UFSM

CIÊNCIA e NATURA

ISSN 2179-460X
OPEN ACCESS

Ci. e Nat., Santa Maria, v. 47, e85042, 2025• https://doi.org/10.5902/2179460X85042
Submitted: 09/11/23 • Approved: 09/26/24 • Published: 03/14/25

Environment

Clustering of spatio-temporal precipitation patterns in the Legal Amazon using deep convolutional autoencoder

Clusterização de padrões espaço-temporais de precipitação na Amazônia Legal via *deep convolutional autoencoder*

Vander Augusto Oliveira da Silva [I], **Raphael Barros Texeira** [II]

[I] Instituto Federal de Educação, Ciência e Tecnologia do Pará, PA, Brazil
[II] Universidade Federal do Pará, PA, Brazil

ABSTRACT

Identifying patterns in precipitation time series in a given region is fundamental for its socioeconomic development. Many studies on this topic have been carried out in Brazil, mainly in the Amazon region. This research aimed at the development of a computational method for analyzing time series of precipitation using machine learning techniques, aiming at a method capable of extracting complex characteristics from the data, obtaining a map of attributes in low dimensionality for pattern recognition and discovery of homogeneous regions with respect to precipitation in the Legal Amazon. The proposed model is trained to learn the main and most complex characteristics of the original data and present them in low dimensionality in latent space. After training, the observations of the reconstructed data showed good performance as evaluated by the RMSE and NRMSE metric with resulting values equal to 0.06610 and 0.3355 respectively. The result of the low-dimensional representation of the data was analyzed by a clustering structure using hierarchical clustering with Ward's method. This methodology carried out consistent groupings characterizing homogeneous regions in relation to precipitation data. In this way, demonstrating that the representation in low dimensionality carried the main characteristics of the time series of the studied data.

Keywords: Machine learning; Deep convolutional autoencoder; Clustering, Pattern recognition; Precipitation time series

RESUMO

Identificar padrões em séries temporais de precipitação em uma determinada região é fundamental para seu desenvolvimento socioeconômico. Muitos estudos dessa temática foram realizados no Brasil, principalmente na região amazônica. Esta pesquisa objetivou o desenvolvimento de um método computacional para análise de séries temporais de precipitação utilizando técnicas de *machine learning*,

visando um método capaz de realizar a extração de características complexas dos dados, obtendo um mapa de atributos em baixa dimensionalidade para reconhecimento de padrões e descoberta de regiões homogêneas com relação a precipitação da Amazônia Legal. O modelo proposto é treinado para aprender as principais e mais complexas características dos dados originais e apresentá-los em baixa dimensionalidade no espaço latente. Após o treinamento, as observações dos dados reconstruídos apresentaram bom desempenho conforme avaliação da métrica de RMSE e NRMSE com valores resultantes iguais a 0.06610 e 0.3355 respectivamente. O resultado da representação dos dados em baixa dimensão foi analisada por uma estrutura de *clustering* usando aglomerativo hierárquico com método de Ward. Essa metodologia realizou agrupamentos consistentes caracterizando regiões homogêneas com relação aos dados de precipitação. Dessa forma, demonstrando que a representação em baixa dimensionalidade carregava as características principais das séries temporais dos dados estudados.

Palavras-chave: Aprendizado de máquina; Autoencoder convolucional profundo; Agrupamentos; Reconhecimento de padrões; Séries temporais de precipitação

1 INTRODUCTION

Machine Learning (ML), is a field of artificial intelligence (AI) that seeks to model data through the application of algorithms, aiming to learn and perform optimizations with the aim of achieving results that satisfy problem solving, whether through supervised, unsupervised, semi-supervised and reinforcement learning Bhavsar et al. (2017).

This technology has been widely used in several areas of knowledge. The reason for the increase of its popularity is directly linked to its exceptional performance in solving complex problems. For example, ML techniques can be used to identify patterns in diverse data extracting characteristics and deep knowledge of which can be tasks, classification, and the development of intelligent systems Baia & Castro (2018); Essien & Giannetti (2020); Huang et al. (2017); Yin et al. (2020).

Complex feature extraction aims to identify and represent inspired aspects of data that may not be readily observable. These characteristics can be quantitative, such as statistical measures, or qualitative, as standards specific behavior. The goal is to capture the essence of the data and provide valuable insights to support the decisions.

In this context, time series have received special attention. A time series is a sequence of observations collected at regular intervals over time. Time series analysis aims to extract relevant information and patterns data, in order to better understand

the temporal behavior of the phenomena studied. This involves the identification of trends, seasonality, cycles and irregular patterns that may be present in the data over time Sá (2023).

According to Maggioni & Silva (2016), the identification of patterns in time series has been awakening, for some time, interest of the scientific community in solving various problems related to different areas of application, whether area of health, finance, industrial and environmental. For Bailão et al. (2020), pattern recognition is the study of the organize the data and can be performed from objects represented in various types of data.

For Cruz et al. (2016), it is not new in the literature that studies ML, the effort used in the search to find representations data of all nature. For the purpose of extracting resources and discovering knowledge of patterns in time series, methods were proposed and applied.

These applications using this approach have been relevant in these scenarios, taking into account that the used for this type of knowledge discovery are unlabeled raw data, the type of learning employed in these cases is the unsupervised. According to Masci et al. (2018), one of the main objectives of unsupervised learning is resource extraction and discovery of unlabeled data patterns, detecting and removing redundancies and characteristics considered weak and maintaining the strongest characteristics of the original data in good representations.

One of the areas of knowledge that uses machine learning techniques to carry out projects and research is the area of environment Dourado et al. (2013). Usually, these studies are performed using as a data set some kind of time series. Every data set obtained through frequent and sequential measurements over time is considered a time series Esling & Agon (2012).

In the Amazon, precipitation is a very important meteorological component for the accomplishment of various human, environmental, industrial, agricultural, scientific, etc., this natural phenomenon has a high variability in time and space Gonçalves et al. (2017). Therefore, it can be stated that the knowledge and understanding of the rainfall characteristics of the regions are important to enable the planning of the use, management and conservation of water resources.

The socioeconomic impacts resulting from the variation of precipitation make it one of the most relevant climatic variables. Rainfall scarcity can compromise food supply, while excess can cause damage to city infrastructure through flooding. In this sense, a method that assists in the analysis and management of water availability, and consequently in socioeconomic development, is the analysis of the behavior of the precipitation time series Severo et al. (2019).

Considering the relevance of the analysis of precipitation time series, especially in the Amazon context, this study focused on the development of a computational method that used machine learning techniques to analyze characteristics and identify patterns in the time series of precipitation of 268 rainfall stations located in the Legal Amazon. The main objective was not only to identify patterns in the annual rainfall cycle and characterize homogeneous regions in relation to precipitation variability, but fundamentally to develop a method that would to analyze, understand and extract knowledge from time series.

The model adopted in this study for the analysis of precipitation data is based on an autocodifying neural network structure, also known as autoencoder (AE), with convolutional layers and deep learning. This approach is highly effective for learning and acquiring knowledge from input data, as well as for extracting relevant characteristics from it (Granzotti, 2020). Another prominent factor of the model is its ability to reduce the dimensionality of the data nonlinearly.

AE is a neural network architecture that aims to approximately reconstruct input data from a latent representation, that is, a compressed version of the original data. The structure consists of two main parts: the Encoder, responsible for mapping the input data to the latent representation, and the Decoder, which reconstructs the data from this representation. During training, the model seeks to minimize the difference between the input data and the reconstructed data by adjusting the weights and hyperparameters of the network.

In this study, in addition to the use of the neural network structure autoencoder, the clustering technique was also used. This machine learning technique aims to analyze precipitation data in low dimensionality and group them taking into account their characteristics and patterns identified. This approach allows the regionalization of rainfall stations based on the similarity of their precipitation characteristics.

The precipitation data used in this study were obtained from the website of the National Water and Sanitation Agency (ANA), available in https://www.gov.br/ana/pt-br. These data represent the values of the daily precipitation cycle of 268 rainfall stations located in the Legal Amazon region, covering the period from 1986 to 2015.

The results of this study are highly promising, showing the effectiveness of the approach adopted. Data analysis using DCAE revealed a low dimensional structure that captured the essential characteristics of the original precipitation data. The approximate reconstruction of the input data presented considerable RMSE and NRMSE values, indicating the ability of the model to efficiently reproduce the important information contained in the data.

The reduction of the dimensionality of the data with the application of the DCAE allowed the accomplishment of a cluster analysis of the studied precipitation data. The results revealed the identification of homogeneous regions in the Legal Amazon in relation to this variable. This means that the proposed methodology was able to group rainfall stations based on similar characteristics of their rainfall patterns, providing valuable information on the spatial variability of precipitation in the region.

These findings are of great value, as they contribute to the understanding of the distribution and behavior of precipitation in the Legal Amazon, assisting in the management of water resources, in the planning of agricultural activities and in coping with extreme events, such as droughts and floods. The ability to identify homogeneous regions based on the precipitation variable opens the way for further studies on the climatic patterns of the region, strategies for adaptation and mitigation of impacts related to rainfall variability.

