequency Analysis (RFFA).

II) Parameters of a hydrological model verified for one catchment within a cluster can be further validated for other catchments within the same cluster (Nourani et al., 2015a).

III) The available data from a catchment can be used to fill in missing data for other catchments within the same cluster.

Traditionally, the classification of catchments has been based on geographic, administrative, or physiographic factors. However, these approaches have proven to be inadequate in capturing the desired level of homogeneity among the regions (Burn, 1997). Alternative attempts were made to incorporate seasonality measures along with physiographic and meteorological characteristics, but obtaining such detailed information for a large number of catchments was deemed challenging (Burn, 1997). However, in the past decade, extensive datasets containing hydrological and geological data have become accessible, providing information on hundreds of catchments worldwide.

To address these limitations, was developed and publicly released a dataset called CAMELS-BR, (Catchment Attributes and MEteorology for Large-sample Studies – Brazil) (Newman et al., 2015; Addor et al., 2017) specifically designed for large-scale hydrological studies in Brazil. This comprehensive dataset comprises daily streamflow time series from 3,679 stream gauges. Additionally, for a selected subset of 897 catchments, it includes daily meteorological time series and 65 catchment attributes related to topography, climate, land cover, geology, soil, and human interventions.

Based on the acknowledged significance, this study presents the application of various clustering methods to the 897 Brazilian basins extracted from the CAMELS-BR database. These methods utilize the dataset's information on climate, land cover, soil, geology, and hydrological signatures. The focus is to classify and group the basins based on their similarities and dissimilarities, thereby enabling a comprehensive analysis of their hydrological characteristics and behaviors.

This study is presented in the form of a scientific article with the following title: Application of Clustering Methods for Hydrological Regionalization using the CAMELS-BR database. The article is structured as:

- Material and Methods, which explains the methodology and data used in the study.
- Results, where the findings of the study are presented.
- Discussion and Conclusion, where the results are discussed in relation to the existing literature, and the conclusions drawn from the study are provided.

2 OBJECTIVES

2.1 Main objective

This study proposed a robust methodology of catchment clustering that incorporate the agreement among multiple clustering methods and address their divergences instead of choosing the best performance clustering method.

2.2 Specific objectives

- Organize the data into distinct attribute classes, including climate, land cover, soil, geologic, and hydrological attributes.

- Apply Principal Components Analysis (PCA) to each attribute class to reduce dimensionality and extract essential information.
- Estimate the optimal number of clusters for each attribute class by evaluating three cluster quality metrics: Silhouette, Elbow, and Gap Statistic.
- Utilize the K-means, Partitioning Around Medoids (PAM), and Fuzzy C-means (FCM) clustering methods on each attribute class, considering the determined optimal number of clusters.
- Evaluate the clustering performance using the average silhouette width as the clustering quality metric.
- Employ a multi-method clustering approach by combining the results of the three clustering methods.

3 APPLICATION OF CLUSTERING METHODS FOR HYDROLOGICAL REGIONALIZATION USING THE CAMELS-BR DATABASE

3.1 Introduction

Data collection is crucial for understanding and analyzing various areas of study. However, certain basins face challenges due to a lack of available information. The ungauged basins can present significant obstacles in understanding and managing their water resources.

Parameter regionalization becomes essential in this context. The idea behind this process is to use data from hydrologically similar basins to estimate parameter values in basins with insufficient information. This approach is based on the premise that basins with similar physical characteristics tend to exhibit similar hydrological behaviors.

To organize and structure the obtained information, a systematic approach is necessary. Parameter clustering is a valuable tool in this regard. It involves grouping similar data into categories or clusters based on their common characteristics. This technique allows for the identification of patterns and trends in datasets, facilitating the understanding and interpretation of results.

The organization and clustering of catchments are crucial for streamflow prediction in ungauged basins, model parameterization, and watershed development and management (Singh et al., 2016). Transferring parameters from one catchment to other similar catchments can obviate the need for model calibration everywhere, saving valuable time and computer resources, and enabling model usage in ungauged catchments. To overcome the challenges of ungauged basins (PUB) due to the lack of funds, suitably trained personnel, and other factors, and to recognize that contiguous regions do not necessarily coincide with political or administrative areas and recognized geographic and/or hydrologic boundaries, greater reliance must be placed on procedures for transferring information from gauged catchment areas to ungauged ones (Sharghi et al., 2018). For the purpose of generalization, clustering and classification techniques can be employed.

Numerous studies have focused on the classification and organization of catchments based on physiographic characteristics (Winter, 2001, Brown et al., 2013; Buttle, 2006; Leibowitz et al., 2016), such as landscape and land use parameters, physical features (area, channel length, channel slope, etc.) (Rao and Srinivas, 2006), river morphology (Poff et al., 2006), and hydrologic similarity indices and signatures (Ley et al., 2011; Olden et al., 2012; Sawicz et al., 2011; Singh et al., 2016; Zhang et al., 2014), among others. These efforts are aimed at understanding the similarities and differences among catchments and their hydrological behavior.

In particular, researchers have attempted to organize catchments using a variety of approaches, including physical and climatic characteristics, and hydrologic signatures. The advantage of the first approach is that physical characteristics, such as topography and land cover, are now available for any location on earth, although the quality of available data may vary. The second approach groups catchments directly by their hydrologic behavior, which is the characteristic of primary concern (Kuentz et al., 2017).

Two major categories exist for classifying conventional clustering methods: hierarchical clustering and partitional clustering. (Rao and Srinivas, 2006). Different methods of clustering have been applied to classify catchments; however, the k-means algorithm is commonly utilized as a partitional clustering approach (Agarwal et al., 2016; Basu and Srinivas, 2016; Beskow et al., 2016; Corporal-Lodangco, 2014; Goyal and Gupta, 2014; Maruyama et al., 2005), due to its simplicity of implementation and computational efficiency (Jain, 2010).

Partitioning Around Medoids (PAM) is another partitional clustering technique that is often used to group similar catchments based on their hydrological characteristics (Snelder and Booker, 2013; Angus Webb et al., 2007; Laaha and Blöschl, 2006). PAM is a partitioning method that is similar to the K-means algorithm. However, instead of using centroids as cluster representatives, PAM defines a set of k objects as medoids or exemplars of each cluster (Laaha and Blöschl, 2006). Compared to the K-means algorithm, the PAM algorithm is known to produce more stable cluster solutions, as it is less sensitive to noise and outliers (Kaufman and Rousseeuw, 1990). This is because PAM uses medoids as cluster representatives, which are less likely to be influenced by extreme values in the data.

The third clustering method used in this study is the Fuzzy c-means (FCM) (Bezdek et al., 1981) which is is similar to the k-means and partitioning around medoids (PAM) algorithms, in that it seeks to partition a dataset into a predetermined number of clusters. However, unlike the K-means and PAM algorithms, FCM allows objects to belong to multiple clusters with varying degrees of membership, rather than being exclusively assigned to a single cluster (Shi et al., 2016; Schröter et al., 2015; Goktepe et al., 2005).

Despite the availability of several clustering methodologies, their application to the same dataset often yields divergent classifications, creating uncertainty and hindering integration into a homogeneous classification. Current literature on clustering catchments commonly employs a single clustering methodology (Jehn et al., 2020; Tongal and Sivakunar, 2017) or multiple methodologies with comparison of their performance (Sharghi et al., 2018).

To apply clustering method in catchments it is necessary to process a massive amount of data. However, access to open and readily available data is still difficult in some regions like South America (Chochemore, 2020), despite the growing number of large-sample datasets worldwide. In Brazil, large-sample hydrological studies face several challenges due to limited access to comprehensive datasets. Brazilian hydrometeorological information is collected and maintained by institutions like ANA (Brazilian National Water Agency) and INMET (National Institute of Meteorology), but accessing the data requires manual acquisition or web-scraping techniques, and the data formats lack consistency. Furthermore, current datasets do not systematically provide catchment attributes, hindering large-sample hydrological studies in Brazil. As a result, nationwide studies are less common than in North America or Europe, and studies in Brazil generally include only a reduced number of stream gauges and catchment attributes and are restricted to specific regions (Chagas et al., 2020).

A new dataset, Catchment Attributes and MEteorology for Large-sample Studies – Brazil (CAMELS – BR), developed by Chagas et al. (2020), has been made available to overcome these limitations in large-sample hydrological studies in Brazil. This dataset includes daily streamflow time series from 3679 stream gauges and daily meteorological time series and 65 catchment attributes for a selected group of 897 catchments. The catchment attributes are derived from various properties, such as topography, climate, land cover, geology, soil, and human intervention.

Despite the increasing availability of large datasets in the field of hydrology, there has been a notable lack of studies that apply clustering methods to catchment analysis in Brazil. Only a few studies have utilized clustering methods in this context (Bork et al., 2021), indicating a potential gap in the literature. When considering the use of CAMELS-BR to clustering none study have been developed.

In this study, it is proposed a robust methodology of catchment clustering that incorporate the agreement among multiple clustering methods and address their divergences instead of choosing the best performance clustering method. The proposed methodology was applied in Brazil using the CAMELS-BR dataset. We believe that is the first clustering study using this dataset.

3.2 Material and Methods

The proposed clustering methodology is detailed in Figure 1. It can be divided in six steps:

i) Data treatment: it consists in the organization of the data into the climate, land cover, soil, geologic, and hydrological attribute classes, the treatment of missing values, the data normalization and also the transforming of the geologic data to fit a classification model format;

ii) Principal Components Analysis: the application of PCA to each attribute class for dimensionality reduction;

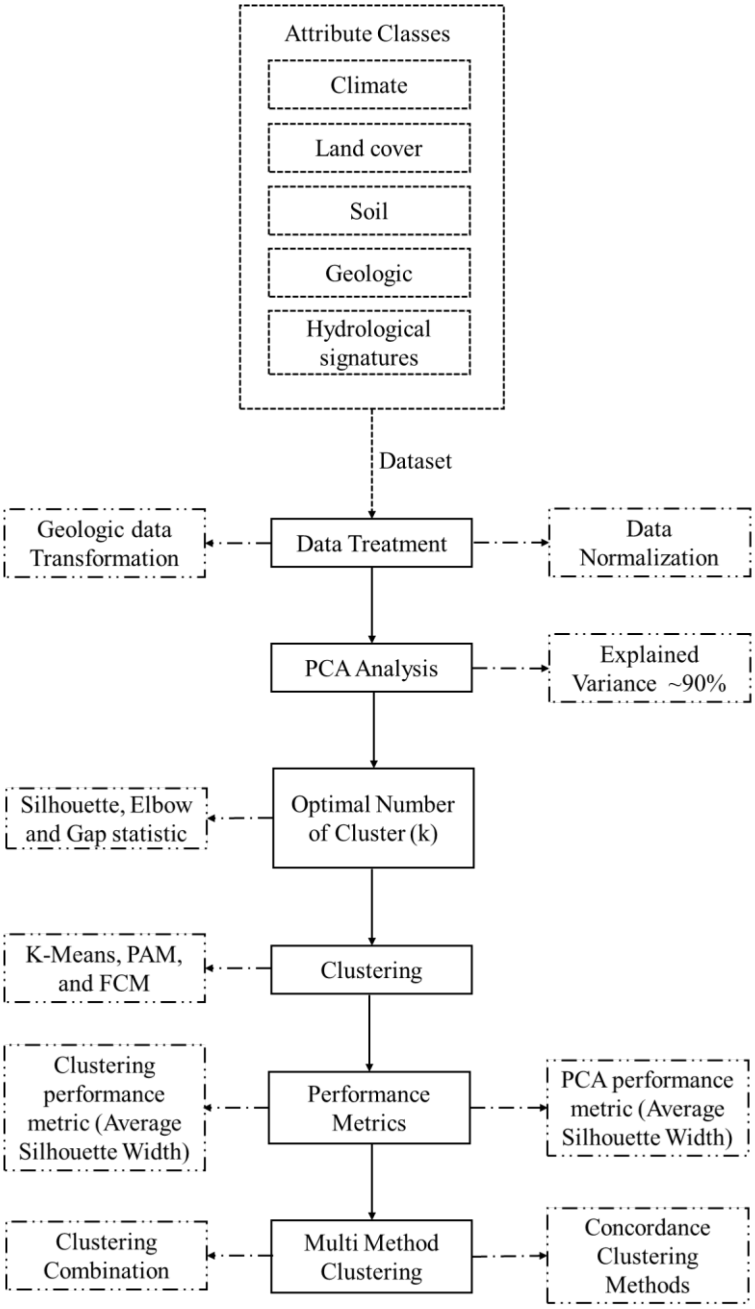
iii) Estimation of the optimal number of clusters for each class: the number of clusters is varied to find the one with optimal performance considering three cluster quality metrics: Silhouette, Elbow and Gap Statistic.

iv) Clustering: the application of the K-means, the Partitioning Around Medoids (PAM), and the Fuzzy C-means (FCM) clustering methods to each class considering the defined optimal number of clusters;

v) Clustering performance analysis: the comparison of the performance of the different clustering methods for each class using the average silhouette width as the clustering quality metric. Also, the use of PCA for dimensionality reduction is evaluated by comparing the overall clustering performance with and without the PCA;

vi) Multi-method clustering: the evaluation of the concordance between the clustering methods in each catchment; and the combination of the results of the three clustering methods for proposing a combined clustering conjecture.

Figure 1. Methodology flowchart



Source: Prepared by the author.

3.2.1 Data Treatment

This study is stablished on catchment atributes analysis and in hydrological signatures information. The CAMELS-BR contains daily time series of observed streamflow from 3679 gauges, as well as meteorological forcing (precipitation, evapotranspiration, and temperature) for 897 catchments, also, includes 65 attributes covering a range of topographic, climatic, hydrologic, land cover, geologic, soil, and human intervention variables (Chagas et al., 2020). All 897 catchments data were used because there are no missing values. These catchments have a wide range of characteristics like mean elevation ranging from 48 to 1691 m a.s.l and catchment areas ranging from 10 to 4.720.020 km². The attributes used per each database class were:

- Climate: Mean daily precipitation; Mean daily potential evapotranspiration (PET); Seasonality and timing of precipitation; Asynchronicity between the annual precipitation and PET cycles; Frequency of high precipitation days; Average duration of high precipitation events; Frequency of dry days; Average duration of low precipitation events.
- Land cover: percentage covered by a mosaic of croplands and natural vegetation; percentage covered by broadleaved or needle leaved forests; percentage covered by shrublands; percentage covered by grasslands or areas with sparse; percentage covered by barren areas; percentage covered by artificial surfaces or urban areas; percentage covered by water bodies or wetlands; percentage covered by permanent snow or ice.
- Geologic: percentage of the catchment covered by the most common geologic class; percentage of the catchment covered by the second most common geologic class; subsurface porosity; subsurface permeability.
- Soil: percentage of sand; percentage of silt; percentage of clay; soil organic carbon content; depth to bedrock; median water table depth.
- Hydrological signatures: mean daily discharge; runoff ratio; streamflow precipitation elasticity; mean half-flow date; 5% flow quantile; 95% flow quantile; frequency of high-flow days; average duration of high-flow events; frequency of low-flow days; average duration of low-flow events; days with zero discharge percentage.

Chagas et al. (2020) present the full description of each of the variables used and how they were calculated.

Before starting the data analysis, it was required to transform the geologic attributes (percentage of the catchment covered by the most common geologic class and by the second most commom) in a data format that was proper to apply clustering methodologies. So, a two steps transformation was used.

First, the aggregation of the ten geologic classes in the five main classes: Siliciclastic Sedimentary Rocks, Carbonate-Rich Sedimentary Rocks, Volcanics, Plutonics, and Metamorphics, accordingly to Hartmann and Moosdorf (2012). This step was done because at the hydrological point of view working with these five classes is satisfactory and makes the analysis easier to be done.

Second, the discretization of each attribute into five parameters. So, the attributes percentage of the catchment covered by the most and by the second most common geologic class becomes five attributes: Siliciclastic Sedimentary Rocks (%), Carbonate-Rich Sedimentary Rocks (%), Volcanics (%), Plutonics (%), and Metamorphics (%).

Before applying the principal components analysis (PCA) it is proper to centralize and scale the variables. Otherwise, the magnitude to certain variables dominates the associations between the variables in the sample. To centralize the data, it is and subtracted the mean of that attribute from each value. Given that the attribute values have different magnitudes it was done a data normalization in all classes by dividing them by their standard deviations.

2.2. Principal component analysis

Clustering methodologies usually are preceded by principal components analysis (PCA) in order to identify key spatial patterns in the data by reducing the number of variables for clustering cases (Shaharudin et al., 2018, Moron et al., 2015, Siva et al., 2014). The selection of the principal components that accounted together for, approximately, 90% of the explained variance resulted in the selection of three to eight principal components depending on the class of the attributes.

3.2.2 Clustering

Three methods of clustering were used: K-means, K-medoids or PAM, and Fuzzy c-means (FCM). K-means clustering is a popular unsupervised machine learning technique developed by MacQueen (1967). It is commonly used for dividing a given data set into k clusters, where k is predetermined by the analyst. The algorithm categorizes objects into multiple groups, aiming to maximize the similarity between objects within the same cluster (intra-class similarity) and minimize the similarity between objects in different clusters (inter-class similarity). K-means clustering also assigns each cluster a centroid, which is the mean of the points within that cluster (Kassambara, 2017).

The PAM algorithm (Partitioning Around Medoids) developed by Kaufman & Rousseeuw (1990) is the most common k-medoids clustering method. The k-medoids is similar to k-means, used for dividing a dataset into k clusters. In k-medoids, each cluster is represented by a data point within the cluster called a medoid, which refers to the object in a cluster with the lowest average dissimilarity to all other members. According to Kassambara (2017), unlike k-means, which calculates cluster centers as the mean value of all points in the cluster, k-medoids is more robust to noise and outliers since it uses medoids instead of means.

The last clustering method, the fuzzy c-means (FCM) (Bezdek et al., 1981), is one of the most widely used fuzzy clustering algorithms. Fuzzy clustering, considered a soft clustering, is a clustering method where each element in a dataset has a probability of belonging to each cluster. Each element has a set of membership coefficients, indicating the degree to which it belongs to each cluster. This differs from k-means and k-medoids clustering, where each object belongs to only one cluster, and is known as hard or non-fuzzy clustering. In fuzzy clustering, the degree to which an element belongs to a cluster can vary from 0 to 1, meaning that points closer to the center of a cluster may have a higher degree of membership than points on the edge of a cluster. In the FCM algorithm the cluster centroid is computed as the mean of all points, weighted by their degree of belonging to the cluster.

3.2.3 Optimal Number of Clusters (k)

One of the major issues in partitional clustering (K-means, PAM method) is to estimate the number of clusters (k) (Dinh et al., 2019). To solve this problem this study proposes a methodology to estimate k by comparing three methods that indicates the optimal number of clusters: silhouette method (Rousseeuw, 1987), elbow method (Cui, 2020sq), and gap statistic (Tibshirani, 2001). The silhouette and the elbow method are worldwide methods (SAPUTRA et al., 2020, Et-taleby et al., 2020, SETIADY, 2021) that are used to determine the value of k or to determine the validity of a cluster method. The gap statistic method was developed by Tibshirani et al. (2001) to efficiently estimate the optimal number of clusters to any cluster technique and distance method.

The Silhouette Method measures how well each point lies within its cluster (Kassambara, 2017). So, the average silhouette returns the average of the silhouette coefficients in each cluster. A high average silhouette width indicates a good clustering. It was computed the average silhouette to all clusters changing the k number and plotted in a graph. Then, it was possible to choose the best k number.

The Elbow method involves analyzing the percentage of variance explained with respect to the number of clusters. This approach is based on the concept that the optimal number of clusters is reached when the addition of another cluster does not significantly improve the data modeling. The percentage of variance explained by the clusters is graphed against the number of clusters, and while the initial clusters contribute a lot of information, there comes a point where the marginal gain drops sharply, forming an angle in the graph. The elbow criterion involves selecting the appropriate "k," or number of clusters, at this juncture (Bholowalia and Kumar, 2014).

The gap statistic method is a statistical technique used to estimate the optimal number of clusters in a dataset. It is similar to the Elbow method, but instead of looking for the point where the marginal gain drops sharply, the gap statistic method compares the observed within-cluster dispersion to a reference distribution of the data generated under a null reference distribution (El-Mandouh et al., 2017). To choose the optimal number of clusters it is plotted the gap statistic against number of cluster and the one where by adding more clusters will not change the gap statistic number significantly.

Therefore, these tree metrics were applied to the three clustering methods: K-means, PAM and Fuzzy c-means, varying the number of clusters between 1 to 20. This was applied in each class of attributes (climate, land cover, geologic, soil, and hydrological signatures). Then, for each class of attributes, it is obtained a 3x3 matrix (three optimal number methods of clusters times three clustering methods) with the values of the optimal number of clusters. From that matrix it is possible to choose the best number of clusters to each catchment attributes class by selecting the optimal number of clusters that most appears in the matrix.

2.5 Cluster Performance Metrics

The quality of the clustering procedure is evaluated using the Average Silhouette Width (ASW) as metric.

Let $X = \{x_1, \dots, x_n\}$ be a data set of n objects from a space χ, d be a dissimilarity or distance over χ. We deal with clusterings C = $\{C_1, \dots, C_k\}$ that are partitions, i.e., non-overlapping and exhaustive. A partition can equivalently be expressed by labels $l(1), \dots, l(n) \in \mathbb{N}_k = \{1, \dots, k\}$ where $l(i) = r \Leftrightarrow x_1 \in C_r,\ i \in \mathbb{N}_n$, and cluster sizes are denoted by $n_r = \sum_{i=1}^{n} 1(l(i) = r),\ r \in \mathbb{N}_k$.

According to Batool and Hennig (2021), the silhouette width for an observation $x_i \in X$ is:

$$s_i(C, d) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{1}$$

Where:

$$a(i) = \frac{1}{n_{l(i)} - 1} \sum_{\substack{l(i)=l(j) \\ i \neq j}} d(x_i, x_j) \tag{2}$$

And:

$$b(i) = min_{r \neq l(i)} \frac{1}{n_r} \sum_{l(j)=r} d(x_i, x_j) \tag{3}$$

In case that $n_r > 1$ for $l(i) = r$. Otherwise $s_i(C, d) = 0$.

The ASW of a clustering C is defined as:

$$\bar{S}(C,d) = \frac{1}{n}\sum_{i=1}^{n} s_i(C,d) \tag{4}$$

The value of $a(i)$ represents the average distance between x_i and the points within its assigned cluster. On the other hand, $b(i)$ represents the average distance between x_i and the points within the nearest cluster that it is not assigned to. When $s_i(C,d)$ has a high value, it indicates that $b(i)$ is significantly larger than $a(i)$. Consequently, x_i is much closer to the observations in its own cluster compared to the neighboring cluster.

In the context of clustering, where the objective is to have homogeneous and well-separated clusters, larger values of si and $\bar{S}$ indicate a higher quality of clustering. An optimal clustering, especially when comparing different values of k, is the one that maximizes $\bar{S}$ according to the ASW (Average Silhouette Width) measure.

The average silhouette width was calculated for all clusters generated by each clustering method. This allowed the analysis of the performance of different clustering methods and to compare them to one another. Furthermore, this analysis was extended to the dataset without using the PCA procedure to determine if the application of PCA improved the overall clustering quality.

3.2.4 Multi-method Clustering

The Multi-method Clustering seeks to propose clusters that are based on the spatial similarity of the three different methods and on the treatment of their disagreement.

First, the results of the different methods are aligned so that clusters representing the same region have the same value. To match the clusters values, it was applied the Euclidean distance between the variables averages between each cluster from different methods resulting in a Euclidean Distance Matrix (Lele, 1993). Then, it was analyzed the minimal distance between clusters from different methods in order to relate different methods and concur the number of clusters which represent basins with similar attributes values of a class. The Euclidean Distance Matrix has this representation:

$$F(X) = \begin{bmatrix} 0 & d(1,2) & d(1,3) & \dots & d(1,K) \\ d(2,1) & 0 & d(2,3) & \dots & d(2,K) \\ \vdots & d(3,2) & 0 & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & 0 \end{bmatrix} \tag{5}$$

Where $d(l,m)$ denotes the Euclidean distance between landmarks l and m.

Then, a catchment-by-catchment analysis is carried out to verify which region each clusters classified the catchment. The concordance of the clustering methods is verified by setting a concordance number varying from 1 to 3. The number 1 means that the three clustering methods disagree for the region of the catchment; the number 2 means that two of the clustering methods agree in the region classification of the catchment and the number 3 means that all three clustering methods agree in the region classification of the catchment.

For catchments with concordance number 3, the multi-method clustering will define the Cluster as the same as the three methods are in agreement. For catchments with concordance number 2, the multi-method clustering will define the Cluster as the one that the 2 methods are in agreement. For catchments with concordance number 1, the multi-method clustering will define the Cluster as the one pointed by the method with the higher average silhouette width in the performance metric analysis.

3.3 Results

3.3.1 Principal component analysis

Principal component analysis (PCA) was conducted by selecting the number of principal components (PCs) that accounted for approximately 90% of the total variance. The results of the PCA results are summarized in Tables 1 to 5, which provide information on the variance of each principal component for each class.

Table 1. Summary PCA to climate variables

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
Standard deviation	1.87	1.61	1.02	0.52	0.52	0.43	0.26	0.20
Proportion of Variance	0.43	0.32	0.13	0.034	0.034	0.023	0.008	0.005
Cumulative Proportion	43.9%	76.4%	89.4%	92.9%	96.3%	98.6%	99.5%	100.0%

Source: Prepared by the author.

Table 2. Summary PCA to land cover variables

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
Standard	1.53	1.32	1.08	0.97	0.89	0.77	0.50	0.31

deviation								
Proportion of Variance	0.29	0.21	0.14	0.11	0.09	0.07	0.03	0.01
Cumulative Proportion	29.6%	51.5%	66.1%	78.0%	87.9%	95.5%	98.7%	100.0%

Source: Prepared by the author.

Table 3. Summary PCA to soil variables

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Standard deviation	1.84	1.08	0.75	0.70	0.58	0.001
Proportion of Variance	0.5697	0.1951	0.0949	0.08341	0.05691	0.00
Cumulative Proportion	57.0%	76.5%	86.0%	94.3%	100.0%	100.0%

Source: Prepared by the author.

Table 4. Summary PCA to geology variables

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
Standard deviation	1.68	1.16	1.00	0.9424	0.9	0.20166	0.17922
Proportion of Variance	0.40	0.19	0.14	0.12	0.11	0.005	0.004
Cumulative Proportion	40.7%	60.2%	74.7%	87.4%	99.0%	99.5%	100.0%

Source: Prepared by the author.

Table 5. Summary PCA to hydrology variables

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC 10**	**PC 11**
Standard deviation	2.26	1.52	1.07	0.93	0.73	0.58	0.50	0.49	0.25	0.23	0.12

Proportion of Variance	0.46	0.21	0.10	0.07	0.04	0.03	0.02	0.02	0.005	0.004	0.001
Cumulative Proportion	46.6%	67.7%	78.2%	86.2%	91.2%	94.2%	96.6%	98.8%	99.4%	99.9%	99.9%

Source: Prepared by the author.

So, it was selected the following number of PCs to each class:

Table 6. Number of PCs to each class

Class	Number of PCs
Climate	3
Land cover	5
Soil	4
Geology	4
Hydrology	5

Source: Prepared by the author.

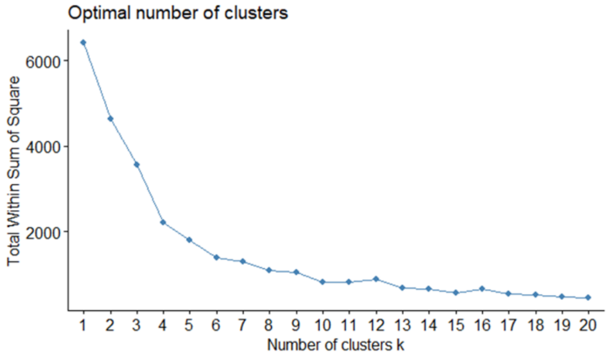
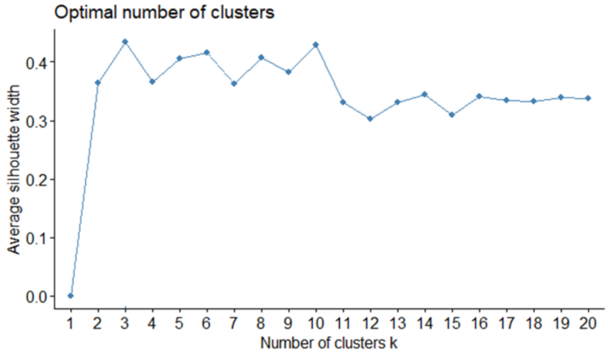
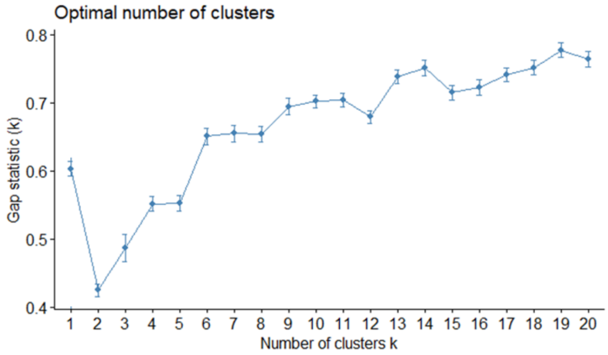
3.3.2 Optimal Number of Clusters

The optimal number of clusters for each class of data was determined using three partitioning algorithms: K-means, PAM, and FCM, and three evaluation methods: elbow, silhouette width, and gap statistic. The appropriate number of clusters was chosen based on the analysis of graphs generated by these methods, which are presented in Annex. Table 7 presents an example of optimal number of clusters results applying K-means in climate data.

The elbow method was used to select the optimal k value from the graph by identifying the point where the slope of the line changes and further clustering does not significantly decrease the total within sum of square. The silhouette width method was employed to choose the highest average silhouette width among the twenty number of clusters, excluding values below 4, which would result in too few clusters for a large country like Brazil.

In contrast, the gap statistic method was used by selecting the point where the line slope changes and adding more clusters does not significantly increase the gap statistic value. However, when applying FCM to the hydrology data, the resulting graphs exhibited non-uniform behavior, making it impossible to determine the optimal number of clusters using the three methods mentioned.

Table 7 – Optimal Number of Clusters results to methods applying K-means in Climate data

Elbow Method	
Silhouette Width	
Gap Statistic	

Source: Prepared by the author.

The selected k values for each evaluation method using three partitional algorithm for each class are presented in Tables 8 to 12. The selected k value for each class represents the number of clusters that appeared most frequently across the different evaluation methods and partitioning algorithms used in this study. Specifically, for the climate, land cover, soil, geology, and hydrology classes, k values of 6, 6, 6, 5, and 10 were selected, respectively.

Table 8 – Optimal number of clusters results to methods in Climate data

Optimal Clusters	Clustering Method		
Number Method	K-Means	PAM	Fuzzy
Elbow	**6**	**6**	4
Silhouette	10	**6**	10
Gap Statistic	**6**	**6**	11

Source: Prepared by the author.

Table 9 – Optimal number of clusters results to methods in Land Cover data

Optimal Clusters	Clustering Method		
Number Method	K-Means	PAM	Fuzzy
Elbow	7	8	5
Silhouette	5	9	**6**
Gap Statistic	**6**	**6**	4

Source: Prepared by the author.

Table 10 – Optimal number of clusters results to methods in Soil data

Optimal Clusters	Clustering Method		
Number Method	K-Means	PAM	Fuzzy
Elbow	6	**6**	4
Silhouette	5	**6**	4
Gap Statistic	**6**	9	9

Source: Prepared by the author.

Table 11 – Optimal number of clusters results to methods in Geology data

Optimal Clusters	Clustering Method		
Number Method	K-Means	PAM	Fuzzy
Elbow	6	**5**	6
Silhouette	4	**5**	**5**
Gap Statistic	9	9	12

Source: Prepared by the author.

Table 12 – Optimal number of clusters results to methods in Hydrology data

Optimal Clusters	Clustering Method

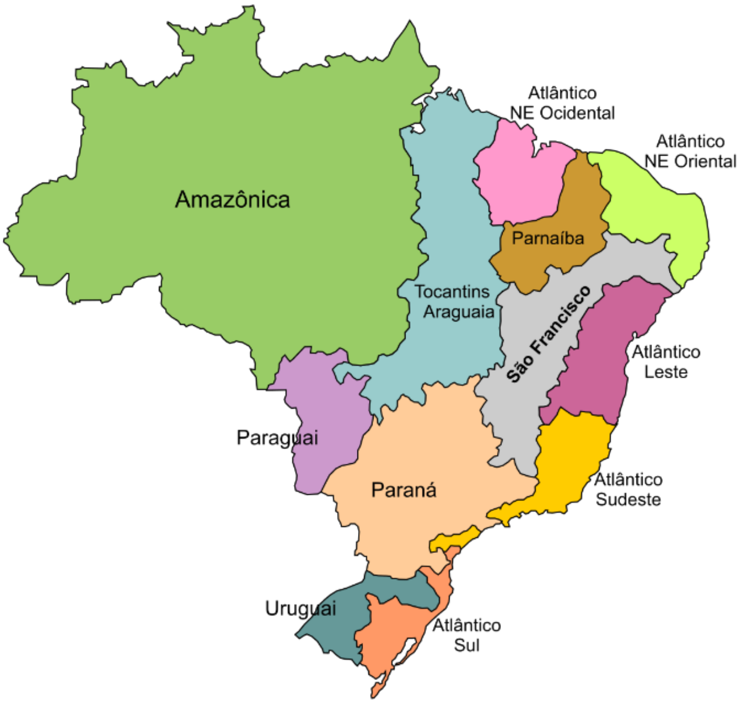
Number Method	K-Means	PAM	Fuzzy
Elbow	9	5	-
Silhouette	4	**10**	7
Gap Statistic	13	**10**	-

Source: Prepared by the author.