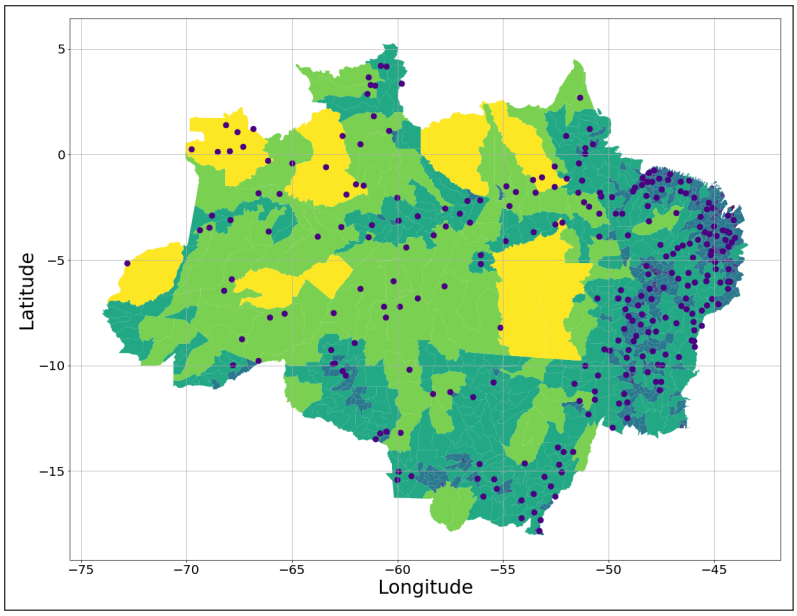
2 METHODOLOGY

In this section, we present in detail the methods that were used to carry out this research. Topics related to data collection, data preprocessing, development environment, configurations and architecture of the proposed model, dynamics of process flows and what else was needed to better understand the methodology addressed.

2.1 Data Collect

The data used in this work are values corresponding to the precipitation volume of 268 rainfall stations located in the Legal Amazon observed in a period of 30 years (1986 - 2015). According to Dourado et al. (2013), it is a standardization of the World Meteorological Organization to use in studies and projects a sample universe equal to 30 years of observations aimed at representing the climate of a given region. Data were collected in electronic spreadsheets that can be found on the website (https://www.gov.br/ana/pt-br) of the National Water and Sanitation Agency (ANA). Figure 1 shows the study area of this research.

Figure 1 – Legal Amazon: study area



Source: the authors (2023)

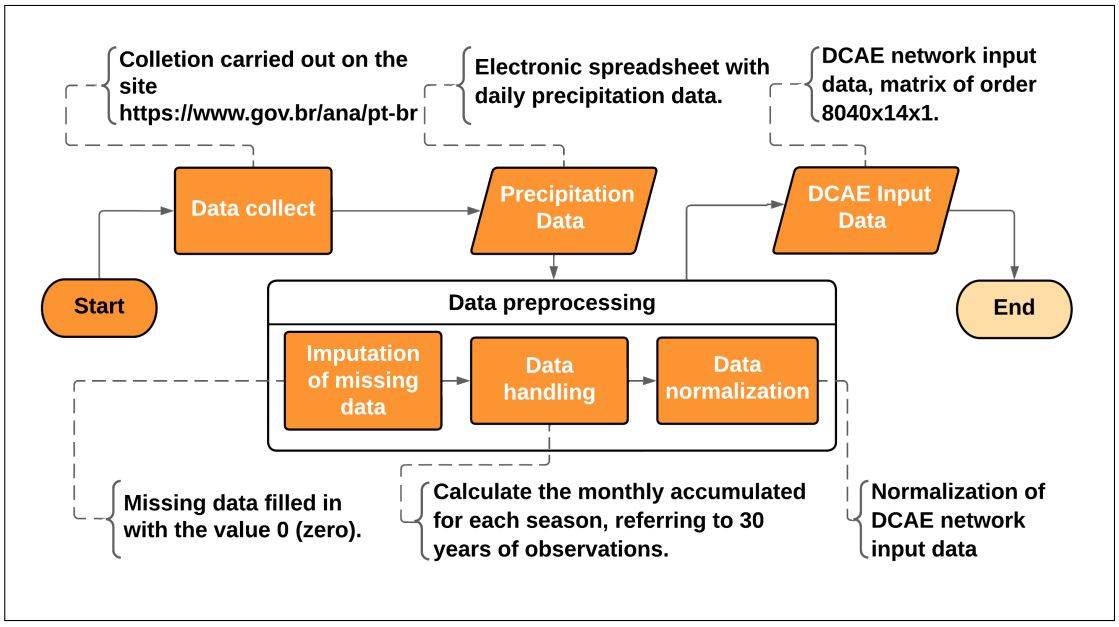
The Legal Amazon currently has in its geographical limits the quantity of 2,029 rainfall stations. Most stations present, within the study period, data with compromised integrity, due to unreliability and inconsistency of information over the available time series, this commitment derives from the excess of missing data and erroneous measurements Lira et al. (2019). After analyzing and performing the treatment of the missing data, we reached the quantitative of 268 rainfall stations installed in the Legal Amazon region, with time series of daily, continuous and complete data, within the study period, between January 1986 and December 2015, totaling 30 years of observations.

2.2 Data Preprocessing

According to Guarienti et al. (2015), rainfall observations using common instruments such as pluviometer or pluviograph are more susceptible to failures, which can occur for several reasons, either by equipment default or by negligence in observation, another factor highlighted is the local representation of the data. In the initial stage of data processing, due to the failures presented in the time series, the missing fields in the daily precipitation table are filled with the value 0 (zero), in the next process, the data is pre-processed for the DCAE network.

In figure 2, follows an illustration of the flow of this process of processing the raw data for transformation into input data of the neural networks.

Figure 2 – Processing flow for handling input data



Source: the authors (2023)

In the data processing, to obtain the input information for the training of the DCAE network, the values of the monthly accumulated for each observed year for each analyzed station and its geographical coordinates (longitude and latitude) are calculated Therefore, each rainfall season will be represented by 30 observations with 14 attributes, one sample for each year analyzed. After this stage is obtained a matrix of 8040 x 14, 8040 observations (30 years for each station) with 14 attributes (accumulated monthly and geographic location coordinates).

2.3 Model Training and Architecture

The deep convolutional autoencoder method is an ML technique that uses artificial neural networks of type AE with convolutional layers to learn and extract a representation of characteristics of low dimensionality of the input data, and later reproduced them roughly at the network exit. This technique is widely used in areas such as computer vision, image processing, speech recognition, etc.

The DCAE architecture consists of two main parts: the encoder and the decoder. The encoder is responsible for mapping the input data into a low-dimensional latent representation, while the decoder is responsible for reconstructing the input data from the latent representation.

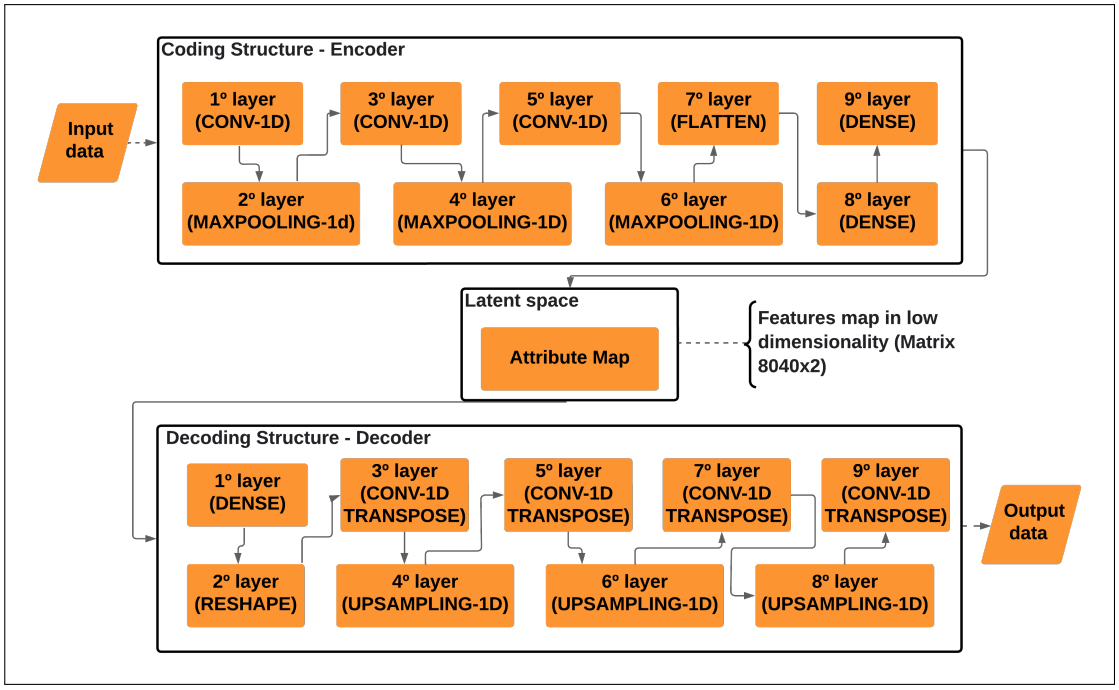
In the coding step, several convolutional and pooling layers are used to extract relevant characteristics from the input data and reduce their dimensionality. These convolutional layers are responsible for detecting patterns in the data, while pooling layers are responsible for compressing the data, preserving the most important characteristics.

In the decoding step, convolutional and pooling layers are used in the opposite way to reconstruct the input data from the latent representation. Pooling layers are replaced by up-sampling layers, which decompress the data, while convolutional layers are responsible for reconstructing the details of the original data.

According to Baia & Castro (2018), in the coding process this step can contain several layers of convolution and pooling, these layers can also be part of the decoding process but performing an opposite operation, aiming at the reconstruction of the input data. Using multiple convolutional and pooling layers in the encoding and decoding process allows DCAE to learn a more robust and compact representation of the input data, making it more efficient and accurate in reconstructing the original data.

For the structure of the DCAE model under study, an input layer, three convolution layers, three pooling layers, one Flatten layer and two Dense layers are implemented in the coding process. Likewise, in the process of decoding the data, a Dense layer, a reshape layer, four transposed convolution layers and three pooling layers are implemented in the model architecture. This structure is represented in figure 3.