3.3.3 Clustering

Following the determination of the optimal number of clusters for each class, three different methods were used to perform cluster analysis, and the resulting clusters were presented in Figures 3 to 7. Additionally, the basins clustering results were compared to the twelve hydrographic regions defined by the National Water Agency (Agência Nacional das Águas, ANA), based on their hydrological features, displayed in Figure 2.

Figure 2 – ANA’s hydrographic regions



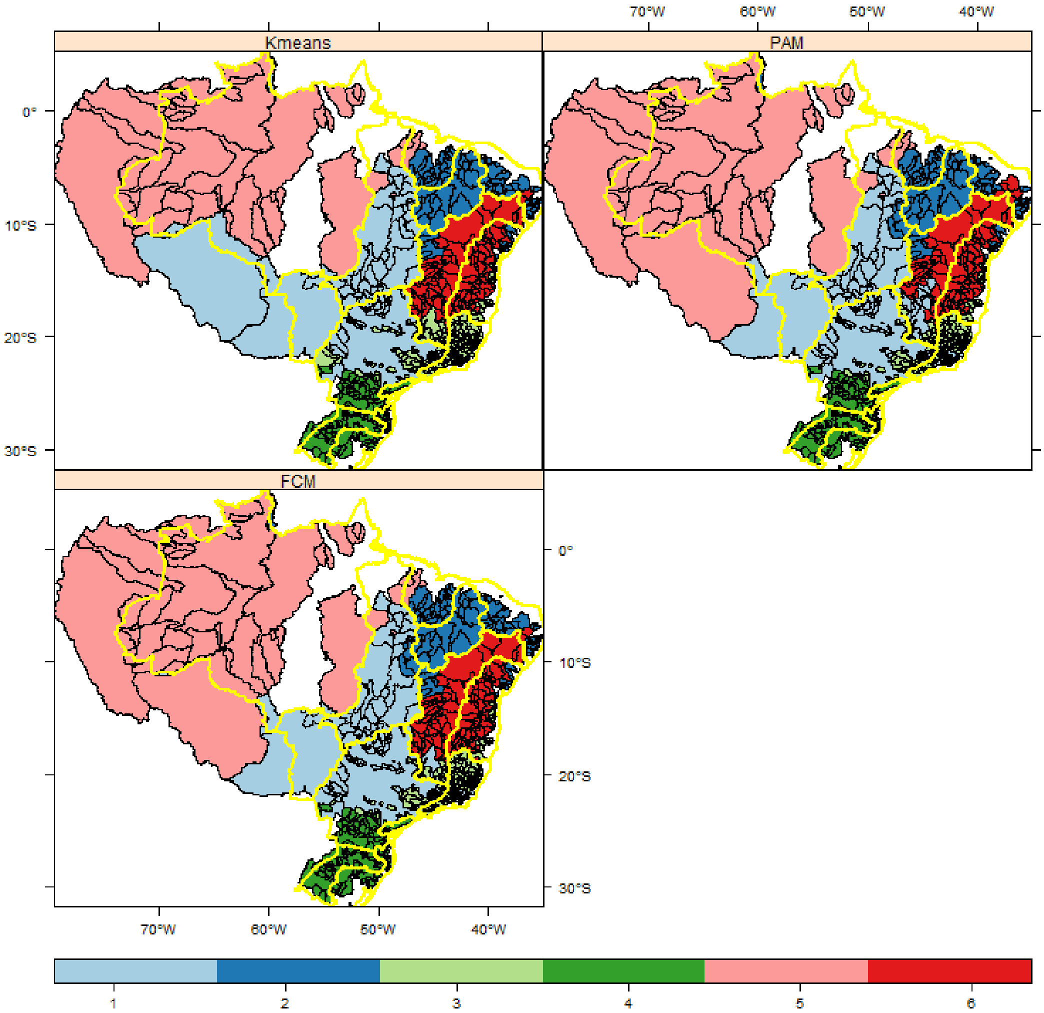
Source: Prepared by the author.

3.3.3.1 Climate Clustering

The climate clustering map revealed a high degree of homogeneity across the basins, enabling the identification of ANA's hydrographic regions in each cluster by comparing the twelve ANA-defined regions to the climate clusters map. Analyzing the Figure 3 it is seen that the three clustering methods showed almost the same result demonstrating high concordance. Specifically, the analysis revealed that the "Amazon" region defined a distinct climate cluster (cluster 5).

Furthermore, the "Paraguai," "Paraná," and "Tocantins Araguaia" regions were almost totally grouped in cluster number 1, having the exception of the more southern basins of "Paraná" that area aggrouped in the cluster number 4. The "Atlantico NE Ocidental," "Parnaiba," and "Atlantico NE Oriental" regions were grouped almost totally together in cluster 2. The medium and lower "São Francisco" and "Atlantico Leste" regions formed the cluster 6. The basins of cluster 3 are almost all located in "Atlantico Sudeste" region. The regions "Uruguai" and "Atlântico Sul" regions comprised cluster 4.

Figure 3 – Maps with clustered basins using climate data



Source: Prepared by the author.

The Tables 13, 14, and 15 contains the average values to each climate attribute to the cluster developed by K-means, PAM, and FCM, respectively. Analyzing the table is possible to see that cluster 1 indicates high values of precipitation and potential evapotranspiration, and seasonality value close to 1 indicating that the peak of precipitation and temperature ate in phase. The cluster 2 represents the basins with low precipitation and high PET, and a high number of days with precipitation above 1mm (low_prec_freq).

The cluster 3 indicates the basins that a high level of precipitation, a medium value of PET, a number of seasonality very close to 1 indicating peaks of precipitation and temperature in phase, and high frequency of dry and humid days. The cluster number 4 represents basins with high precipitation, low PET, and high frequency of dry and humid days.

The cluster number 5 that mainly represents the Amazon basin have the highest precipitation average, a high number of PET, a seasonality number close to 0 that indicates uniform precipitation throughout the year, and low frequency of high and low precipitation values.

The cluster 6 that represents the São Francisco basin have low average precipitation and high PET value, which it is explained by being located in a semiarid region. Furthermore, this cluster represent basins with high frequency of humid and dry days, and the highest average of duration of days with low precipitation.

Table 13 – Climate attribute average values in cluster applying K-means (p_mean: Mean daily precipitation, pet_mean: Mean daily potential evapotranspiration, p_seasonality: Seasonality and timing of precipitation, asynchronicity: Asynchronicity between the annual precipitation and PET cycles, high_prec_freq: Frequency of high precipitation days, high_prec_dur: Average duration of high precipitation events, low_prec_freq: Frequency of dry days, high_prec_freq: Average duration of high precipitation events)

Attribute	**1**	2	3	4	5	6
p_mean (mm/day)	4.19	2.76	3.98	4.60	5.93	2.83
pet_mean (mm/day)	3.14	3.38	2.77	2.44	3.09	3.11
p_seasonality (adm)	0.90	-0.41	1.03	0.16	-0.28	1.16
Asynchronicity (adm)	0.15	0.31	0.04	0.03	0.17	0.11
high_prec_freq (days/year)	12.95	18.8	21.85	22.15	4.95	22.15
high_prec_dur (days)	1.16	1.27	1.27	1.17	1.03	1.49
low_prec_freq (days/year)	197.65	253.05	248.8	245.25	127.95	261.3
low_prec_dur (days)	5.39	6.05	5.43	4.39	2.91	7.03

Source: Prepared by the author.

Table 14 – Climate attribute average values in cluster applying PAM (p_mean: Mean daily precipitation, pet_mean: Mean daily potential evapotranspiration, p_seasonality: Seasonality and timing of precipitation, asynchronicity: Asynchronicity between the annual precipitation and PET cycles, high_prec_freq: Frequency of high precipitation days, high_prec_dur: Average duration of high precipitation events, low_prec_freq: Frequency of dry days, high_prec_freq: Average duration of high precipitation events)

Attribute	**1**	2	3	4	5	6
p_mean (mm/day)	4.09	2.96	3.98	4.61	5.82	2.66
pet_mean (mm/day)	3.11	3.39	2.76	2.45	3.09	3.11
p_seasonality (adm)	0.94	-0.44	1.03	0.17	-0.27	1.11
Asynchronicity (adm)	0.14	0.31	0.04	0.04	0.18	0.11
high_prec_freq (days/year)	15.63	18.48	22.20	22.15	4.95	22.53
high_prec_dur (days)	1.19	1.25	1.29	1.18	1.03	1.50
low_prec_freq (days/year)	210.33	244.00	249.95	245.45	127.95	265.65
low_prec_dur (days)	5.43	5.92	5.44	4.40	2.94	7.13

Source: Prepared by the author.

Table 15 – Climate attribute average values in cluster applying FCM (p_mean: Mean daily precipitation, pet_mean: Mean daily potential evapotranspiration, p_seasonality: Seasonality and timing of precipitation, asynchronicity: Asynchronicity between the annual precipitation and PET cycles, high_prec_freq: Frequency of high precipitation days, high_prec_dur: Average duration of high precipitation events, low_prec_freq: Frequency of dry days, high_prec_freq: Average duration of high precipitation events)

Attribute	**1**	2	3	4	5	6

p_mean (mm/day)	4.11	2.94	4.00	4.61	5.82	2.87
pet_mean (mm/day)	3.11	3.40	2.76	2.45	3.09	3.10
p_seasonality (adm)	0.93	-0.44	1.03	0.17	-0.27	1.14
Asynchronicity (adm)	0.15	0.31	0.04	0.04	0.18	0.11
high_prec_freq (days/year)	15.08	18.55	22.15	22.15	4.95	22.30
high_prec_dur (days)	1.19	1.26	1.28	1.18	1.03	1.48
low_prec_freq (days/year)	207.73	246.95	249.70	245.35	127.95	262.05
low_prec_dur (days)	5.37	5.96	5.42	4.40	2.94	6.92

Source: Prepared by the author.

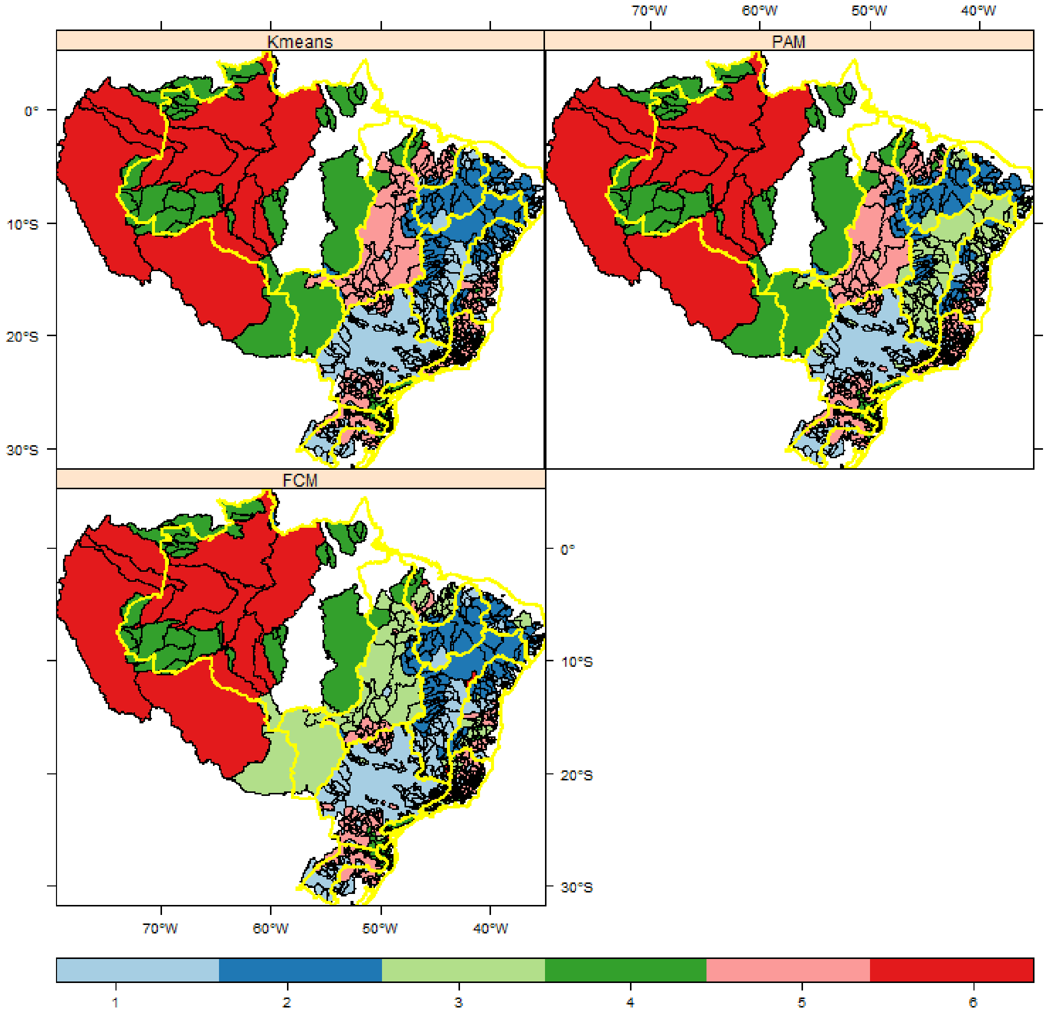
3.3.3.2 Land Cover Clustering

Upon comparing the land cover clustering (see Figure 4) with the hydrographic regions defined by ANA, notable distinctions in homogeneaty arise when contrasted with the climate clustering. Furthermore, variations emerge in the outcomes of different clustering methods.

In both K-means and PAM results, the basins within the "Tocantins-Araguaia" region are classified into Cluster 5, which includes basins from the "Atlantico Sudeste" region and certain basins in the southern part of Brazil. However, the FCM result indicates that the "Tocantins-Araguaia" basins are grouped within Cluster 3, implying greater similarity to the "Paraguai" region.

Another discrepancy pertains to the "São Francisco" region, where the PAM clustering assigns almost all basins to an isolated cluster, which contrasts with the K-means and FCM results, where the "São Francisco" region is divided into one northern cluster and one southern cluster.

Figure 4 – Maps with clustered basins using land cover data



Source: Prepared by the author.

Tables 16, 17, and 18 provide the mean values for each land cover attribute within the clusters generated by the respective clustering methods. Across all methods, Cluster 1 consistently comprises basins characterized by a high percentage of croplands, which aligns with the agricultural-centric nature of the central region of Brazil.

Cluster 2 represents basins with a prominent presence of shrublands, characteristic of the "caatinga" biome located in the northeastern part of Brazil. Cluster 3 exhibits varying behavior among the clustering methods: in K-means, it has limited representativeness; in PAM, it predominantly represents basins within the São Francisco region, which exhibit a high percentage of cropland (Table 31); and in FCM, it represents the Tocantins-Araguaia and

Paraguai regions while featuring a higher proportion of forest cover (Table 32).

Cluster 4 uniformly represents basins within the Amazon Forest, showcasing a substantial forest cover (above 80%) across all methods. Cluster 5 exhibits comparable attribute averages between methods, with cropland and forest percentages ranging from 30% to 40%. Notably, the FCM method excludes the basins of Tocantins-Araguaia within Cluster 5, distinguishing it from the K-means and PAM methods.

Cluster 6 demonstrates a similar definition across all methods and primarily represents basins along the Amazonas River, characterized by high forest percentages and a more pronounced presence of water bodies.

Table 16 – Land cover attribute average values in cluster applying K-means (crop_mosaic_perc: Percentage covered by a mosaic of croplands and natural vegetation, forest_perc: Percentage covered by broadleaved or needleleaved forests, shrub_perc: Percentage covered by shrublands, grass_perc: Percentage covered by grasslands, barren_perc: Percentage covered by barren áreas, imperv_perc: Percentage covered by artificial surfaces or urban áreas, wet_perc: Percentage covered by water bodies or wetlands, snow_perc: Percentage covered by permanent snow or ice)

Attribute	1	2	3	4	5	6
crop_mosaic_perc (%)	58.69	33.54	33.28	11.64	35.02	4.69
forest_perc (%)	11.83	14.94	25.11	81.92	39.43	74.33
shrub_perc (%)	8.68	39.91	24.74	2.44	4.93	5.90
grass_perc (%)	0.01	0.09	0.37	0.00	0.00	2.61
barren_perc (%)	0.00	0.03	0.05	0.00	0.00	0.19
imperv_perc (%)	0.00	0.00	3.38	0.00	0.00	0.01
wet_perc (%)	0.02	0.05	0.45	0.01	0.05	6.67
snow_perc (%)	0.00	0.00	0.00	0.00	0.00	0.07

Source: Prepared by the author.

Table 17 – Land cover attribute average values in cluster applying PAM (crop_mosaic_perc: Percentage covered by a mosaic of croplands and natural vegetation, forest_perc: Percentage covered by broadleaved or needleleaved forests, shrub_perc: Percentage covered by shrublands, grass_perc: Percentage covered by grasslands, barren_perc: Percentage covered by barren áreas, imperv_perc: Percentage covered by artificial surfaces or urban áreas,

wet_perc: Percentage covered by water bodies or wetlands, snow_perc: Percentage covered by permanent snow or ice)

Attribute	1	2	3	4	5	6
crop_mosaic_perc (%)	60.18	31.10	52.73	12.17	35.28	4.69
forest_perc (%)	13.84	20.25	11.69	81.58	39.87	74.33
shrub_perc (%)	4.42	42.05	22.61	2.46	4.48	5.90
grass_perc (%)	0.00	0.08	0.09	0.00	0.00	2.61
barren_perc (%)	0.00	0.02	0.02	0.00	0.00	0.19
imperv_perc (%)	0.00	0.00	0.00	0.00	0.00	0.01
wet_perc (%)	0.01	0.06	0.08	0.02	0.04	6.67
snow_perc (%)	0.00	0.00	0.00	0.00	0.00	0.07

Source: Prepared by the author.

Table 18 – Land cover attribute average values in cluster applying FCM (crop_mosaic_perc: Percentage covered by a mosaic of croplands and natural vegetation, forest_perc: Percentage covered by broadleaved or needleleaved forests, shrub_perc: Percentage covered by shrublands, grass_perc: Percentage covered by grasslands, barren_perc: Percentage covered by barren áreas, imperv_perc: Percentage covered by artificial surfaces or urban áreas, wet_perc: Percentage covered by water bodies or wetlands, snow_perc: Percentage covered by permanent snow or ice)

Attribute	1	2	3	4	5	6
crop_mosaic_perc (%)	59.54	34.46	32.37	11.93	39.12	20.02
forest_perc (%)	10.81	14.88	40.91	81.92	37.89	50.89
shrub_perc (%)	8.73	38.94	17.38	2.29	2.61	12.97
grass_perc (%)	0.01	0.09	0.05	0.00	0.00	0.54
barren_perc (%)	0.00	0.03	0.01	0.00	0.00	0.17
imperv_perc (%)	0.00	0.00	0.00	0.00	0.00	0.03
wet_perc (%)	0.02	0.04	0.32	0.01	0.01	2.57
snow_perc (%)	0.00	0.00	0.00	0.00	0.00	0.00

Source: Prepared by the author.

3.3.3.3 Soil Clustering

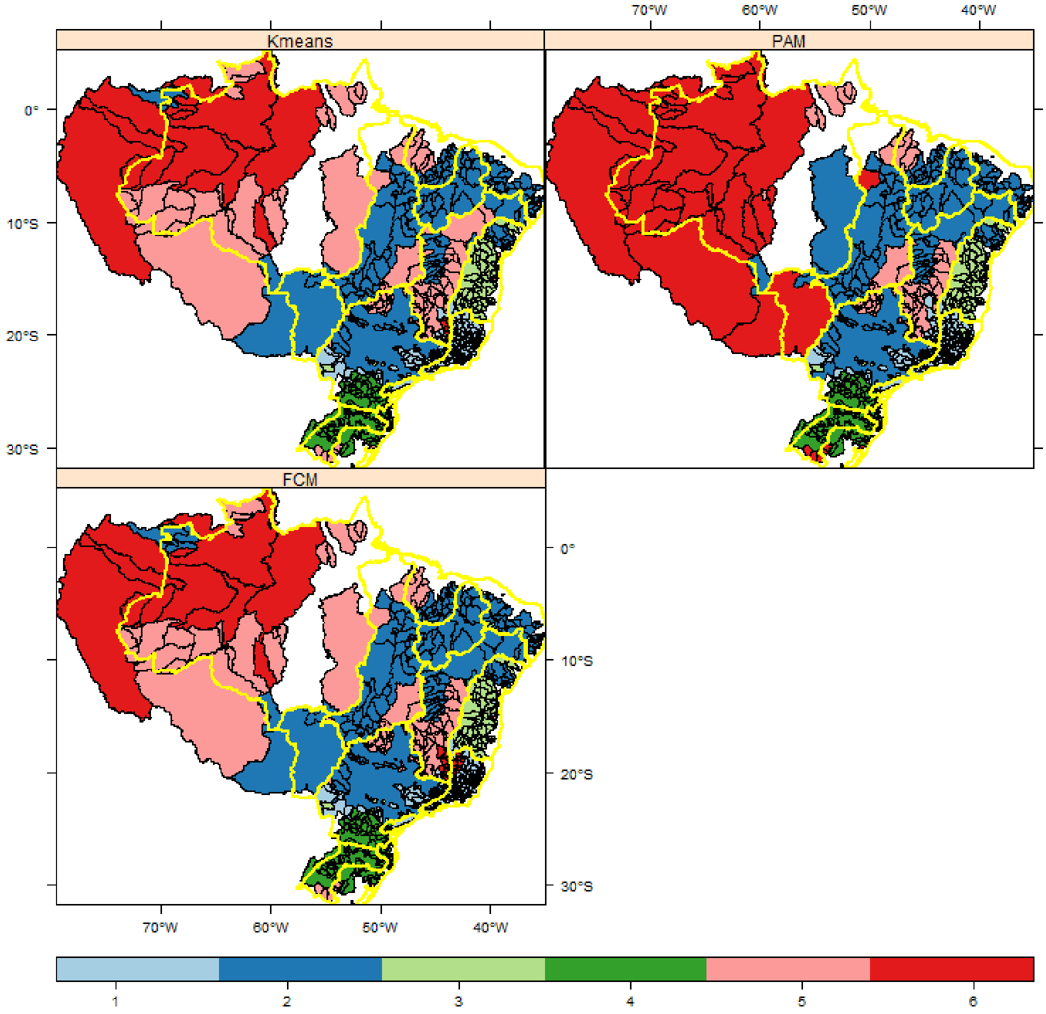
Upon analyzing Figure 5, which presents the clustering maps of soil data, a high level of concordance among the methods is observed. The primary difference arises in the assignment of more basins from the Amazon and Paraguai regions to Cluster 6 in the PAM clustering method.

Cluster 1 predominantly comprises basins located in the "Atlântico Sudeste" region and the southwestern part of the Parana region. Cluster 2 encompasses basins situated in the western portion of the "Atlantico NE Ocidental" region, the entire "Atlantico NE Oriental" region, the complete "Parnaíba" region, the lower São Francisco region, the northern part of the "Atlantico Leste" region, the western portion of the "Tocantins-Araguaia" region, the complete "Paraguai" region (in the K-means and FCM methods), as well as some basins in the "Parana" region.

Basins assigned to Cluster 3 are predominantly found throughout the "Atlantico Leste" region. Cluster 4 indicates basins located in the southern region of Brazil, encompassing the "Uruguai" and "Atlântico Sul" regions, as well as the southern part of the "Parana" region.

Finally, Cluster 6 represents basins within the "Amazon" region.

Figure 5 – Map with clustered basins using soil data



Source: Prepared by the author.

Tables 19 to 21 contain the average values of each attribute within the clusters generated by the respective clustering methods. Clusters 1 to 5 exhibit similar attribute averages across all three methods, whereas Cluster 6 displays some variations among them.

Cluster 1 is characterized by high and comparable percentages of sand and clay (around 40%), a lower percentage of silt (approximately 17%), a significantly high content of soil organic carbon, and elevated values of bedrock and median water table depth.

Cluster 2 represents basins with the highest proportion of sand, lower percentages of silt and clay, a low amount of organic soil, a deep bedrock depth, and a shallow median water table depth. Cluster 3 indicates basins with more similar values of sand and clay compared to Cluster 2, but with a predominant presence of sand, low levels of organic soil, and the highest median water table depth. Basins in Cluster 4 exhibit a predominance of clay, distinguishing them from the other clusters, along with a low bedrock depth and a high amount of organic soil.

Cluster 5 displays attribute averages similar to Cluster 3 in terms of sand, silt, clay, and organic soil percentages, but with a greater bedrock depth and lower median water table depth. The attribute averages of Cluster 6 differ among the methods: K-means and FCM indicate similar percentages of sand and clay, a lower value of silt, a considerably high level of organic soil, and a moderate median water table depth. In contrast, the PAM method suggests a higher percentage of silt, a lower percentage of clay, significantly lower levels of organic soil, and median water table depth.

Table 19 – Soil attribute average values in cluster applying K-means (sand_perc: Percentage of sand, silt_perc: Percentage of silt, clay_perc: Percentage of clay, org_carbon_content: Soil organic carbon content at a soil depth of 30 cm, bedrock_depth: Depth to bedrock, water_table_depth: Median water table depth)

Attribute	1	2	3	4	5	6
sand_perc (%)	40.56	55.30	46.37	27.32	43.81	41.16
silt_perc (%)	16.99	16.09	15.44	24.86	17.55	18.76
clay_perc (%)	41.62	29.04	38.49	47.54	37.88	37.27
org_carbon_content (g/kg)	16.76	8.29	10.82	18.73	10.78	31.60
bedrock_depth (cm)	2099.01	2650.66	2462.09	932.63	3080.86	2237.67
water_table_depth (cm)	2601.69	1646.74	3781.50	2112.28	1571.36	1774.44

Source: Prepared by the author.

Table 20 – Soil attribute average values in cluster applying PAM (sand_perc: Percentage of sand, silt_perc: Percentage of silt, clay_perc: Percentage of clay, org_carbon_content: Soil organic carbon content at a soil depth of 30 cm, bedrock_depth: Depth to bedrock, water_table_depth: Median water table depth)

Attribute	1	2	3	4	5	6
sand_perc (%)	40.40	55.10	45.15	27.07	43.94	41.82
silt_perc (%)	17.27	16.28	15.55	24.93	17.38	23.88

clay_perc (%)	41.65	29.27	39.63	47.95	38.92	33.27
org_carbon_content (g/kg)	17.45	8.28	11.90	18.97	10.37	15.01
bedrock_depth (cm)	2133.37	2678.30	2225.74	923.36	3199.63	2119.78
water_table_depth (cm)	2543.10	1665.14	3687.39	2105.61	1844.28	326.56

Source: Prepared by the author.

Table 21 – Soil attribute average values in cluster applying FCM (sand_perc: Percentage of sand, silt_perc: Percentage of silt, clay_perc: Percentage of clay, org_carbon_content: Soil organic carbon content at a soil depth of 30 cm, bedrock_depth: Depth to bedrock, water_table_depth: Median water table depth)

Attribute	1	2	3	4	5	6
sand_perc (%)	40.61	55.12	46.26	28.04	43.62	40.43
silt_perc (%)	16.95	16.29	15.43	24.78	17.61	18.14
clay_perc (%)	41.61	29.08	38.56	47.24	38.14	40.97
org_carbon_content (g/kg)	16.23	8.31	10.94	18.68	10.81	25.97
bedrock_depth (cm)	2088.39	2658.53	2445.82	939.04	3033.18	2278.64
water_table_depth (cm)	2587.26	1628.22	3785.43	2112.28	1634.95	2480.85

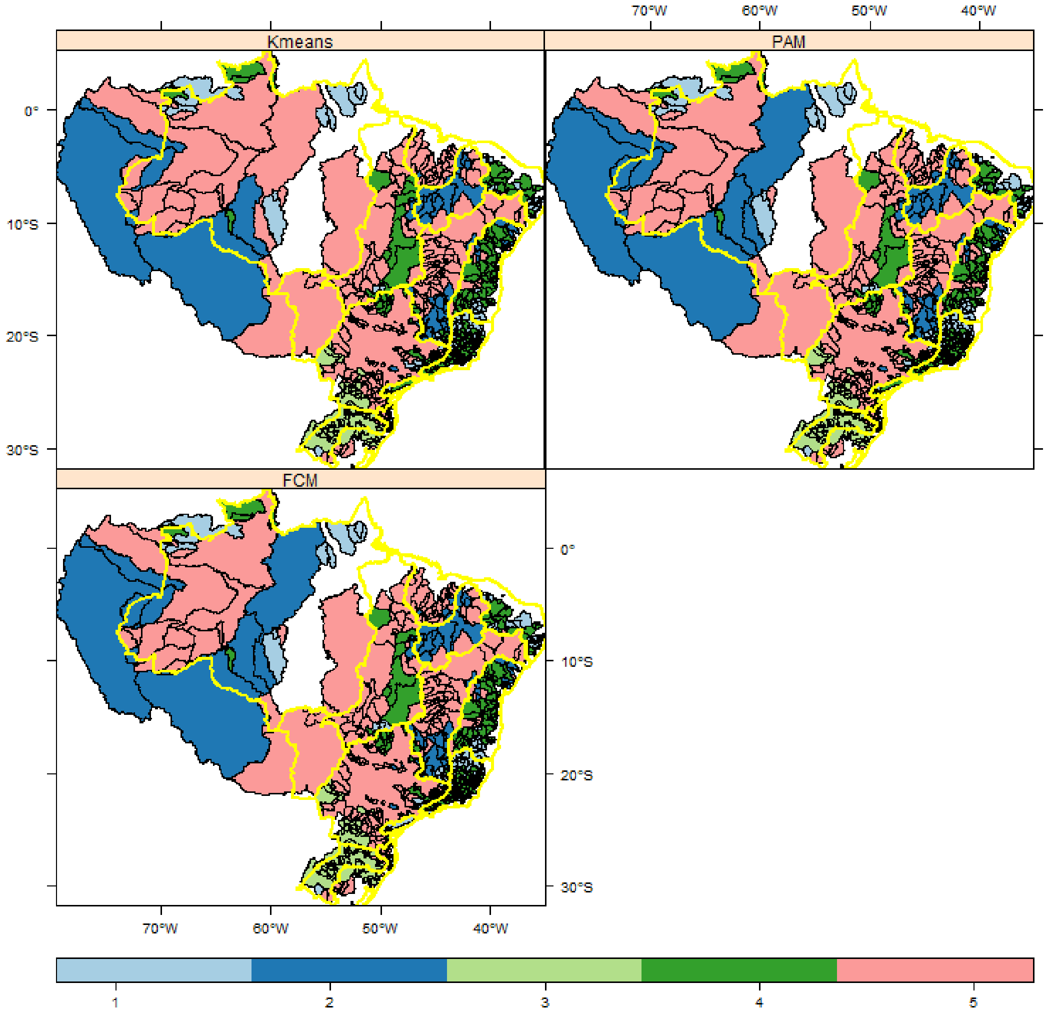
Source: Prepared by the author.

3.3.3.4 Geology Clustering

Upon analyzing Figure 6, which depicts the maps of geology data clustering, a notable level of concordance among the methods is observed. Cluster 5 predominates in six out of twelve hydrological regions. However, the regions "Atlantico NE Oriental," "Parnaíba," "Atlantico Leste," "Atlantico Sudeste," "Uruguai," and "Atlantico Sul" are characterized by a predominance of clusters 4, 2, 4, 4, 3, and 3, respectively.

Cluster 1 is predominantly located in the northern part of the "Atlantico Sudeste" region and in certain basins within the "Amazon" region. Additionally, Cluster 2 is present not only in the "Parnaiba" region but also in the upper São Francisco and upper Amazon basins.

Figure 6 – Maps with clustered basins using geology data



Source: Prepared by the author.

The Tables 22 to 24 contains the average values to each attribute to the cluster developed to each clustering method.

Cluster 1 is characterized by basins exhibiting a high percentage of plutonic rock and a notable presence of metamorphic rock, along with a low porosity value. Cluster 2 encompasses basins with a significant proportion of siliciclastic sedimentary rocks and carbonate rocks, accompanied by a high porosity value. Basins within Cluster 3 predominantly consist of volcanic rocks.

Cluster 4 includes basins predominantly composed of metamorphic rocks and a lower occurrence of plutonic rocks. The dominant Cluster 5 across Brazil is associated with basins featuring a high percentage of siliciclastic sedimentary rocks and exhibiting the highest average porosity.

Table 22 – Geology attribute average values in cluster applying K-means (siliciclastic_sedimentary: Percentage of Siliciclastic Sedimentary Rocks, carbonate: Percentage of Carbonate-Rich Sedimentary Rocks, plutonics: Percentage of Plutonic Rocks, volcanics: Percentage of Volcanic Rocks, metamorphic: Percentage of Metamorphic Rocks, geol_porosity: Subsurface porosity of the catchment, geol_permeability: Subsurface permeability)

Attribute	1	2	3	4	5
siliciclastic_sedimentary (%)	0.00	20.53	0.00	0.00	77.76
carbonate (%)	0.00	48.96	0.00	0.00	0.00
plutonics (%)	57.91	0.00	0.00	9.88	0.00
volcanics (%)	0.00	0.00	99.98	0.00	0.00
metamorphics (%)	33.46	0.00	0.00	68.84	0.00
geol_porosity (adm)	0.01	0.17	0.09	0.02	0.18
geol_permeability (m²)	-13.45	-13.17	-12.72	-13.27	-13.10

Source: Prepared by the author.