Figure 3 – Proposed deep convolutional autoencoder processing flow



Source: the authors (2023)

The entire development and training process of the model was carried out in a Python 3.8 environment (PYTHON, 2021), using the API Keras 2 (Keras, 2021), with the TensorFlow 2.6.0 library as a backend (TENSORFLOW, 2021) and the aid of a Geforce 8400 GS GPU supporting CUDA technology (NVIDEA, 2021). Other features of the development and training platform are i5-4590 3.30 GHz processor and 16GB of RAM.

2.3.1 Configuration of model hyperparameters

The API Keras 2 (KERAS, 2021) used in this model has some layer modules that are necessary for the configuration of the network structure, as well as its full functioning. These modules adopt the use of hyperparameters that each iteration seeks to adjust the model so that the loss is minimized.

With regard to obtaining ideal values of hyperparameters, we sought empirically a configuration that presented reasonable results to the model evaluation metric. The process was started with low values, in this case the value two, and equal in all filters and kernels of the transposed convolutional and convolutional layers.

After the first execution, only the values of the filters were changed, increasing them gradually with multiple values of the number of months under study. After assigning the value 36, it was observed that the reduction of the error of the evaluation

metric of the model stabilizes, and compared to the computational cost generated with the increase of the values of this parameter, this value was adopted as a reference.

After finding a reference value for the filters, we tried to adjust the kernel values of the convolutional layers. The same methodology adopted to determine the values of the filters was used to obtain the values of the kernels. Thus, the reference value found was 18, where similarly, it was noted a stability in reducing the error of evaluation of the model. With the reference values for the filters and kernels defined, new experiments were carried out with the values of the filters and kernels varying between the layers of the model, and the values found as the maximum values. Thus, we found the best configuration for the proposed model, according to the evaluation metric.

Architectures that use convolutional layers receive in their input layer data in three dimensions, therefore, the number of channels was added to the data matrix under study, establishing the order matrix 8040x14x1.

For model training, a process with 2000 iterations with lots of size equal to 30 was used, causing all the observations of each station to be applied in a single lot. The NONEx14x1 matrix was presented to the DCAE input layer, where ”NONE” represents the size of the data lot, so the model input data are matrices of order 30x14x1.

Convolutional layers in architectures for time series analysis perform one-dimensional convolutions. In this operation, a filter is slid along the input vector to extract relevant characteristics at each position. The pooling operation can also be performed one-dimensional, reducing the size of the vector without compromising the relevant information. These adaptations in convolutional layers allow convolutional architectures to be used for time series analysis, with applications in areas such as time series forecasting, detection of anomalies in time series, signal processing, among others.

As an optimizer of the training system was applied the process that performs the implementation of the adaptive Moment estimation algorithm (ADAM), which is a stochastic gradient optimizer that combines the idea of moment update of the stochastic gradient Descent (SGD) with an adaptive estimate of second moment of gradient variations. It calculates an individual adaptive learning rate for each parameter of the model, which helps to deal with scale problems in different directions.

Table 1 – Values adopted for the hyperparameters of the model

Description	**Filters**	***Kernel Size***	***Activation***	***Strides***	***Padding***
Coding Process					
Input Layer	(8040, 14, 1)				
1ª Layer *Conv1D*	36	18	Selu	1	Causal
2ª Layer *Maxpooling*	6			1	
3ª Layer *Conv1D*	24	12	Selu	1	Causal
4ª Layer *Maxpooling*	5			1	
5ª Layer *Conv1D*	12	6	Selu	1	Causal
6ª Layer *Maxpooling*	4			1	
7ª Layer *Flatten*					
8ª Layer *Dense*	6		Selu		
9 ª Layer *Dense*	2				
Latente Space	(8040, 2)				
Decoding Process					
Input Layer (espaço latente)	(8040,2)				
1ª Layer *Dense*	168		Selu		
2ª Layer *Reshape*	(14,1)				
3ª Layer *Conv1DTranspose*	12	6	Selu	1	Same
4ª Layer *Upsampling*	1				
5ª Layer *Conv1DTranspose*	24	12	Selu	1	Same
6ª Layer *Upsampling*	1				
7ª Layer *Conv1DTranspose*	36	18	Selu	1	Same
8ª Layer *Upsampling*	1				
9ª Layer *Conv1DTranspose*	1	1	Selu		Same
Output Layer	(8040, 14, 1)				

Source: the authors (2023)

In general, the ADAM algorithm is a powerful method for optimizing deep neural networks in problems with large volumes of data and complex parameter configurations. However, it is important to carefully adjust the hyperparameters of the algorithm, such as the learning rate, momentum coefficient and regularization parameters, to obtain the best results. According to Kingma & Ba (2014), it is a method with low computational cost, since it requires little memory consumption, ideal for solving problems with large volumes of data and complex parameter configurations.

In table 1, shows the values adopted for the hyperparameters of the coding and decoding layers of the proposed model.

2.3.2 Model performance evaluation

The evaluation of the performance of a model is a critical step in any modeling process, as it allows to evaluate the ability of the model to perform specific tasks. The common approach for model evaluation in many areas is the calculation of normalized mean square error (MSNRE), which is a measure of the difference between the values predicted by the model and the actual values of the test data.

Before you evaluate the model with the NRMSE metric, you must configure the model to assess the loss in rebuilding the input data, which is commonly called Loss. This is a measure of the difference between the values predicted by the model and the actual values of the training data. The goal of the training process is to minimize this Loss, so that the model is able to generalize well to new data.

Once the model has been trained and Loss has been minimized, it is possible to evaluate its performance using the NRMSE metric. NRMSE is calculated by dividing the root Mean square error (RMSE) by the breadth of the test data. It is normalized to allow comparison of model performance across different datasets.

Thus, it is possible to evaluate the best configuration for the model proposed in this study. The Loss is calculated using the Mean square error (MSE), of possession of the MSE we calculate the RMSE that is given by equation 3.

$$e_i = \hat{x}_i - \bar{x}_i \tag{1}$$

$$\|e_i\| = \|\hat{x}_i - \bar{x}_i\|^2 \tag{2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|e_i\|^2} \tag{3}$$

Then the NRMSE is given by equation 4.

$$NRMSE = \frac{RMSE}{\sqrt{var(x)}} \tag{4}$$

The RMSE calculation is based on the mean of the differences between the reconstructed data and the original squared observations. In order to evaluate the error between the predicted values and the original inputs, having its result varying from zero. On the other hand, the NRMSE considers the scale of the observed values, facilitating the comparison between models that have different scales, in this way one can reach the information of the proximity between the result and the original data. The variation of its values is part of one.

3 RESULTS

In this research, we analyzed precipitation data for 30 years from 268 rainfall stations located in the Legal Amazon. Using machine learning techniques with unsupervised learning, the discovery of knowledge and identification of patterns of the data studied was performed. Deep neural networks techniques of the deep convolutional autoencoder type were adopted for extraction of knowledge and representation in low dimensionality of data, and clustering techniques, such as hierarchical agglomerative with Ward's binding method, for the task of grouping and better analysis of station data.

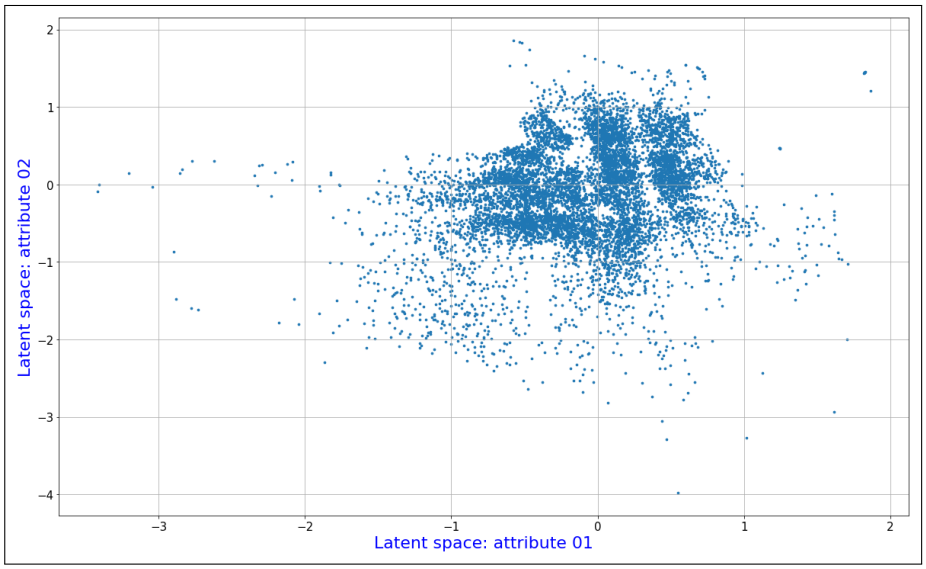
3.1 Results of data analysis using the DCAE model and application of the clustering technique

As a result of the coding step performed by the DCAE model, the precipitation data of the 268 selected stations in the Legal Amazon are represented in low dimensionality in the latent space, generating the data with two-dimensional structure. This representation allows the visualization of 8040 observations of rainfall stations in a dispersion plot, where each point represents the annual precipitation cycle of the 30 years studied for each selected season. Its position in the graph is determined by the values of the attributes of the two dimensions of the latent space. Figure 4, presents the scatter plot which is an important tool for exploratory data analysis and identification of patterns or groupings of rainfall stations with similar characteristics.

After the analysis of the attributes present in the latent space, it was noticed that most of the observations are concentrated, in relation to the x and y axes, between the values -1.0 and 1.0. This result is evidenced by the dispersion graph produced, in which

it was possible to observe the considerable agglutination of the majority of the observations in low dimensionality. These observations represent the accumulated monthly precipitation of the selected rainfall stations within the period of 30 years under study.