Table 23 – Geology attribute average values in cluster applying PAM (siliciclastic_sedimentary: Percentage of Siliciclastic Sedimentary Rocks, carbonate: Percentage of Carbonate-Rich Sedimentary Rocks, plutonics: Percentage of Plutonic Rocks, volcanics: Percentage of Volcanic Rocks, metamorphic: Percentage of Metamorphic Rocks, geol_porosity: Subsurface porosity of the catchment, geol_permeability: Subsurface permeability)

Attribute	1	2	3	4	5
siliciclastic_sedimentary (%)	0.00	21.72	0.00	0.00	76.78
carbonate (%)	0.00	47.02	0.00	0.00	0.00
plutonics (%)	56.92	0.00	0.00	9.88	0.00
volcanics (%)	0.00	0.00	100.00	0.00	0.00

metamorphics (%)	34.41	0.00	0.00	70.69	0.00
geol_porosity (adm)	0.01	0.17	0.09	0.02	0.17
geol_permeability (m^2)	-13.44	-13.10	-12.71	-13.25	-13.14

Source: Prepared by the author.

Table 24 – Geology attribute average values in cluster applying FCM (siliciclastic_sedimentary: Percentage of Siliciclastic Sedimentary Rocks, carbonate: Percentage of Carbonate-Rich Sedimentary Rocks, plutonics: Percentage of Plutonic Rocks, volcanics: Percentage of Volcanic Rocks, metamorphic: Percentage of Metamorphic Rocks, geol_porosity: Subsurface porosity of the catchment, geol_permeability: Subsurface permeability)

Attribute	1	2	3	4	5
siliciclastic_sedimentary (%)	0.00	23.20	0.00	0.00	78.96
carbonate (%)	0.00	45.69	0.00	0.00	0.00
plutonics (%)	56.33	0.00	0.00	7.70	0.00
volcanics (%)	0.00	0.00	100.00	0.00	0.00
metamorphics (%)	35.27	0.00	0.00	71.00	0.00
geol_porosity (adm)	0.01	0.17	0.09	0.02	0.18
geol_permeability (m^2)	-13.46	-13.17	-12.71	-13.26	-13.11

Source: Prepared by the author.

3.3.3.5 Hydrology Clustering

Figure 7 presents the maps with clustered basins using hydrology data. In the Amazon region, Cluster 1 represents the basins of the Amazon River, while Cluster 3 denotes the peripheral Amazon basins, and Cluster 8 corresponds to the southernmost basin.

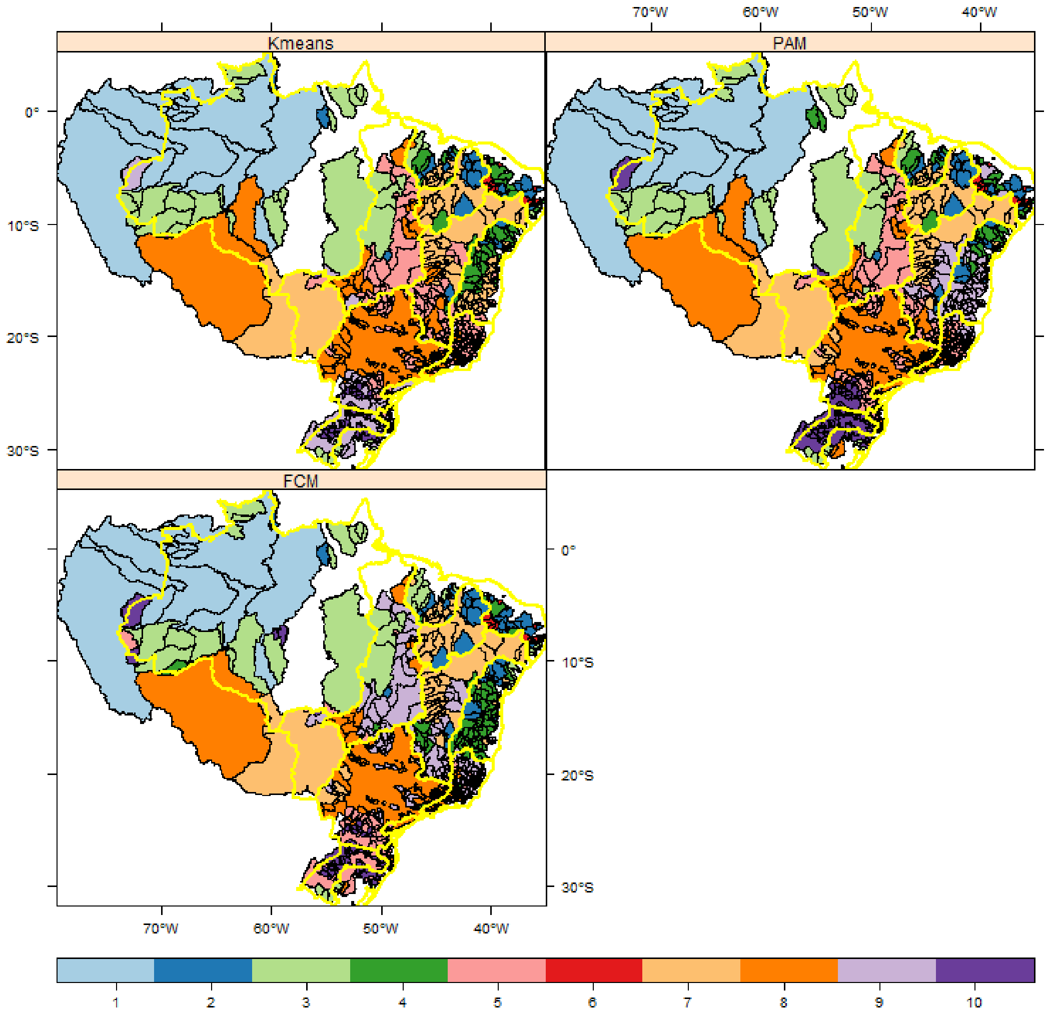
The "Tocantins-Araguaia" region is partitioned into Cluster 3 in the western region, Cluster 8 in the extreme north and south, and Cluster 5 (Kmeans and PAM) or Cluster 9 (FCM) in the eastern region. The "Atlantico NE Ocidental" region exhibits substantial similarity across methods and is divided into Clusters 2, 3, 4, and 7. Within the "Parnaíba" region, Cluster 7 predominates in the upper Parnaíba River, while Cluster 2 prevails in the lower section.

The "Atlantico NE Oriental" region is divided among Clusters 2, 4, and 6. In the "São

Francisco" region, all three methods consistently identify the western section and lower São Francisco as part of Cluster 7. However, the east medium São Francisco and upper São Francisco are assigned to different clusters: 4, 5, 8, and 9. For the "Atlantico Leste" region, Kmeans clustering divides it between Clusters 4 and 7, PAM clustering between Clusters 4 and 9, and FCM clustering between Clusters 2 and 4. In the "Atlantico Sudeste" region, Kmeans clustering delineates it between Clusters 5 and 8, PAM clustering between Clusters 5, 7, and 9, and FCM clustering between Clusters 8 and 9. The northern part of the "Paraná" region is divided into Clusters 5 and 8 in the Kmeans result, Clusters 5, 7, and 8 in the PAM result, and Clusters 7, 8, and 9 in the FCM result.

The southern part of the "Paraná" region includes Cluster 9 and Cluster 10 in the Kmeans result, Cluster 10 in the PAM result, and Cluster 5 and Cluster 10 in the FCM result.

Figure 7 – Maps with clustered basins using hydrology data



Source: Prepared by the author.

Tables 25 to 27 provide the average values of each attribute for the clusters generated by each clustering method. Cluster 1 comprises basins characterized by the highest mean daily discharge, Q5, and Q95, as well as low frequencies and durations of extreme events (dry and humid). These attributes indicate a consistent flow pattern in the basins.

Cluster 2 represents basins with low mean daily discharge, Q5, and Q95, accompanied by a high frequency of dry days and a significant occurrence of humid days. This pattern suggests basins with precipitation concentrated within a limited time of the year.

Cluster 3 predominantly includes basins within the Amazon Basin, which exhibit lower mean discharge, more heterogeneous precipitation distribution throughout the year, and a notable number of dry days.

Clusters 4, represented similarly by Kmeans and PAM methods, demonstrate high similarity in the average attribute values. They represent basins with low mean daily discharge, a high frequency of dry days, and a significant occurrence of humid days, indicating precipitation concentrated within a specific period of the year. However, FCM indicates basins with higher mean discharge and a lower frequency of extreme events. This distinction is reflected in the map, where Cluster 4 in FCM includes basins located further south, thus exhibiting similarities to subtropical climates.

Cluster 5 in the Kmeans and PAM methods represents basins east of the "Tocantins-Araguaia" region. These basins exhibit moderate values of mean daily discharge, Q5, and Q95, along with a low frequency of extreme events. FCM indicates a similar behavior for Cluster 5 but with slightly higher values of discharge attributes.

Cluster 6 denotes basins in the northeastern extreme of Brazil, characterized by the lowest discharge values and the highest frequency of extreme events. Cluster 7 represents basins with low discharge values and a low frequency of extreme events, indicating a consistent but relatively low discharge throughout the year. Cluster 8 represents basins with moderate discharge values and a low frequency of extreme events, signifying a steady medium discharge throughout the year.

Cluster 9, as indicated by Kmeans and FCM, exhibits similarities in attribute values, indicating basins with a significant mean daily discharge, Q95, and a low frequency of extreme events. However, PAM's Cluster 9 represents basins with lower discharge values and a more pronounced frequency of dry days. Cluster 10 encompasses basins located in the southernmost region of Brazil, with a higher mean daily discharge and Q95, a lower value of Q5, and a significant frequency of dry days. This cluster is more prevalent in the PAM result.

Table 25 – Hydrology attribute average values in cluster applying K-means (q_mean: Mean daily discharge, runnof_ratio: Ratio of mean daily discharge to mean daily precipitation, stream_elas: Streamflow precipitation elasticity, hfd_mean: Mean half-flow date, Q5: 5% flow quantile, Q95: 95% flow quantile, high_q_freq: Frequency of high-flow days, high_q_dur: Average duration of high-flow events, low_q_freq: Frequency of low-flow days, low_q_dur: Average duration of low-flow events, zero_q_freq: Percentage of days with zero discharge)

Attribute	1	2	3	4	5	6	7	8	9	10
q_mean (mm/ day)	4.87	0.16	1.96	0.19	1.19	0.06	0.44	1.49	2.15	2.77
runoff_ ratio (adm)	0.67	0.08	0.40	0.08	0.31	0.04	0.16	0.36	0.48	0.54
stream_ elas (adm)	0.99	2.99	1.67	2.16	2.05	2.85	1.07	1.12	1.77	1.91
hfd_mean (days)	223.6	206.3	204.0	191.9	164.8	211.1	173.8	174.0	165.0	163.9
Q5 (mm/day)	2.03	0.00	0.17	0.01	0.32	0.00	0.17	0.62	0.63	0.37
Q95 (mm/day)	9.68	0.74	6.04	0.64	3.19	0.23	0.98	3.34	5.72	9.15
high_q_ freq (days/ year)	0.00	50.08	3.62	22.27	2.21	138.8	1.00	0.05	1.35	10.38
high_q_ dur (days)	0.00	10.64	2.15	5.68	2.19	60.77	2.00	1.00	1.81	2.12
low_q_ freq (days/	0.83	210.2	83.68	114.8	3.97	269.2	0.00	0.00	1.40	44.50

year)										
low_q_ dur (days)	3.06	50.05	19.27	20.86	6.07	70.94	0.00	0.00	3.77	8.99
zero_q_ freq (%)	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00

Source: Prepared by the author.

Table 26 – Hydrology attribute average values in cluster applying PAM (q_mean: Mean daily discharge, runnof_ratio: Ratio of mean daily discharge to mean daily precipitation, stream_elas: Streamflow precipitation elasticity, hfd_mean: Mean half-flow date, Q5: 5% flow quantile, Q95: 95% flow quantile, high_q_freq: Frequency of high-flow days, high_q_dur: Average duration of high-flow events, low_q_freq: Frequency of low-flow days, low_q_dur: Average duration of low-flow events, zero_q_freq: Percentage of days with zero discharge)

Attribute	1	2	3	4	5	6	7	8	9	10
q_mean (mm/ day)	4.41	0.16	1.96	0.23	1.30	0.06	0.62	1.69	0.59	2.59
runoff_ ratio (adm)	0.66	0.08	0.39	0.09	0.33	0.04	0.19	0.41	0.20	0.54
stream_ elas (adm)	1.00	3.01	1.78	2.05	2.09	2.85	0.68	1.26	1.67	1.90
hfd_mean (days)	223.6	200.2	199.3	223.5	165.5	211.1	179.8	170.4	152.8	163.9
Q5 (mm/day)	1.98	0.00	0.17	0.01	0.37	0.00	0.32	0.67	0.11	0.44
Q95 (mm/day)	9.30	0.74	6.04	0.76	3.54	0.23	1.06	3.99	1.80	8.13
high_q_ freq (days/	0.00	46.63	5.01	18.17	1.80	138.86	0.00	0.40	7.12	6.15

year)										
high_q_dur (days)	0.00	10.23	2.44	5.18	2.17	60.77	0.00	1.37	3.36	2.00
low_q_freq (days/ year)	0.00	209.14	92.20	98.93	3.25	269.24	0.00	0.00	24.14	30.43
low_q_dur (days)	0.00	44.16	19.77	19.73	5.80	70.94	0.00	0.00	12.32	7.98
zero_q_freq (%)	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00

Source: Prepared by the author.

Table 27 – Hydrology attribute average values in cluster applying FCM (q_mean: Mean daily discharge, runnof_ratio: Ratio of mean daily discharge to mean daily precipitation, stream_elas: Streamflow precipitation elasticity, hfd_mean: Mean half-flow date, Q5: 5% flow quantile, Q95: 95% flow quantile, high_q_freq: Frequency of high-flow days, high_q_dur: Average duration of high-flow events, low_q_freq: Frequency of low-flow days, low_q_dur: Average duration of low-flow events, zero_q_freq: Percentage of days with zero discharge)

Attribute	1	2	3	4	5	6	7	8	9	10
q_mean (mm/ day)	4.87	0.14	1.29	0.52	1.87	0.06	0.58	1.67	1.16	2.70
runoff_ratio (adm)	0.67	0.07	0.29	0.18	0.44	0.04	0.19	0.40	0.30	0.54
stream_elas (adm)	0.99	2.99	1.63	1.91	2.00	2.85	0.71	1.18	1.97	1.84
hfd_mean	223.6	201.1	224.2	153.0	163.2	211.1	179.9	171.4	166.6	164.6

(days)										
Q5 (mm/day)	2.03	0.00	0.08	0.07	0.49	0.00	0.31	0.69	0.32	0.40
Q95 (mm/day)	9.68	0.51	3.71	1.60	5.35	0.23	1.06	3.86	3.04	8.67
high_q_ freq (days/ year)	0.00	40.13	6.43	11.08	1.70	138.86	0.00	0.38	1.65	8.22
high_q_ dur (days)	0.00	8.91	2.91	4.36	2.17	60.77	0.00	1.27	2.00	2.05
low_q_ freq (days/ year)	0.83	200.36	68.79	42.84	4.25	269.24	0.00	0.00	1.85	39.77
low_q_ dur (days)	3.06	39.12	18.57	17.10	6.60	70.94	0.00	0.00	4.00	8.58
zero_q_ freq (%)	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00

Source: Prepared by the author.

3.3.4 Cluster Performance Metrics

Upon analyzing Table 28, which display the average silhouette values for raw and PCA-transformed data across various clustering methods and classes, it becomes evident that applying PCA to the data leads to an overall increase in average silhouette values. This improvement is particularly prominent in the climate data, as shown in Table 42.

Specifically, when examining the geology class, the average silhouette value for Kmeans is considerably lower at 0.34 compared to PAM and FCM, which both have values of 0.49.

Consequently, the results obtained from Kmeans were not considered for generating the multi-method clustering, and only the results from PAM and FCM were taken into account.

Table 28 – Average Silhouette Values to Climate data

Average Silhouette Width						
		Climate	Land cover	Soil	Geology	Hydrology
K-MEANS	Raw Data	0.34	0.29	0.3	0.28	0.26
	PCA Data	0.42	0.35	0.31	0.34	0.31
PAM	Raw Data	0.33	0.24	0.24	0.43	0.22
	PCA Data	0.41	0.3	0.34	0.49	0.30
FUZZY	Raw Data	0.33	0.3	0.3	0.43	0.21
	PCA Data	0.41	0.32	0.35	0.49	0.28

Source: Prepared by the author.

3.3.5 Multi-method Clustering

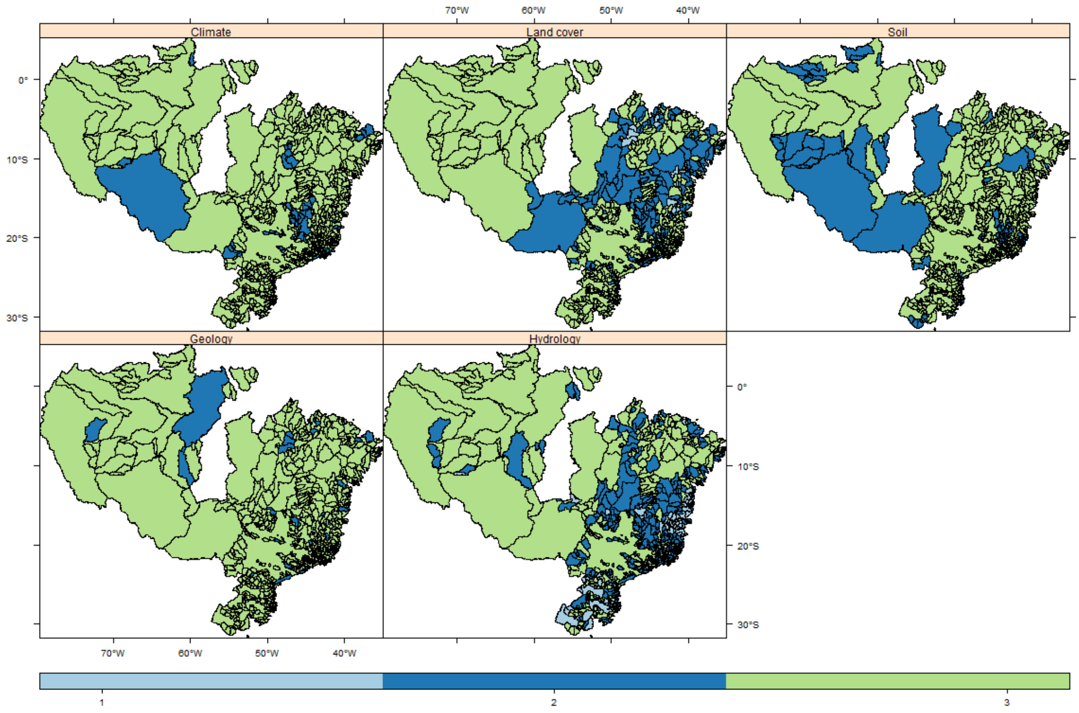
Table 29 presents the distribution of basin area proportions across various levels of concordance. The findings reveal that climate data demonstrated a higher degree of concordance, suggesting lower uncertainty among the clustering methods. Conversely, land cover and hydrological data exhibited higher levels of uncertainty, indicating the need for a multi-method approach to mitigate such uncertainty. Consequently, it can be deduced that climate data did not significantly benefit from the utilization of different clustering methods, whereas land cover and hydrology data showed greater potential for improvement through the adoption of a multi-method approach.

Table 29 – Average Silhouette Values to Climate data

Concordance	Climate	Land cover	Soil	Geology	Hydrology
1	0.00%	3.63%	0.00%	0.00%	3.96%
2	6.83%	31.40%	18.39%	16.65%	26.22%
3	93.17%	64.96%	81.60%	83.35%	69.82%

Source: Prepared by the author.

Figure 8 – Maps with basins concordance number to each class



.

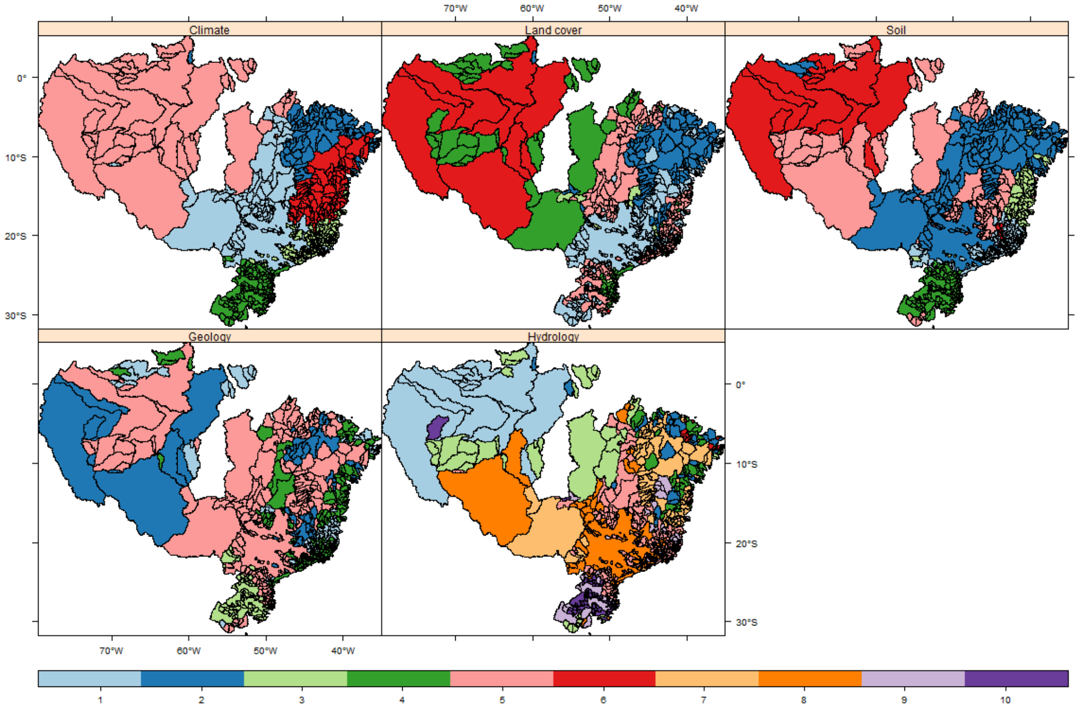
Source: Prepared by the author.

Figure 9 illustrates the final clustering maps for each class. The multi-method maps of climate data and geology exhibit a high degree of similarity to the maps generated by the three individual methods, indicating a strong concordance among the methods.

The land cover multi-method map aligns with the Kmeans and PAM methods for the basins in the “Paraguai,” “Tocantins-Araguai,” and “Atlantico NE Ocidental” regions. In addition, the “São Francisco” region concurs with the Kmeans and FCM methods, while the remaining regions demonstrate a higher consensus among all three methods. The soil multi-method results closely resemble the outcomes of the Kmeans and FCM methods.

The multi-method map of hydrology data reveals that the basins in the "Tocantins-Araguaia" region align with the Kmeans and FCM methods, while the "São Francisco" region agrees with the PAM method. Furthermore, the "Atlantico Leste," "Uruguai," and "Atlantico Sul" regions exhibit similarities to the Kmeans method.

Figure 9 – Multi-method clustering maps to each class



Source: Prepared by the author.

3.3.6 Discussions and Conclusions

This study introduces a novel approach for catchment clustering that tackles the uncertainties associated with classification methods and the selection of an optimal number of clusters. Instead of relying on a single best-performing method, the proposed methodology emphasizes consensus among multiple clustering methods while addressing their divergences. The methodology was specifically applied to the CAMELS-BR dataset in Brazil.

The clustering methodology proposed in this study involved organizing the data into attribute classes, applying Principal Components Analysis (PCA) for dimensionality reduction, estimating the optimal number of clusters for each class using performance metrics, applying different clustering methods to each class with the determined optimal number of clusters, comparing the performance of the clustering methods using quality metrics, evaluating the impact of PCA on clustering performance, assessing the concordance between clustering methods within each catchment, and combining the results of the clustering methods to propose a combined clustering conjecture.

In the field of hydrology, it is common for researchers to utilize a single clustering approach when analyzing the hydrological characteristics of basins (Jehn et al., 2020; Latt, 2014; Sawicz et al., 2014). Alternatively, some studies have explored the use of different clustering methods and compared their respective outcomes (Boscarello et al., 2016), or employed diverse clustering techniques to derive hydrological classifications and subsequently evaluated performance metrics to determine the most optimal method (Shargui, 2017). However, the existing literature does not appear to have investigated the establishment of a consensus among these methods to determine the classification, as proposed in the methodology of this study. The concept of consensus, in this context, aims to derive a classification based on the collective agreement among multiple methods, rather than relying solely on a single method and disregarding others based solely on performance metrics. This novel approach seeks to address the limitations associated with individual clustering methods and promote a more comprehensive and robust methodology for classification.

The authors JEHN et al. (2020) conducted hydrological clustering and aimed to compare it with the Koppen-Geiger climate classification, anticipating a high degree of similarity due to the influential role of climate in hydrological processes. However, their findings revealed significant disparities between the two classifications. Similarly, in the present study, when examining the classification of hydrological signatures in relation to climate, some concurrence was observed in the southern region and the Amazon River basins, but not in the other regions.

The classification of hydrographic regions by the National Water Agency (ANA, year), which forms the basis for the management instruments of Brazil's basins, heavily relies on geographical boundaries. As discovered by several authors (JEHN et al., 2020; Burn, 2017; Santos et al., 2013), geographical boundaries are inadequate in capturing the heterogeneity of hydrological features. Consequently, the hydrographic regions defined by ANA demonstrate low concordance with the hydrological clustering performed in this study.

This study represents the pioneering utilization of the CAMELS-BR database for the classification of Brazilian basins, thus highlighting the significance of comprehensive databases, such as CAMELS, that facilitate the application of classification techniques relying on a large number of attributes.

As a suggestion for future research, it is recommended to explore more detailed approaches to consensus that consider the varying performance of clusters. Also, a streamflow regionalization study that incorporates the hydrological classification showed in the study.

REFERENCES

AGARWAL, Ankit et al. Hydrologic regionalization using wavelet-based multiscale entropy method. **Journal of Hydrology**, v. 538, p. 22-32, 2016.

ANGUS WEBB, J. et al. Bayesian clustering with AutoClass explicitly recognises uncertainties in landscape classification. **Ecography**, v. 30, n. 4, p. 526-536, 2007.

BASU, Bidroha; SRINIVAS, V. V. Regional flood frequency analysis using entropy-based clustering approach. **Journal of Hydrologic Engineering**, v. 21, n. 8, p. 04016020, 2016.

BATOOL, Fatima; HENNIG, Christian. Clustering with the average silhouette width. **Computational Statistics & Data Analysis**, v. 158, p. 107190, 2021.

BESKOW, Samuel et al. Artificial intelligence techniques coupled with seasonality measures for hydrological regionalization of Q90 under Brazilian conditions. **Journal of Hydrology**, v. 541, p. 1406-1419, 2016.

BEZDEK, James C.; EHRLICH, Robert; FULL, William. FCM: The fuzzy c-means clustering algorithm. **Computers & geosciences**, v. 10, n. 2-3, p. 191-203, 1984.

BHOLOWALIA, Purnima; KUMAR, Arvind. EBK-means: A clustering technique based on elbow method and k-means in WSN. **International Journal of Computer Applications**, v. 105, n. 9, 2014.

BORK, Carina K. et al. Minimum streamflow regionalization in a Brazilian watershed under different clustering approaches. **Anais da Academia Brasileira de Ciências**, v. 93, 2021.

BROWN, Alice E. et al. Impact of forest cover changes on annual streamflow and flow duration curves. **Journal of Hydrology**, v. 483, p. 39-50, 2013.

Burn, D. H. (1989). Cluster analysis as applied to regional flood frequency. Journal of Water Resources Planning and Management, 115(5), 567-582.

BURN, Donald H. Cluster analysis as applied to regional flood frequency. **Journal of Water Resources Planning and Management**, v. 115, n. 5, p. 567-582, 1989.

BUTTLE, Jim. Mapping first-order controls on streamflow from drainage basins: the T3 template. **Hydrological Processes: An International Journal**, v. 20, n. 15, p. 3415-3422, 2006.

CHAGAS, Vinícius BP et al. CAMELS-BR: Hydrometeorological time series and landscape attributes for 897 catchments in Brazil. **Earth System Science Data**, v. 12, n. 3, p. 2075-2096, 2020.

CORPORAL-LODANGCO, Irenea L. et al. Cluster analysis of North Atlantic tropical cyclones. **Procedia Computer Science**, v. 36, p. 293-300, 2014.

CROCHEMORE, Louise et al. Lessons learnt from checking the quality of openly accessible river flow data worldwide. **Hydrological Sciences Journal**, v. 65, n. 5, p. 699-711, 2020.

CUI, Mengyao et al. Introduction to the k-means clustering algorithm based on the elbow method. **Accounting, Auditing and Finance**, v. 1, n. 1, p. 5-8, 2020.

DINH, Duy-Tai; FUJINAMI, Tsutomu; HUYNH, Van-Nam. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In: **Knowledge and Systems Sciences: 20th International Symposium, KSS 2019, Da Nang, Vietnam, November 29–December 1, 2019, Proceedings 20**. Springer Singapore, 2019. p. 1-17.

EL-MANDOUH, Amira M. et al. Optimized K-means clustering model based on gap statistic. **International Journal of Advanced Computer Science and Applications**, v. 10, n. 1, 2019.

ET-TALEBY, Abdelilah; BOUSSETTA, Mohammed; BENSLIMANE, Mohamed. Faults detection for photovoltaic field based on k-means, elbow, and average silhouette techniques through the segmentation of a thermal image. **International Journal of Photoenergy**, v. 2020, p. 1-7, 2020.

GOKTEPE, A. B.; ALTUN, Selim; SEZER, Alper. Soil clustering by fuzzy c-means algorithm. **Advances in Engineering Software**, v. 36, n. 10, p. 691-698, 2005.

GOYAL, Manish Kumar; GUPTA, Vivek. Identification of homogeneous rainfall regimes in Northeast Region of India using fuzzy cluster analysis. **Water Resources Management**, v. 28, p. 4491-4511, 2014.

HARTMANN, Jens; MOOSDORF, Nils. The new global lithological map database GLiM: A representation of rock properties at the Earth surface. **Geochemistry, Geophysics, Geosystems**, v. 13, n. 12, 2012.

JAIN, Anil K. Data clustering: 50 years beyond K-means. **Pattern recognition letters**, v. 31, n. 8, p. 651-666, 2010.

JEHN, Florian U. et al. Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. **Hydrology and Earth System Sciences**, v. 24, n. 3, p. 1081-1100, 2020.

KASSAMBARA, Alboukadel. **Practical guide to cluster analysis in R: Unsupervised machine learning**. Sthda, 2017.

KAUFMAN, Leonard. Partitioning around medoids (program pam). **Finding groups in data**, v. 344, p. 68-125, 1990.

KUENTZ, Anna et al. Understanding hydrologic variability across Europe through catchment classification. **Hydrology and Earth System Sciences**, v. 21, n. 6, p. 2863-2879, 2017.

LAAHA, Gregor; BLÖSCHL, Günter. A comparison of low flow regionalisation methods—catchment grouping. **Journal of Hydrology**, v. 323, n. 1-4, p. 193-214, 2006.

LATT, Zaw Zaw; WITTENBERG, Hartmut; URBAN, Brigitte. Clustering hydrological homogeneous regions and neural network based index flood estimation for ungauged catchments: an example of the Chindwin River in Myanmar. **Water resources management**, v. 29, p. 913-928, 2015.

LEIBOWITZ, Scott G. et al. Hydrologic landscape characterization for the Pacific Northwest, USA. **JAWRA Journal of the American Water Resources Association**, v. 52, n. 2, p. 473-493, 2016.

LELE, Subhash. Euclidean distance matrix analysis (EDMA): estimation of mean form and mean form difference. **Mathematical Geology**, v. 25, p. 573-602, 1993.

LEY, Rita et al. Catchment classification by runoff behaviour with self-organizing maps (SOM). **Hydrology and Earth System Sciences**, v. 15, n. 9, p. 2947-2962, 2011.

MCDONNELL, Jeffrey J.; WOODS, Ross. On the need for catchment classification. Journal of Hydrology, v. 299, n. 1, p. 2-3, 2004.

MACQUEEN, James et al. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. 1967. p. 281-297.

MARUYAMA, Takeo; KAWACHI, Toshihiko; SINGH, Vijay P. Entropy-based assessment and clustering of potential water resources availability. **Journal of hydrology**, v. 309, n. 1-4, p. 104-113, 2005.

MORON, Vincent et al. Weather types across the Maritime Continent: From the diurnal cycle to interannual variations. **Frontiers in Environmental Science**, v. 2, p. 65, 2015.

NOURANI, Vahid; PARHIZKAR, Masoumeh. Conjunction of SOM-based feature extraction method and hybrid wavelet-ANN approach for rainfall–runoff modeling. **Journal of Hydroinformatics**, v. 15, n. 3, p. 829-848, 2013.

OLDEN, Julian D.; KENNARD, Mark J.; PUSEY, Bradley J. A framework for hydrologic classification with a review of methodologies and applications in ecohydrology. **Ecohydrology**, v. 5, n. 4, p. 503-518, 2012.

POFF, N. LeRoy et al. Placing global stream flow variability in geographic and geomorphic contexts. **River Research and Applications**, v. 22, n. 2, p. 149-166, 2006.
RAO, A. Ramachandra; SRINIVAS, V. V. Regionalization of watersheds by hybrid-cluster analysis. **Journal of Hydrology**, v. 318, n. 1-4, p. 37-56, 2006.

ROUSSEEUW, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, v. 20, p. 53-65, 1987.