Figure 4 – Scatter plot with latent space data



Source: the authors (2023)

Each point in the plane represents one year of observation of each station, and the farthest points can be considered representation of years with distinct observations of the patterns. These observations may have occurred due to failures in obtaining the data or by atypical characteristics of the natural phenomenon studied. Therefore, it is important to analyze these outliers carefully and understand their causes.

In addition to data visualization, the latent space generated by the model, the MSE metric was also used as a network analysis parameter to assess the loss of information in the reconstruction of the original data. This was a way of evaluating the behavior of the model, taking into account that for approximate reconstruction of the original data the decoding process requires an input, which is the data of the latent space. This entry shall have a good representation of the characteristics learned from the original data. In this sense, it could be considered that the MSE metric, used in the approximate reconstruction of the original data, would also be an option to evaluate the performance of the proposed model.

In order to improve the error metric adopted in the model, the MSE was taken as the root of the mean square error, RMSE, and then as the root of the mean square error normalized, NRMSE. Table 2 shows the RMSE and NRMSE values generated by the proposed DCAE.

Table 2 – RMSE and NRMSE generated by the model

RMSE	NRMSE
0.06610	0.3355

Source: the authors (2023)

After obtaining the representation in low dimensionality of the time series of precipitation of the clustering techniques were used with the data generated for analysis and identification of patterns and mapping of homogeneous regions in the Legal Amazon in relation to monthly precipitation values.

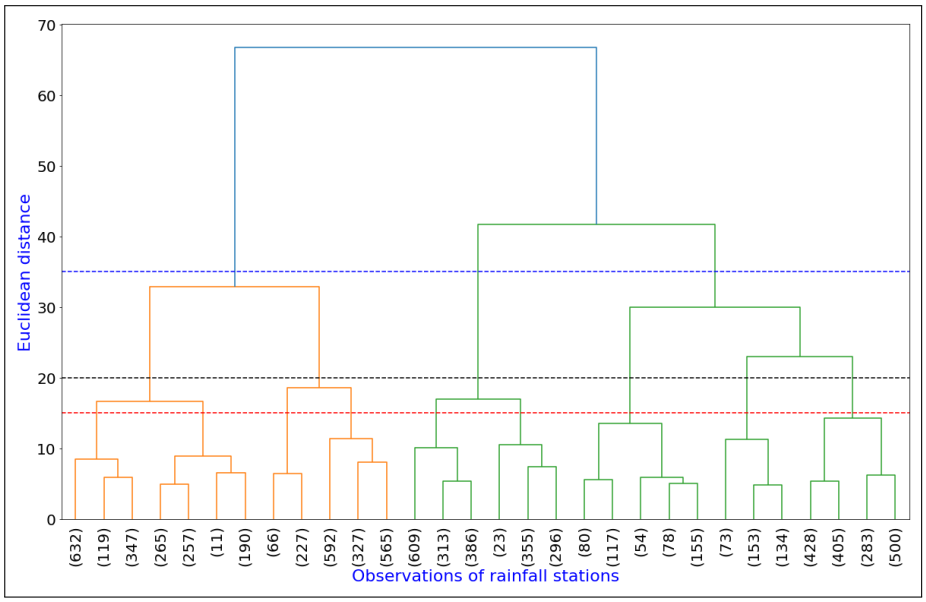
According to Neves et al. (2017), Menezes et al. (2015) and Amanajás & Braga (2012), in the context of climate data analysis, the clustering technique allows grouping rainfall stations with similar precipitation patterns, identifying geographic or climatic regions with consistent rainfall characteristics. By grouping rainfall stations based on similar characteristics, one can identify regional weather patterns and better understand spatial and temporal variations of precipitation. This information is essential to know the rainfall regimes, providing valuable information for decision making related to agriculture, water resource management and other areas, as well as for the elaboration of strategies for adaptation to local climatic conditions.

Clustering is an unsupervised learning technique that aims to group objects with similar attributes into groups. In this study, the hierarchical agglomerative technique was used with the Ward binding method to analyze and group data from the latent space. It is important to emphasize that this methodology requires prior knowledge of the number of clusters that must be formed. In this way, to obtain the information of the quantitative clusters necessary for analysis of the attributes of the latent space and realization of unsupervised learning, the technique of dendrogram was applied, which analyzes the Based on the Euclidean distance metric, separate the data into clusters.

The application of dendrogram is a technique commonly used in studies involving data analysis and ML. This technique allows to identify the appropriate

number of clusters for the analysis of the attributes of the latent space and the realization of unsupervised learning. In addition, the Euclidean distance metric is a popular measure to calculate the similarity between data, and thus group them into clusters. Figure 5 shows the dendrogram formed.

Figure 5 – Dendrogram for cluster formation using euclidean distance as a metric.



Source: the authors (2023)

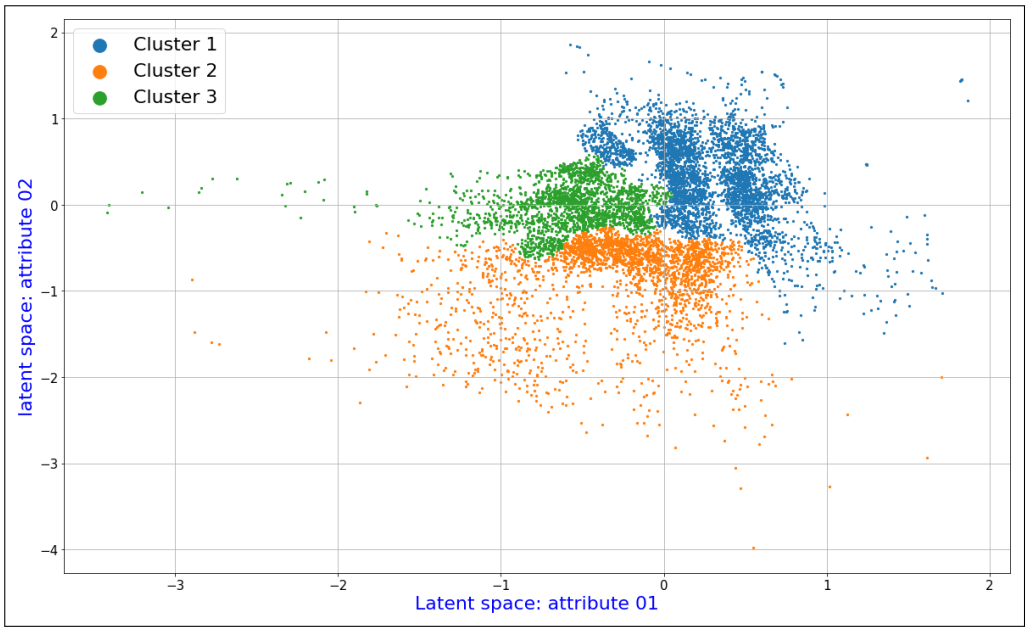
Analyzing the generated dendrogram, it was observed among the possible formations, a formation with three clusters on the precipitation data presented the proposed technique to obtain the number of clusters. According to Santos et al. (2015), three homogeneous regions are sufficient to represent the different precipitation patterns in the Amazon. Because the regions are consistent with the performance of the main atmospheric systems responsible for the formation of rain in the Amazon. Regarding the climatology of the state of Pará, which belongs to the Legal Amazon, there are three climatic subtypes: ”Af”, ”Am”, ”Aw”, these subtypes were determined with the climatic classification of Köppen and are related to the tropical rainy climate Crispim et al. (2020).

The dotted line in blue corresponds to the Euclidean distance adopted for separation into three groups of observations. For better analysis, understanding and comparison of the data, other formations with different numbers of clusters were performed. Thus, the formations adopted for the study were those that presented clusters with 3, 6 and 9 distinct groups.

3.1.1 Result of applying the clustering technique for formation with 3 groups

After defining the number of clusters that were used in the study, we applied the clustering techniques chosen in the study for analysis and comparison of the results. Therefore, the clustering process was initiated with the hierarchical agglomerative method with Ward binding. In figure 6, the latent space is represented in a scatter plot with the definition of clusters, in this case, with the formation of 3 groups.

Figure 6 – Scatter plot with latent space data: Formation with 3 clusters



Source: the authors (2023)

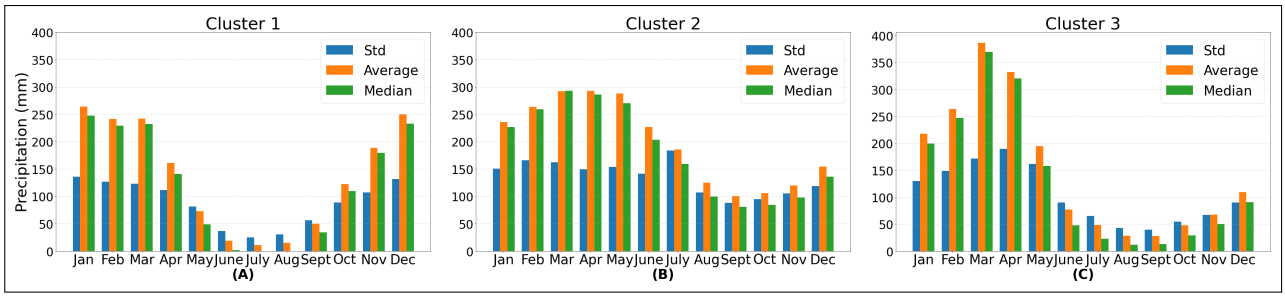
Based on this analysis, it was possible to identify the formation of three distinct clusters, with cluster-1 being the most numerous, with 3598 observations, followed by cluster-2 with 2460 observations, and finally, cluster-3 with 1982 observations. In total, these clusters represent a set of 8040 observations, allowing a more accurate analysis of precipitation data and its variations over the years. This information is valuable for understanding natural phenomena and contributes to the development of models for forecasting and analysis of climate trends.

The cluster-1, represented by the blue dots, concentrated most of its observations between the values 0 and 1 on the X-axis of the graph, and presented similar values for the Y-axis. Already the cluster-2, represented by the orange dots, presented most of his observations concentrated between -1 and 1 on the X axis, and between -2 and 0 on the Y axis.

The cluster-3 is represented by the green dots scattered in the plane, and presented a higher level of agglutination than the other clusters. The observations of this cluster are concentrated between -1 and 0 on the X axis and between -0.5 and 0.5 on the Y axis. From this information, it was possible to identify clear differences between the three clusters and better understand the characteristics and patterns present in the data.

After the formation of clusters and assignment of each observation to their respective cluster, it was necessary to perform statistical analysis of the data to better understand the patterns that determined the formation of each group. In this sense, it was essential to calculate the standard deviation measures, mean and median of the observations belonging to each cluster. Thus, it was possible to obtain an overview of the distribution of data in each group and make comparisons between the clusters represented in the graphs of figures 7, 8, 11, 12, 16 and 17. These statistical analyzes allow to identify distinct characteristics in each cluster and can be useful for decision making in several areas, such as urban planning, agribusiness and water resources management, among others.

Figure 7 – Bar graphs with statistics (standard deviation, average and median) of the clusters: Formation with 3 clusters



Source: the authors (2023)

This information is important for understanding the behavior of clusters and helping to identify seasonal weather patterns. In addition, the analysis of the standard deviation, mean and median of each cluster in graphs 7.A, 7.B and 7.C, also provided information on the variations of the data within each cluster. These statistical indicators were useful to understand the consistency of observations within the same cluster and their relationship with the other clusters. The comparison between the different clusters helped to identify differences and similarities in precipitation patterns, allowing more accurate and robust analysis.

The graph 7.A, which represents the cluster-1, showed that rainfall volume is more significant in the months of December to March, while in the months of May to September there is a decrease in rainfall volume. This pattern is consistent with the climatic characteristics of the studied region, where the rainy season begins in November, peaks in January and begins to decrease in April. From May, the driest period begins.

In the graph 7.B representing the cluster-2, it was observed that the months of March, April and May have the highest rainfall volumes, while August, September, October and November have the lowest volumes. It was noted that the month of December can be considered as a transitional period, with the onset of the most intense rains that extend until July, when the transition to the dry period begins. This analysis is important for understanding cluster-2 weather patterns.

The study carried out by Lira et al. (2020b), analyzed the historical series of the city of Belém and found that the distribution of rainfall throughout the year shows significant variations, forming a very rainy period that comprises the months of December to May and a less rainy period that comprises the months of August to November. This rainfall temporal variation is influenced by the main atmospheric systems that act in the region. These results corroborate the findings of the present study, which after analysis of the time series data by the DCAE system and application of the clustering method, identified that the pluviometric station representing the city of Belém was grouped within the cluster-2. This cluster presented the same characteristics of the distribution of rainfall both in the rainy and dry periods, as identified in Chart 7.B. Thus, the results of the two studies complement each other and reinforce the importance of considering the influence of atmospheric systems on the temporal variability of rainfall in the region.

The cluster-3, represented by the 7.C graph, presented distinct characteristics of the other clusters analyzed. In this cluster, it was possible to observe a longer period with the lowest rainfall volumes, which extends from June to December, especially the months of August and September, which have the lowest precipitation values. In contrast, the month of March stands out as the wettest. The statistical analysis of the historical series showed that the beginning of the increasing volume of rainfall in this cluster occurs in December, reaching the peak in March and beginning to decrease in

May. In June, the driest period begins.

The cluster analysis was fundamental to understand the temporal and spatial variability of rainfall in a given region. However, it is important to highlight that the same rainfall season could present observations in more than one cluster, being explained for several reasons, such as an atypical year or failures in the capture of the precipitation volume.

To solve this question, it was necessary to create a mechanism for evaluation and ranking of observations of each rainfall station, in order to define which cluster a given observation belonged to. Thus, an X station was included in the Y cluster, when most observations of this X station belonged to the Y cluster. This methodology allowed a more accurate analysis of the distribution of rainfall in a given region and the identification of temporal and spatial variations. After the ranking, the stations were distributed among the clusters according to table 3.

Table 3 – Number of rainfall stations per formed cluster

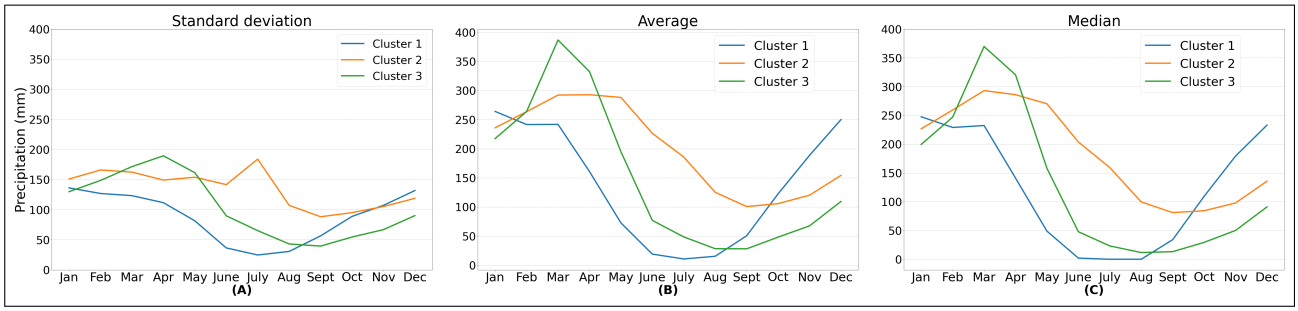
Cluster description	**Number of stations per cluster**
$Cluster - 1$	132
$Cluster - 2$	85
$Cluster - 3$	51

Source: the authors (2023)

With the definition of which cluster each of the rainfall stations belonged, a statistical analysis of the median, mean and standard deviation of the clusters formed was performed. Figure 8, presents comparative graphs between these clusters, the standard deviation measures are shown in graph 8.A , the mean in graph 8.B, and the median in graph 8.C.

The results presented in the graphs demonstrated the distinction of the standard values for each cluster, highlighting the characteristics of each group. The behavior of the initial months was similar between the clusters, while the months between April and August showed greater difference in precipitation values. It was possible to notice that the months that make up the dry season are those that presented greater dissimilarity between the clusters, indicating that the identification of the distinct characteristics between the groups is clearer in this period.

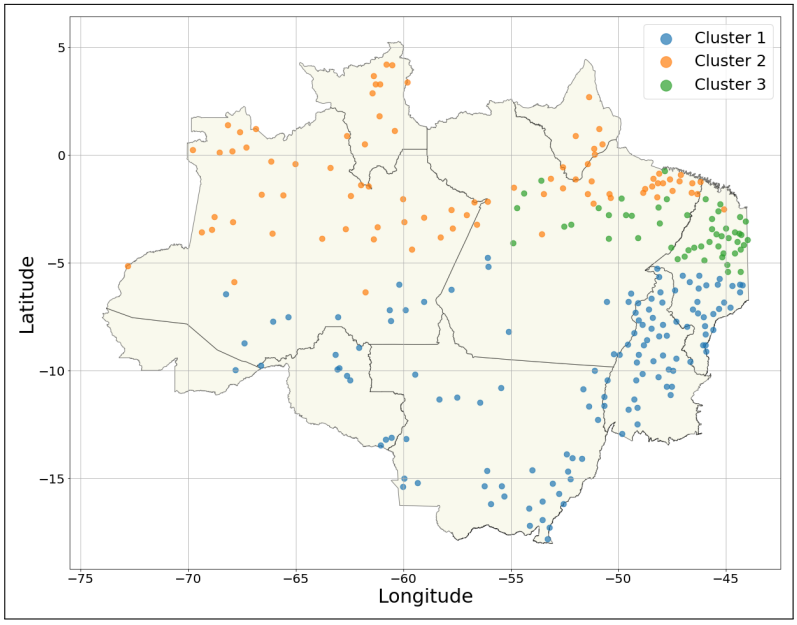
Figure 8 – Comparison chart of the standard statistical measures identified in each cluster: Formation with 3 clusters



Source: the authors (2023)

With the data in reduced dimensionality, clustering techniques were applied to evaluate them in the search for patterns and, consequently, separate the stations according to the patterns found. The use of clustering techniques allows the identification of homogeneous regions in relation to precipitation, which can be extremely useful for several areas, including agriculture, water resource management and climate forecasting.

Figure 9 – Map of the Legal Amazon with the arrangement of pluviometric stations formed by cluster: Formation with 3 clusters



Source: the authors (2023)

The use of dimensionality reduction models and clustering techniques is a common practice in the field of ML in general, as they allow the identification of patterns and characteristics that may be difficult to detect otherwise. In the analysis of

precipitation time series, this approach is especially important because it allows the identification of homogeneous regions and the understanding of precipitation trends and patterns over time. In this sense, figure 9 presented the geographic layout of the rainfall stations in the Legal Amazon, as well as the configuration of the clusters formed by the stations.

The clustering analysis, performed with data from the rainfall stations of the legal Amazon, allowed us to observe a north-south division with a smaller grouping in a band to the east between the two largest clusters. This information is valuable to understand the distribution of precipitation in the region and can be used for various practical purposes.

A previous study conducted by Lira et al. (2020a), highlighted the importance of two climatic stations with higher volume of rain in relation to total precipitation in the state of Pará. When the clustering analysis was performed with the formation in two clusters, it was possible to observe the rainfall separation between north and south in the state of Pará.