SAPUTRA, Danny Matthew; SAPUTRA, Daniel; OSWARI, Liniyanti D. Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method. In: **Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)**. Atlantis Press, 2020. p. 341-346.

SAWICZ, K. A. et al. Characterizing hydrologic change through catchment classification. **Hydrology and Earth System Sciences**, v. 18, n. 1, p. 273-285, 2014.

SAWICZ, Keith et al. Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. **Hydrology and Earth System Sciences**, v. 15, n. 9, p. 2895-2911, 2011.

SCHRÖTER, Ingmar et al. Estimation of catchment-scale soil moisture patterns based on terrain data and sparse TDR measurements using a fuzzy C-means clustering approach. **Vadose Zone Journal**, v. 14, n. 11, 2015.

SETIADY, DANIEL ADRIAN. **IMPLEMENTATION OF K-MEANS ALGORITHM ELBOW METHOD AND SILHOUETTE COEFFICIENT FOR RAINFALL CLASSIFICATION**. 2021. Tese de Doutorado. Universitas Katholik Soegijapranata Semarang.

SHARGHI, Elnaz et al. Application of different clustering approaches to hydroclimatological catchment regionalization in mountainous regions, a case study in Utah State. **Journal of Mountain Science**, v. 15, n. 3, p. 461-484, 2018.

SHAHARUDIN, S. M. et al. Identification of rainfall patterns on hydrological simulation using robust principal component analysis. **Indonesian Journal of Electrical Engineering and Computer Science**, v. 11, n. 3, p. 1162-1167, 2018.

SIVA, G. Samba; RAO, V. Srinivasa; BABU, D. Ratna. Cluster Analysis Approach to Study the Rainfall Pattern in Visakhapatnam District. **Weekly Science Research Journal**, v. 1, p. 31, 2014.

SINGH, Shailesh Kumar et al. Nonparametric catchment clustering using the data depth function. **Hydrological Sciences Journal**, v. 61, n. 15, p. 2649-2667, 2016.

SRINIVAS, V. V. et al. Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. **Journal of Hydrology**, v. 348, n. 1-2, p. 148-166, 2008.

SNELDER, T. H.; J. BOOKER, D. Natural flow regime classifications are sensitive to definition procedures. **River Research and Applications**, v. 29, n. 7, p. 822-838, 2013.

TIBSHIRANI, Robert; WALTHER, Guenther; HASTIE, Trevor. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 63, n. 2, p. 411-423, 2001.

TONGAL, Hakan; SIVAKUMAR, Bellie. Cross-entropy clustering framework for catchment classification. **Journal of Hydrology**, v. 552, p. 433-446, 2017.

WINTER, Thomas C. The concept of hydrologic landscapes 1. **JAWRA Journal of the American Water Resources Association**, v. 37, n. 2, p. 335-349, 2001.

ZHANG, Yongqiang et al. Predicting hydrological signatures in ungauged catchments using spatial interpolation, index model, and rainfall–runoff modelling. **Journal of Hydrology**, v. 517, p. 936-948, 2014.

APPENDIX A – ADDITIONAL OPTIMAL NUMBER OF CLUSTERES RESULTS

Table A. 1 – Optimal Number of Clusters results to methods applying PAM in Climate data

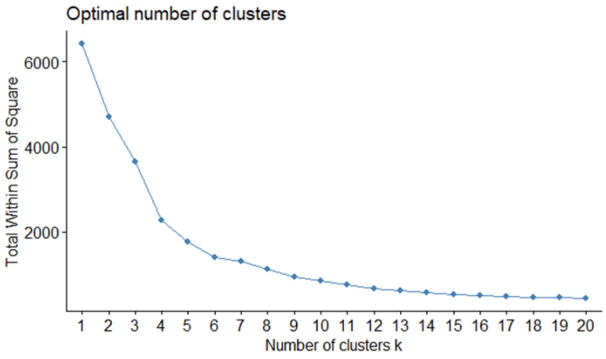
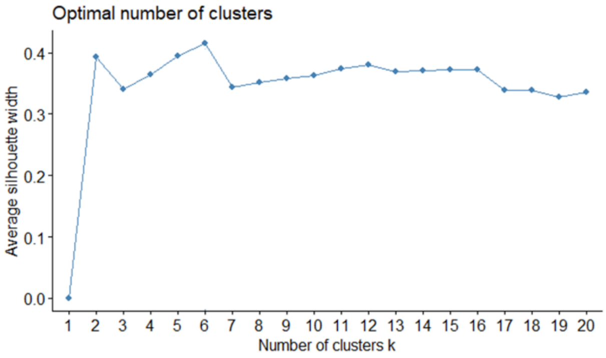
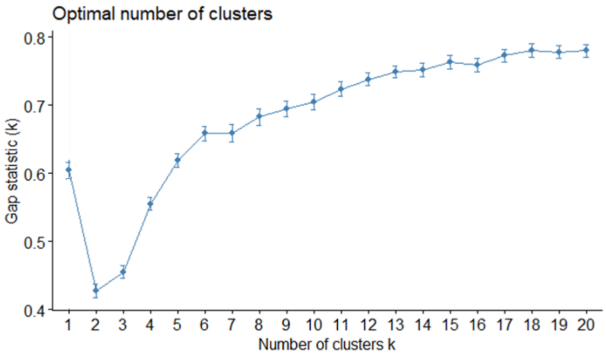
Elbow Method	
Silhouette Width	
Gap Statistic	

Table A. 2 – Optimal Number of Clusters results to methods applying FCM in Climate data

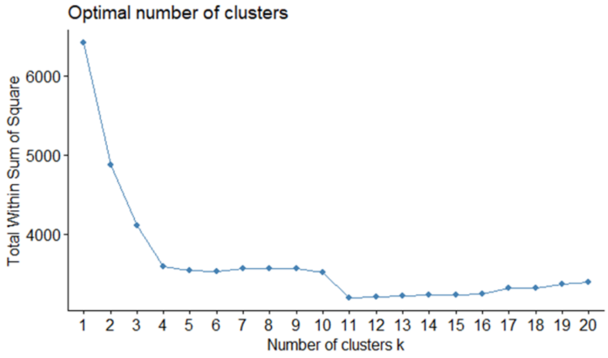
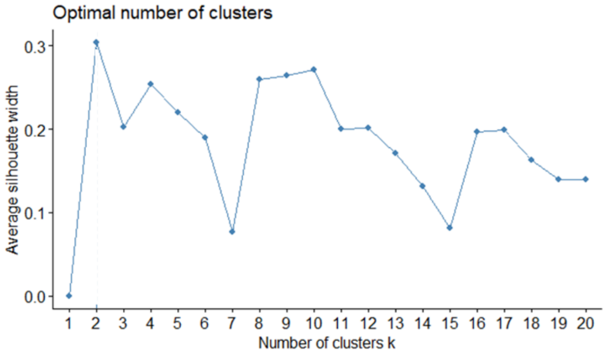
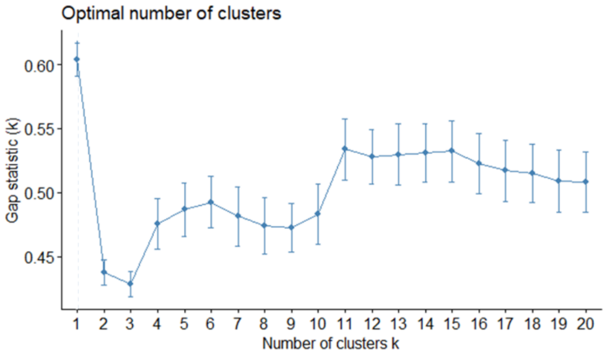
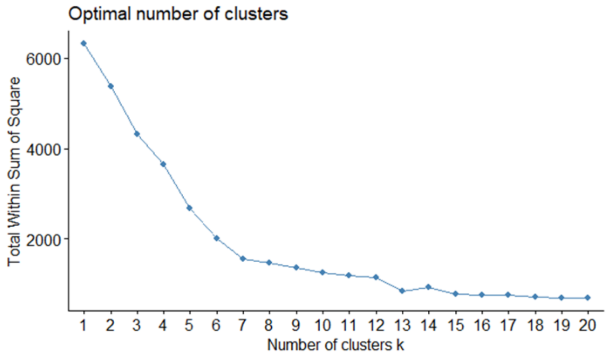
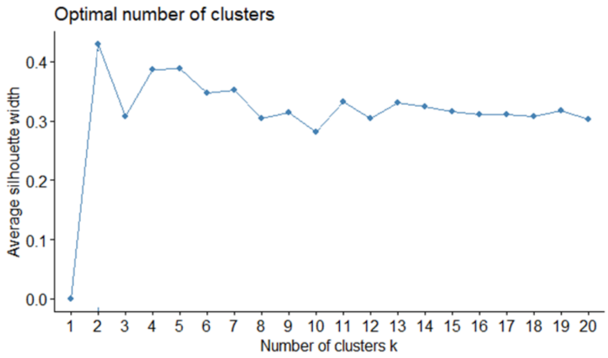
Elbow Method	
Silhouette Width	
Gap Statistic	

Table A. 3 – Optimal Number of Clusters results to methods applying K-means in Land Cover data

Elbow Method



Silhouette Width



Gap Statistic

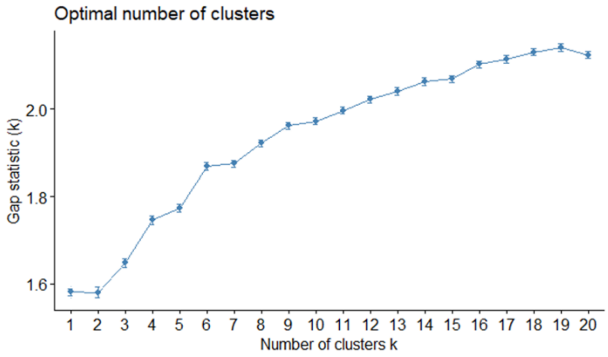


Table A. 4 – Optimal Number of Clusters results to methods applying PAM in Land Cover data

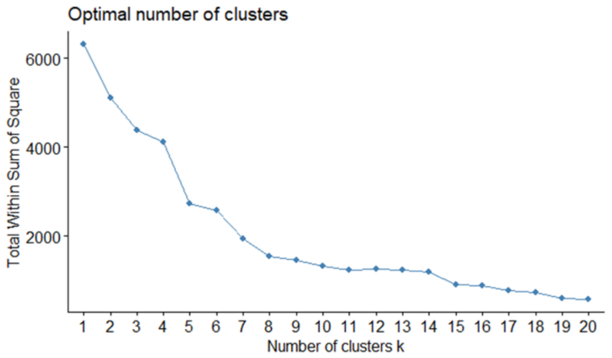
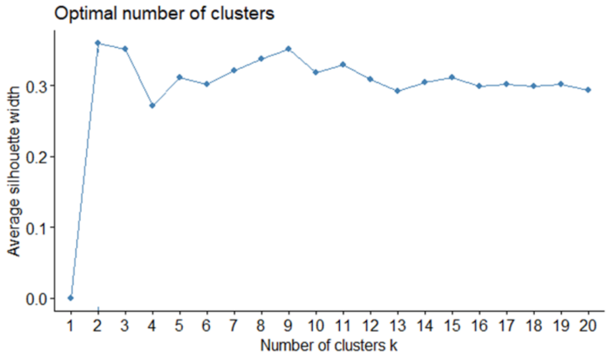
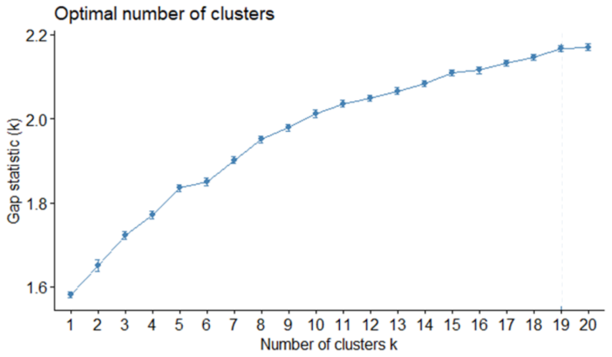
Elbow Method	
Silhouette Width	
Gap Statistic	

Table A. 5 – Optimal Number of Clusters results to methods applying FCM in Land Cover data

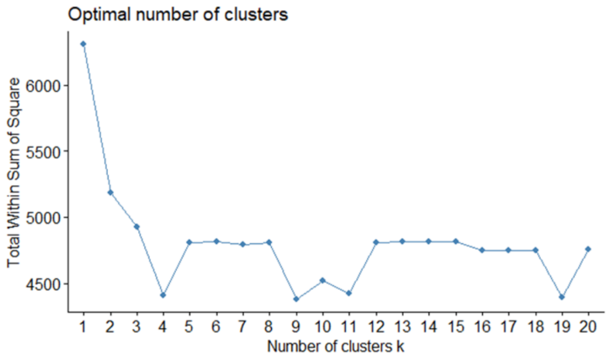
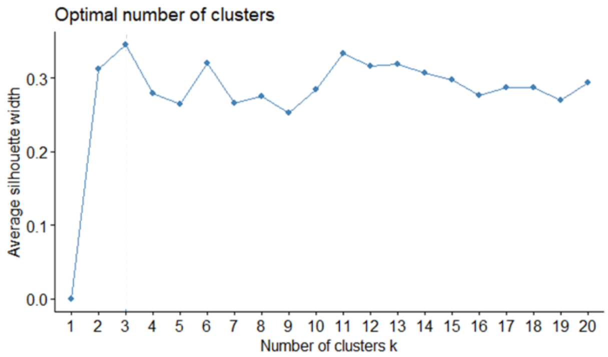
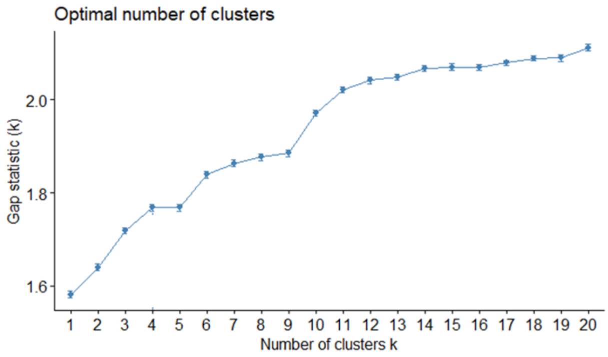
Elbow Method	
Silhouette Width	
Gap Statistic	

Table A. 6 – Optimal Number of Clusters results to methods applying K-means in Soil data

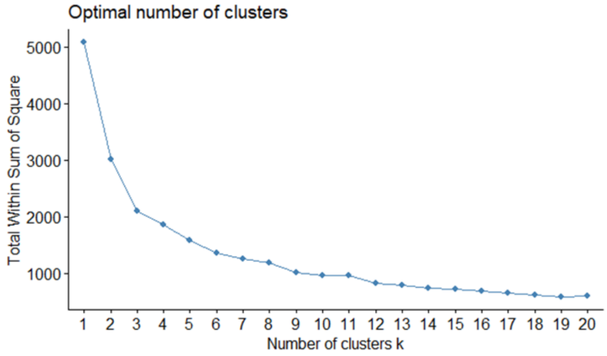
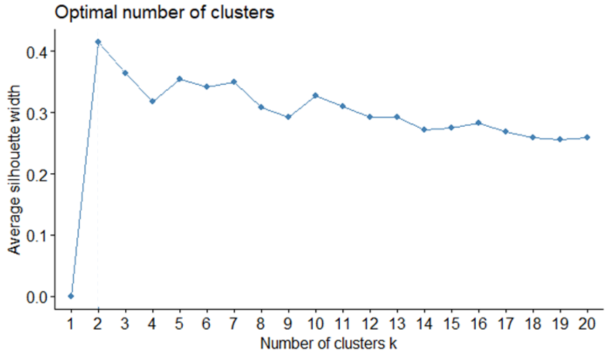
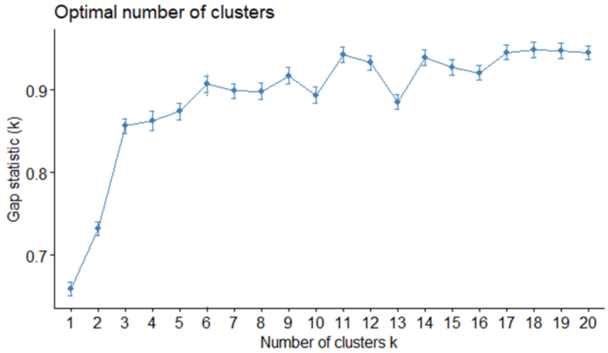
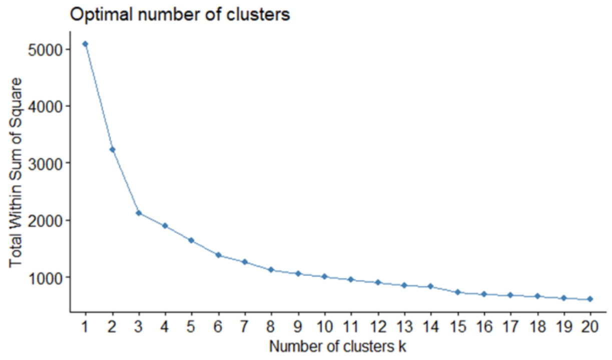
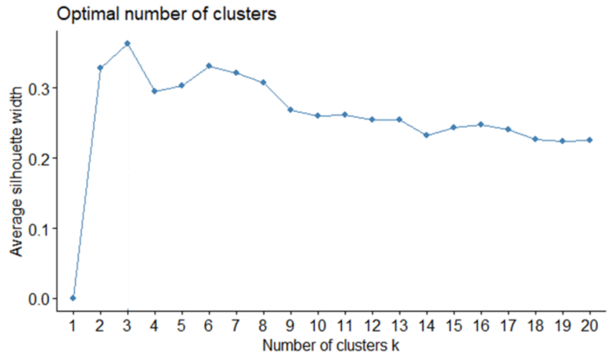
Elbow Method	
Silhouette Width	
Gap Statistic	