As a result of the analysis, three clusters representing different regions in the Legal Amazon were found: the cluster-1, represented by blue dots, occupied the southernmost area; the cluster-2, represented by orange dots, concentrated in the strip to the north; and the cluster-3, represented by green dots, was located in an easternmost range between the two other clusters. This spatial configuration of clusters suggests that the distribution of precipitation in the Legal Amazon region presents a clear north-south division, with a transition zone to the east.

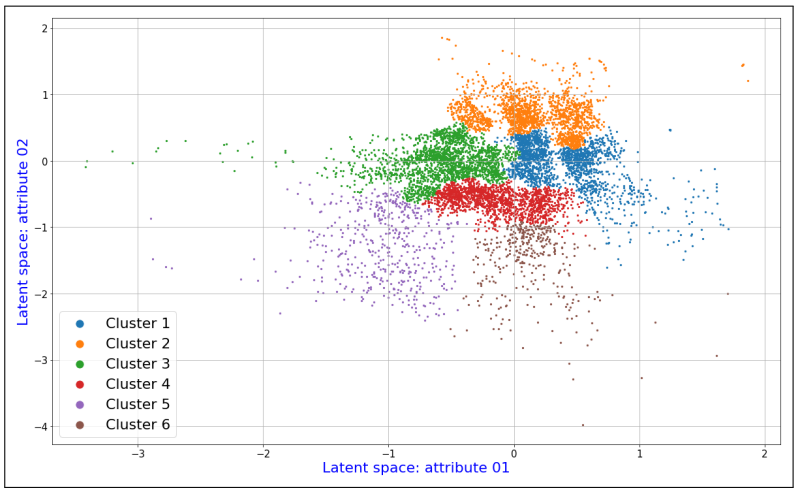
3.1.2 Result of applying the clustering technique for formation with 6 groups

When analyzing the graph provided in figure 10, it was observed that six clusters of data were identified, each represented by a different color. Through the observation of the distribution patterns of the points of each cluster, it was possible to infer relevant information about the structure of the data.

When comparing the methodologies with formations with 3 and 6 groups, it was identified that only clusters 1 and 2 of the first methodology underwent significant changes. Cluster 1 of the formation with 3 groups was divided into clusters 1 and 2 in the formation with 6 groups. Suggesting the existence of two distinct subpopulations

in this cluster. Cluster 2 of the formation with 3 groups was divided into clusters 4, 5 and 6 in the formation with 6 groups, and cluster 4 obtained the largest share of the division. This division demonstrated that the observations present in this cluster had different characteristics, and therefore could be subdivided into more homogeneous groups.

Figure 10 – Scatter plot with latent space data: Formation with 6 clusters



Source: the authors (2023)

After analyzing the observations and the clusters formed in the configuration with six groups, the ranking of the observations was performed to determine which group each rainfall station belonged to. Thus, the quantitative of the rainfall stations by clusters was defined as shown in table 4.

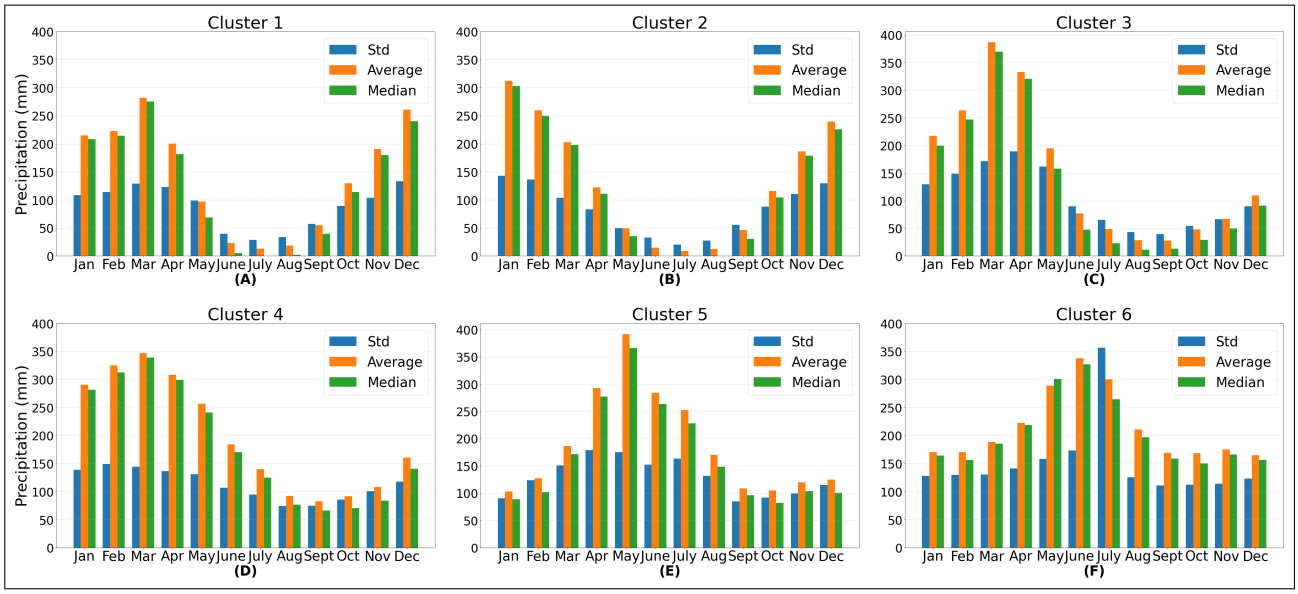
Table 4 – Number of rainfall stations per formed cluster

Cluster description	Number of stations per cluster
$Cluster - 1$	48
$Cluster - 2$	69
$Cluster - 3$	70
$Cluster - 4$	60
$Cluster - 5$	13
$Cluster - 6$	8

Source: the authors (2023)

The graphs presented in figure 11, show the statistical measures of median, standard deviation and mean of the clusters formed in the clustering analysis with six clusters. When analyzing these measures, valuable information was obtained about the characteristics of each group formed.

Figure 11 – Bar graphs with statistics (standard deviation, average and median) of the clusters: Formation with 6 clusters



Source: the authors (2023)

In the graph 11.A, which represents cluster 1, formed by the hierarchical clustering method with six clusters, it was noticed that the wettest month in this cluster corresponds to the month of March. In addition, it is interesting to note that this period of many rains began in November and the fall in the volume of rainfall began in May, extending until October, July being the month with the lowest volume of precipitation.

Analyzing the information presented in the text and the available graphics, it was possible to identify the existence of a seasonal pattern in the behavior of rainfall in the region represented by cluster 2, as indicated in Chart 11.B. Through the observation of the data, it was noticed that the months of May to September are characterized as the period of greatest drought, with July presenting the lowest volume of rain.

From the month of October, however, a process of gradual increase in the volume of rainfall in the region began, reaching its peak in January, with an average volume of more than 300mm. This phase characterizes a period of greater rainfall in the region, with a wetter climate and frequent rainfall. From April, in turn, the rains began to gradually decrease, and the region began a transition to a drier period. This phase characterizes a period of lower rainfall, with a drier climate and less frequent rainfall.

When analyzing Figure 11.C, which represents cluster 3, it was possible to establish that this is one of the clusters that presented the highest number of months in a dry period, in relation to the volume of precipitation. During the dry season, the average volume of rain in cluster 3 varied between 25mm and 100mm, which certainly indicated a challenge for activities that depended on water, such as agriculture.

The rainy season in cluster 3 began in January and ended in May, featuring a phase of greater rainfall in the region. The period with the lowest volume of rain lasted seven months, starting in June and running until December. Among the months analyzed, the month of March stood out as the one with the highest rainfall volume and August as the one with the lowest rainfall volume.

The Graph 11.D represents cluster 4, which has a very distinct precipitation pattern from the other clusters. In this cluster, the month of March stood out with the highest volume of precipitation, with an average of approximately 350mm, while the month of September was the lowest volume. The driest period covers the months of August to November. The month of December was marked by the beginning of the strongest rains in this cluster, characterized as a transition period for the period of more intense rains, which extended from January to June. In turn, the month of July was characterized by the beginning of the transition to the months of lower precipitation volume.

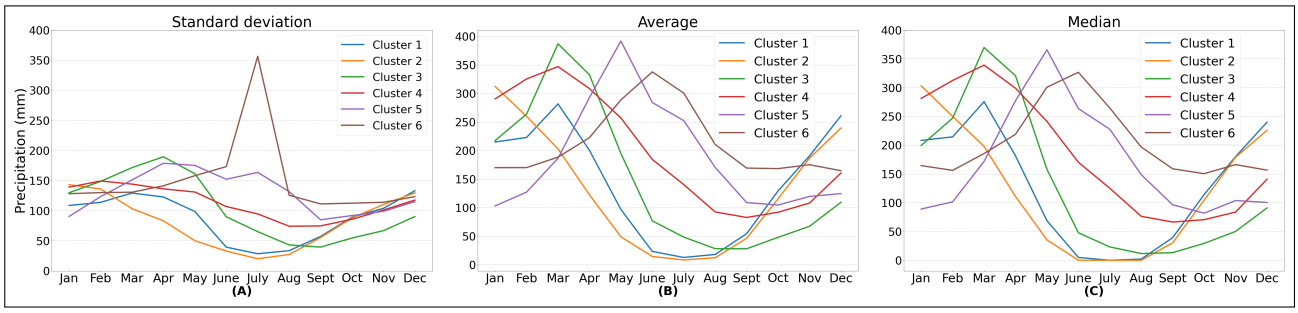
The cluster 5, shown in figure 11.E, represents a distinct climatic pattern, was characterized by a period with intense rainfall predominant from April to July, with the month of May as the and March as a month of transition to the beginning of the strongest rains. It is important to highlight that January was one of the months with the lowest average rainfall volume in this cluster. In addition, August could be considered as the beginning of the transition to the driest period, which extended from September to February.

The cluster 6, described in Figure 11.F, presents a very extensive dry season, which lasted from September to March. In this cluster, the month of April could be considered the transition to the period of greater precipitation volume. This cluster was also characterized by a stronger rainfall interval, which occurred between the months of May to July, with August as the month of transition to the onset of droughts.

Another important feature of cluster 6 was the average monthly precipitation volume in the driest period, which is higher than the other clusters, measuring between 150mm and 200mm. Suggesting that even during periods considered dry, it was still possible to record heavy rainfall in the coverage of cluster 6.

The same statistics were generated using line graph to obtain a different approach in the analysis of this information and a comparison between the clusters formed. Figure 12 represents this approach.

Figure 12 – Comparison chart of the standard statistical measures identified in each cluster: Formation with 6 clusters



Source: the authors (2023)

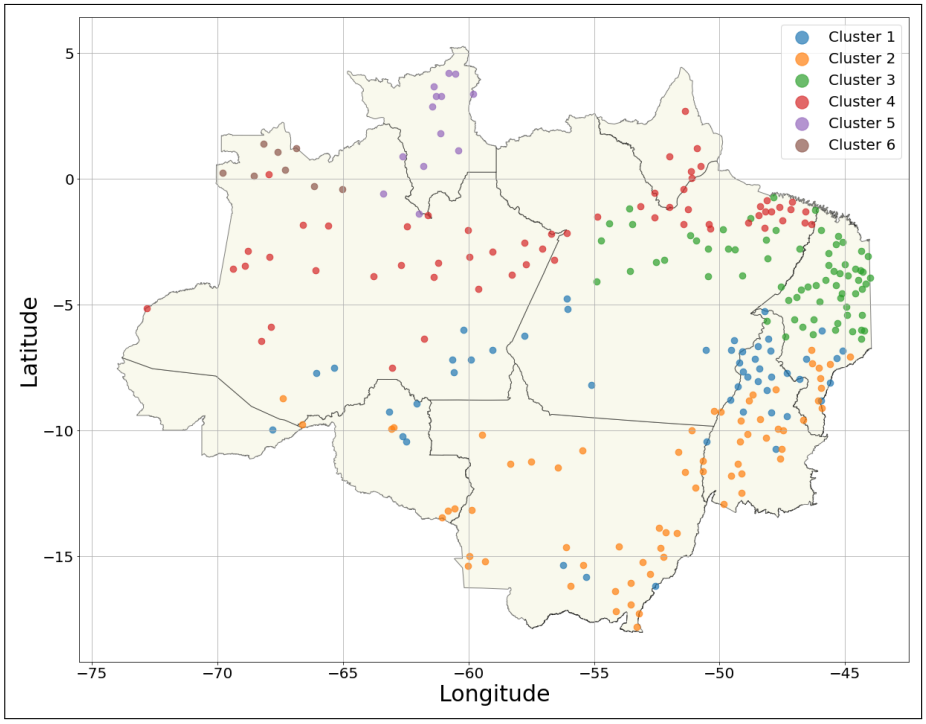
Similarly, as performed in the formation with three clusters, it was realized the impression of the geographic disposition of the rainfall stations in the Legal Amazon, as well as the configuration of the clusters formed by the stations, in the format with six clusters. Figure 13, shows the arrangement of the stations for the formation of six groups.

The cluster 1 was represented by the blue dots that are located in a central strip of the Legal Amazon, which extends from southern Maranhão to Acre. This strip passes through the north of Tocantins and south of Amazonas, in addition to traveling the south and southeast of Pará and the north of Rondônia. It is important to highlight that, even being a central strip, there were some rainfall stations that were far from this region, being south of Mato Grosso. This cluster was considered one of the most important, since it presented a wide variety of stations and covered an extensive geographical area of the Legal Amazon.

When observing the distribution of rainfall stations in the Legal Amazon, it was observed that cluster 2, represented by the orange dots, presented some uniformity in its geographical arrangement. Most of the stations of this cluster were located in the

state of Mato Grosso, with some stations in the center-south of Tocantins and south of Maranhão.

Figure 13 – Map of the Legal Amazon with the arrangement of pluviometric stations formed by cluster: Formation with 6 clusters



Source: the authors (2023)

It stands out in this formation with 6 clusters cluster 3, which practically maintained the same arrangement of the stations of cluster 3 of the formation with 3 clusters. Its stations remained in a region east of the Legal Amazon, with most stations located in Maranhão, 43 stations in total. The other stations were located as follows: 24 stations north of Pará; 2 stations in the northern region of Tocantins; and 1 in the southern region of Amazonas, which is the farthest from the group.

Another important information is the increase in the number of stations belonging to this cluster in relation to the previous formation. In the first formation the cluster was composed of 51 rainfall stations, already in this new scenario the cluster has 70 stations. Fifteen of these new stations came from cluster 1, and four of these new stations are from cluster 2, based on the formation with 3 groups.

It could be noted that in the formation with 6 clusters, cluster 2 of the previous formation was divided into three distinct clusters, 4, 5 and 6. Cluster 4, represented by the red dots, stood out for having rainfall stations distributed in very different regions,

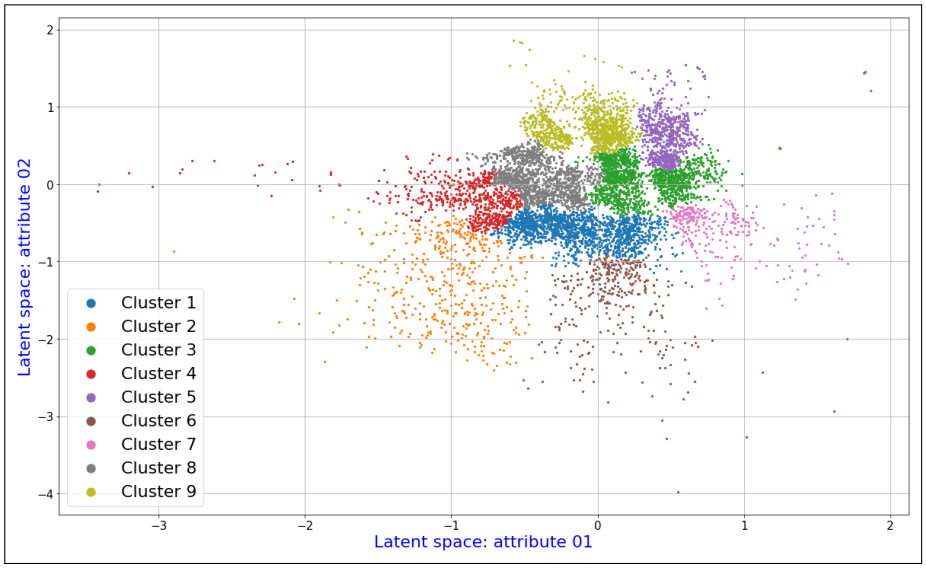
north, northeast and west of Pará, in addition to all stations in the state of Amapá and most stations in the Amazon.

The cluster 5 covered all ten rainfall stations belonging to the state of Roraima, as well as three other stations located north of Amazonas, very close to the border with Roraima. The eight rainfall stations that make up the cluster 6, were positioned north of the state of Amazonas, near the borders with Colombia and Venezuela.

3.1.3 Result of applying the clustering technique for formation with 9 groups.

When analyzing the scatter plot printed in figure 14, it was possible to notice the division of the observations into nine distinct clusters. This division is important for understanding precipitation patterns in the Legal Amazon. It is noteworthy that in this graph were plotted the 8040 observations corresponding to each year, from the period of 30 years, of the 268 stations studied.

Figure 14 – Scatter plot with latent space data: Formation with 9 clusters



Source: the authors (2023)

Of the 8040 observations evaluated, cluster 1 had the largest number of samples, with 1616 observations. Cluster 3 then received 1484 observations, and cluster 8 received 1308 observations. On the other hand, the clusters with the least number of observations were cluster 6, with only 360 samples, and cluster 7, with 293 samples. Cluster 2 received 484 observations, cluster 4 received 674 observations, cluster 5 received 723 observations, and cluster 9 received 1098 observations.

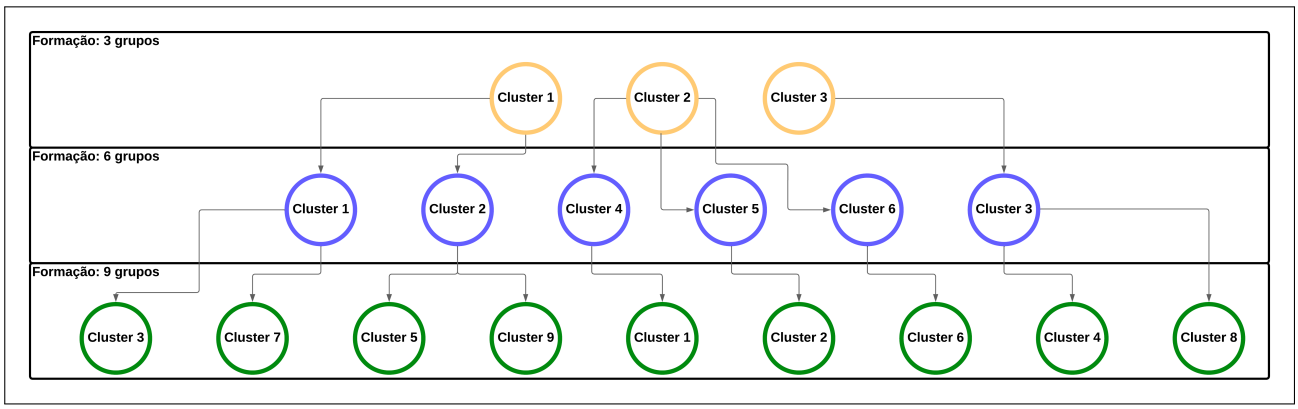
When comparing these different configurations, it was possible to notice that some groups of formation with nine clusters can be considered subgroups of clusters different from the current one, present in other formations. This observation suggests that the use of nine clusters generated an excessive segmentation of the data, which may not be ideal in certain analysis contexts.