Table A. 7 – Optimal Number of Clusters results to methods applying PAM in Soil data

Elbow Method



Silhouette Width



Gap Statistic

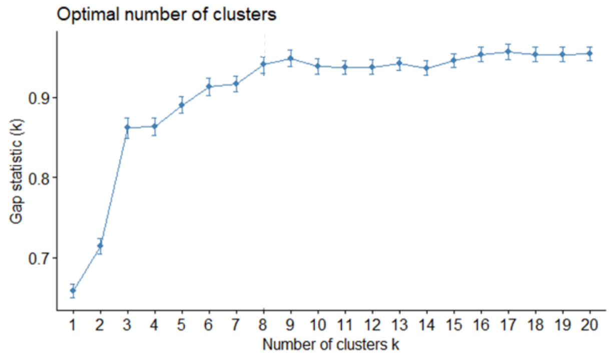


Table A. 8 – Optimal Number of Clusters results to methods applying FCM in Soil data

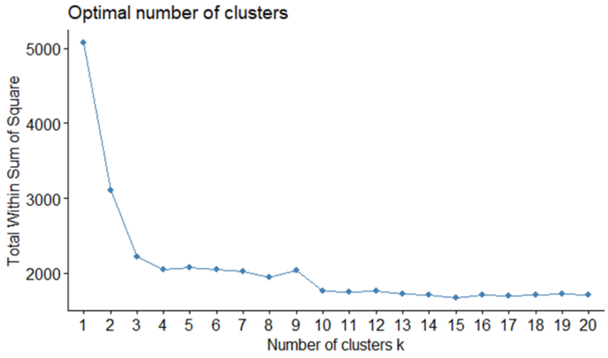
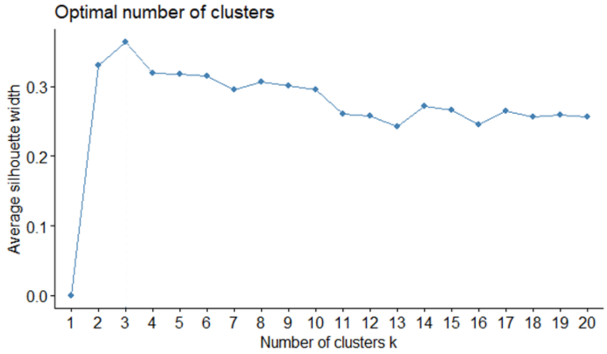
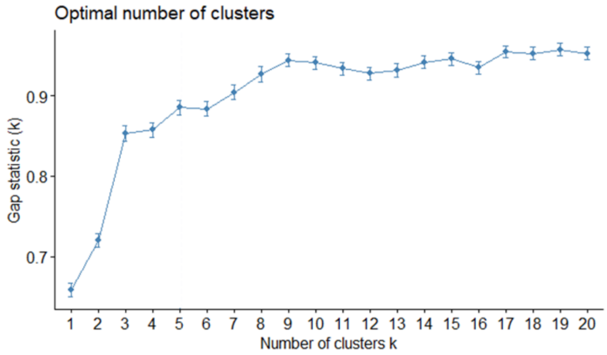
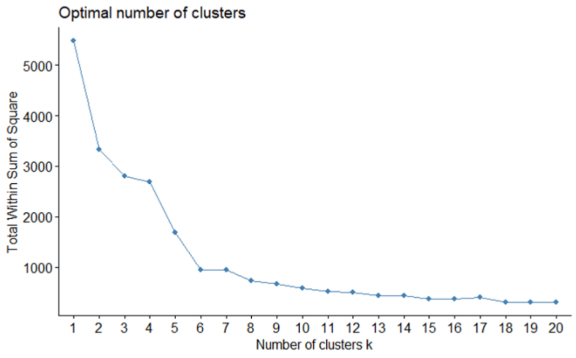
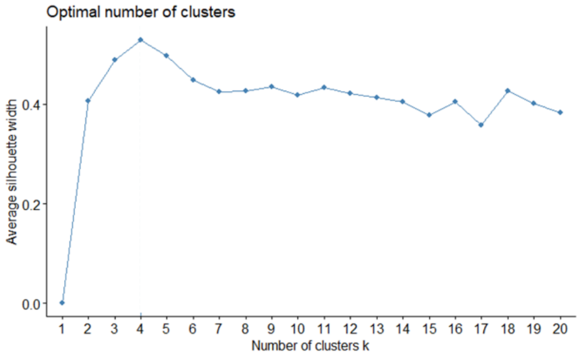
Elbow Method	
Silhouette Width	
Gap Statistic	

Table A. 9 – Optimal Number of Clusters results to methods applying K-means in Geology data

Elbow Method



Silhouette Width



Gap Statistic

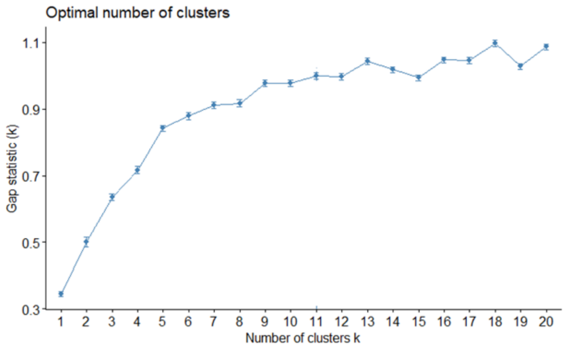


Table A. 10 – Optimal Number of Clusters results to methods applying PAM in Geology oil data

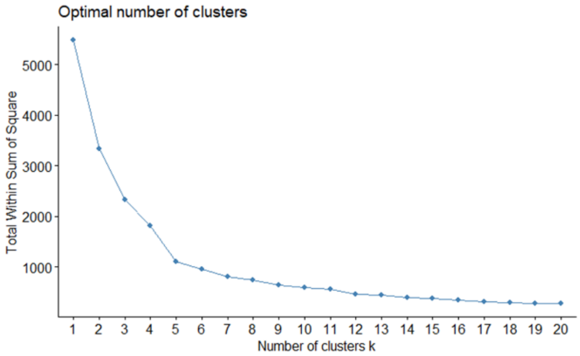
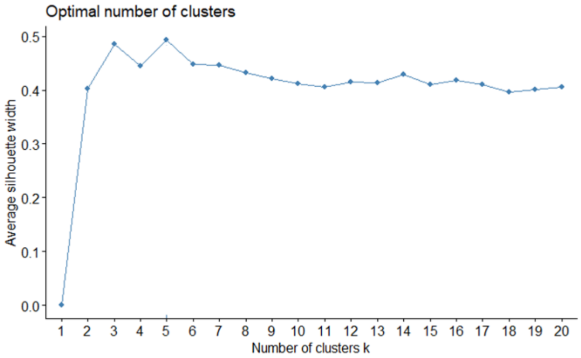
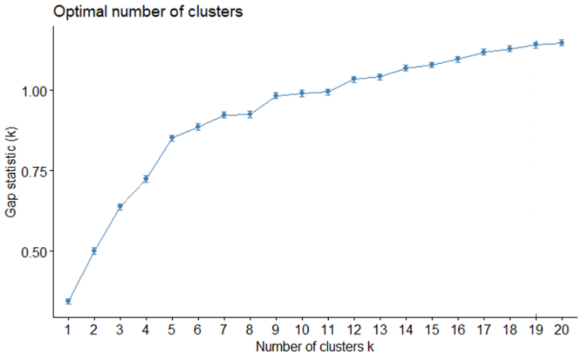
Elbow Method	
Silhouette Width	
Gap Statistic	

Table A. 11 – Optimal Number of Clusters results to methods applying FCM in Geology oil data

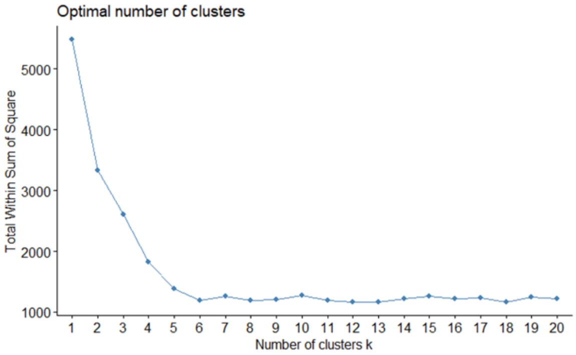
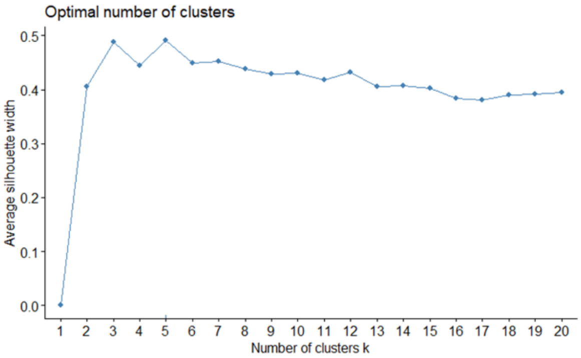
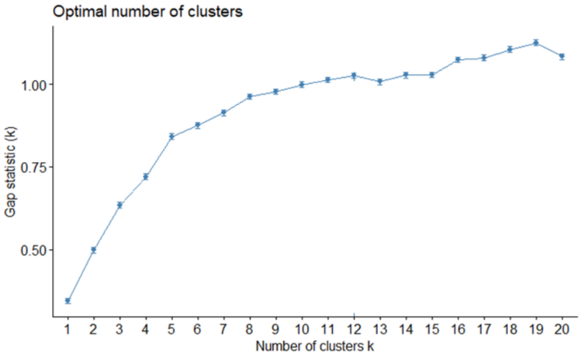
Elbow Method	
Silhouette Width	
Gap Statistic	

Table A. 12 – Optimal Number of Clusters results to methods applying K-means in Hydrology data

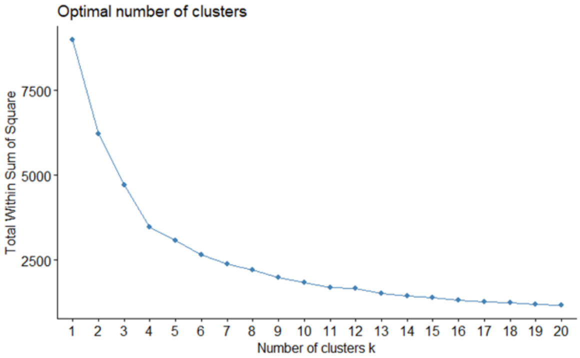
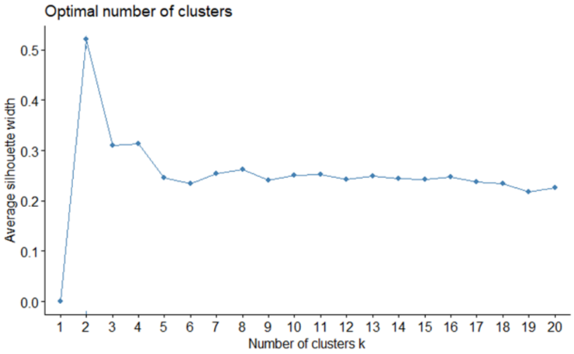
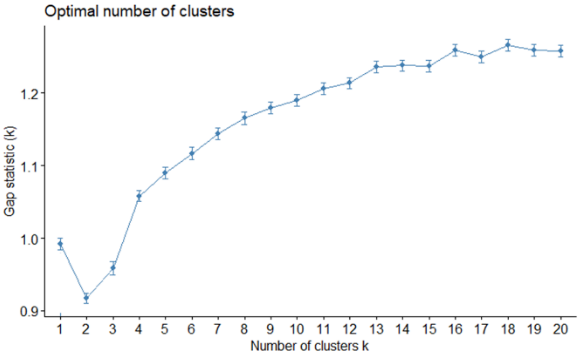
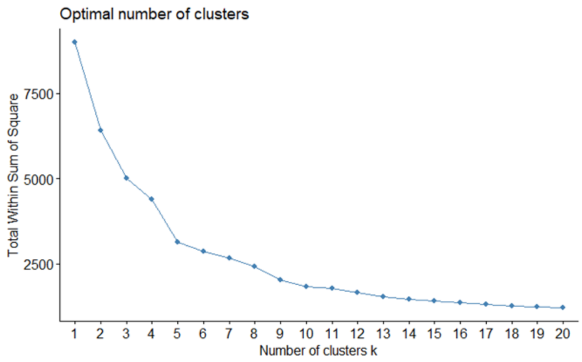
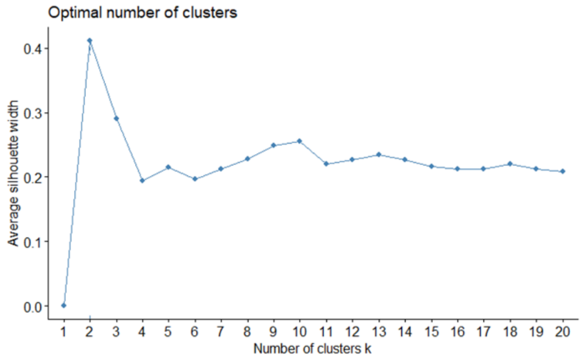
Elbow Method	
Silhouette Width	
Gap Statistic	

Table A. 13 – Optimal Number of Clusters results to methods applying PAM in Hydrology data

Elbow Method



Silhouette Width



Gap Statistic

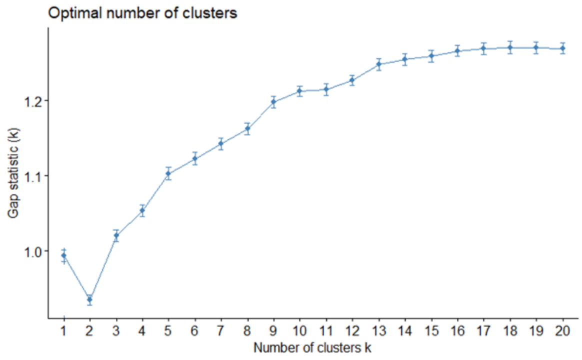
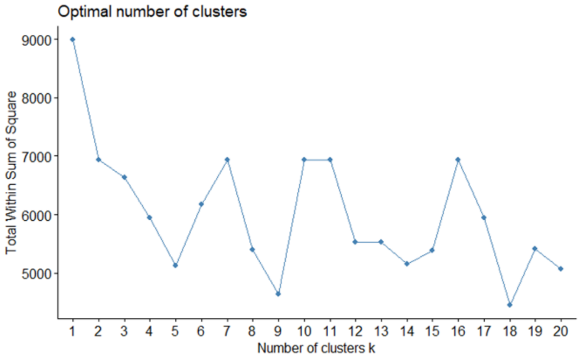
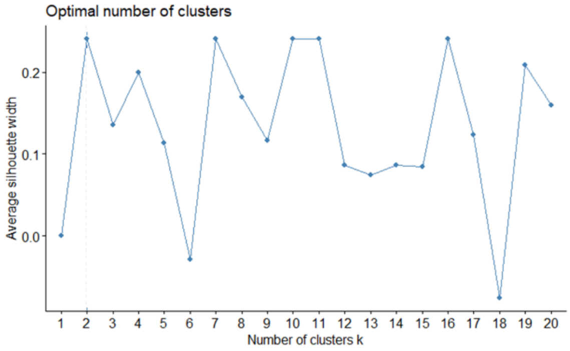


Table A. 14 – Optimal Number of Clusters results to methods applying FCM in Hydrology data

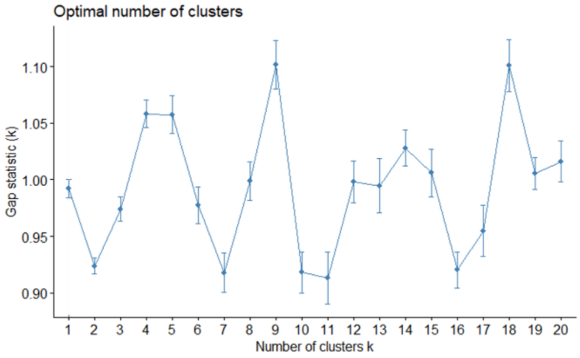
Elbow Method



Silhouette Width



Gap Statistic


Optimal number of clusters
Gap statistic (k)
Number of clusters k