Analyzing the formations of the clusters obtained through the clustering technique in this study, it was possible to identify some interesting relationships between the different configurations used. Comparing the formation of clusters of this topic with the other formations that used three and six clusters, it was noticed that cluster 1 of the formation with three groups was divided into two clusters in the formation with six groups, clusters 1 and 2.

In the formation with nine clusters, these clusters 1 and 2 were divided into four clusters, with cluster 3 formed from cluster 1 of formation with six groups and cluster 9 formed from cluster 2 of formation with six groups. In addition, cluster 5 of the formation with nine groups was formed from cluster 2 of the formation with six groups, while cluster 7 was formed from cluster 1 of the formation with six groups.

These relationships between the different grouping formations provided valuable insights for understanding the precipitation data of the Legal Amazon, helping to identify regions with similar rainfall characteristics. In figure 15, this relationship between the clusters was demonstrated.

Figure 15 – Comparison between the formation clusters with 3, 6 and 9 groups



Source: the authors (2023)

According to the comparative table in figure 15, it was possible to perceive the division of clusters and their origins related to the other formation. These subgroups

formed in the nine-cluster split configuration can be embedded in the formations with 3 and 6 groups, which can be seen as an advantage of this approach.

By applying the methodology of ranking the observations in the formation with nine distinct groupings, it was possible to establish a division of the rainfall stations in the Legal Amazon into groups with similar characteristics. With the formation of clusters, it was possible to understand the relationship between the rainfall stations and the meteorological variable, as well as their geographic distributions. The ranking of the observations also allowed to identify the rainfall stations with the most similar characteristics, facilitating the comparative analysis and the identification of climatic patterns in the region.

The number of rainfall stations present in the clusters is described in table 5.

Table 5 – Number of rainfall stations per formed cluster

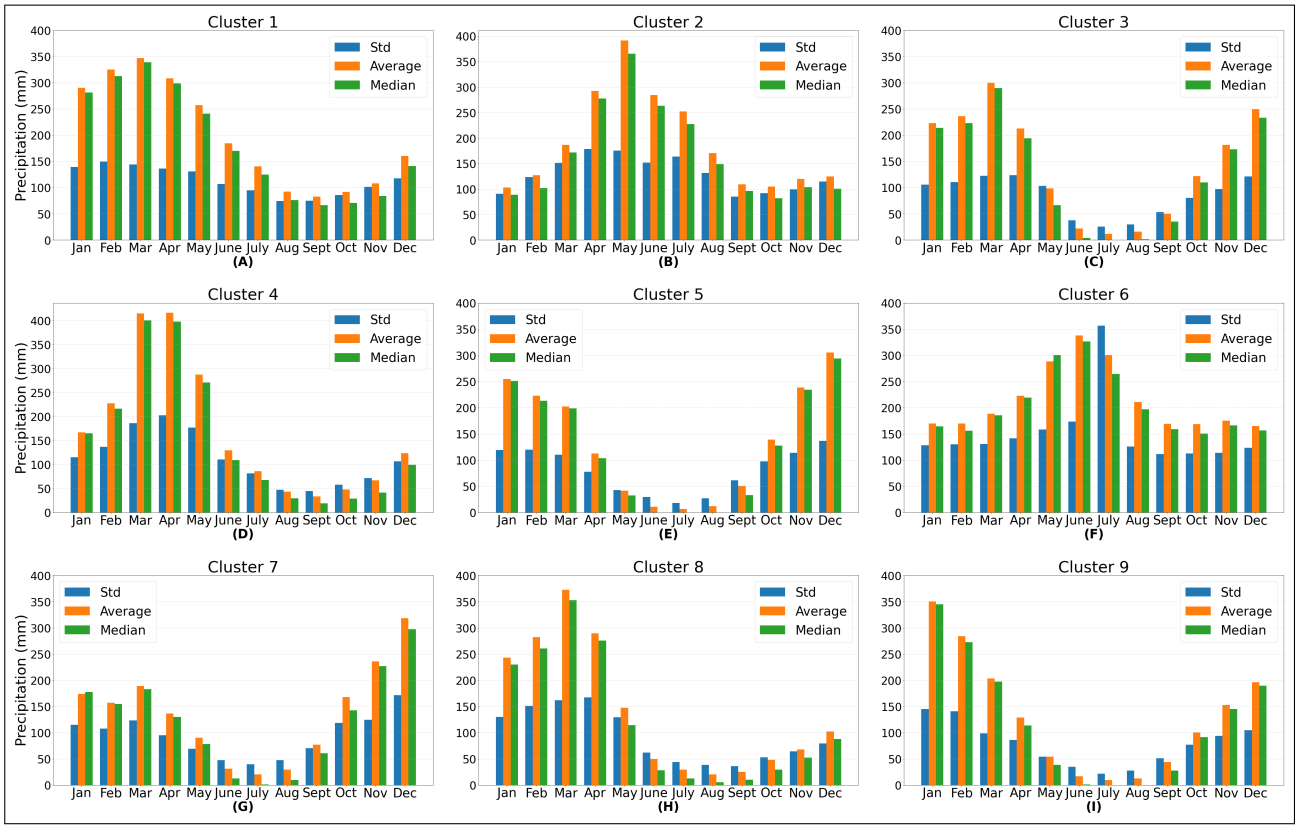
Cluster description	Number of stations per cluster
$Cluster-1$	71
$Cluster-2$	13
$Cluster-3$	75
$Cluster-4$	10
$Cluster-5$	11
$Cluster-6$	8
$Cluster-8$	49
$Cluster-9$	31

Source: the authors (2023)

By presenting statistical measures for each cluster, it was possible to identify rainfall patterns that could be common to a given group of stations. This is important for understanding the climate of the region under study and for planning activities that depend on rainfall conditions. In figure 16, we presented, through bar graphs, the statistics of each cluster formed in the configuration with nine groups.

In particular, cluster 1 of the formation with nine groups showed similarities with clusters 4 and 2 of the formations with six and three clusters, respectively. These clusters exhibited a pattern of more intense rainfall in the first six months of the year, followed by a dry period in the following six months. It is worth mentioning that the month of March stood out as the period of greatest average volume of precipitation, while September was the month of lowest average volume. This discovery indicated the presence of a consistent trend in these regions, regardless of the number of clusters considered.

Figure 16 – Bar graphs with statistics (standard deviation, average and median) of the clusters: Formation with 9 clusters



Source: the authors (2023)

When analyzing the formation under study, it was observed that cluster 2 had a greater statistical similarity with cluster 5 of formation with six clusters. When comparing the statistics presented in the bar graphs of figures 7, 11 and 16, it was not possible to identify, in the formation with three clusters, a cluster that presented patterns similar to cluster 2 under analysis.

However, it is important to note that cluster 5 of formation with six groups and cluster 2 of formation with nine groups originate from the division of cluster 2 of formation with three clusters. Therefore, it could be considered that the pattern found in clusters 2 and 5, which presented similar characteristics, are subgroups that were present in cluster 2 of the formation with three groups.

It was observed that cluster 3, represented by graph 16.C, presented statistical patterns approximated to those demonstrated in graphs 7.A and 11.A, which represented clusters of formation with three and six clusters respectively. These clusters shared similar characteristics in relation to precipitation patterns. In the

regions represented by these clusters, a period of less precipitation was evident, characterized by a considerable drought in the months of June, July and August. On the other hand, the highest rainfall volume averages are recorded in the period from December to March.

the cluster 4 showed similarity in its characteristics with clusters of other formations analyzed. Its pattern revealed that the months of March and April stood out with the highest average volume of rain, with the most intense rainy period between January and May. On the other hand, the dry season occurred from July to November, with the months of August and September registering the lowest rainfall volumes.

Analyzing the other formations, we identified clusters with characteristics close to cluster 4, represented in graphs 7.C and 11.C. These clusters also exhibited a similar pattern, with the months of March and April standing out for the higher average of precipitation and the period from January to May characterized by more intense rainfall. Similarly, the months of August and September are those with the lowest volume of rainfall in these regions.

The cluster 5, represented by graph 16.E, presented statistical characteristics similar to cluster 1 of formation with three groups and cluster 2 of formation with six clusters, represented by graphs 7.A and 11.B, respectively. A significant difference between these clusters is in the month characterized as the largest volume of rain. In the cluster of graph 16.E, the month of December recorded the highest average rainfall volume, with an average of 300 mm. On the other hand, cluster 2, represented by graph 11.B, also presented an average rainfall volume around 300 mm, but the month of occurrence is January. The cluster represented by the graph 7.A indicated that January is the wettest month, although with an average rainfall volume slightly higher than 250 mm.

The cluster 6, from the formation with nine clusters, represented by graph 16.F, presented interesting statistical characteristics. When analyzing the pattern of this cluster, it was observed that it had a prolonged dry season, which runs from August to April. During these months, the average volume of precipitation varies between 150 mm and 200 mm. This average rainfall is considerable, especially when compared to other clusters that presented similar averages in their wettest periods. In cluster 6, the

months of May, June and July stood out as those with the highest volume of precipitation. Specifically, the month of June is characterized as the most intense rainy season in this cluster.

As in the clusters 1, 2, 3, 4, 5 and 6, which showed similar characteristics to other clusters originating from the other formations, clusters 8 and 9 also showed some similarity to clusters of other formations. However, a case was identified in which there was no similarity, which occurred with cluster 7.

The cluster 7 was composed of sixteen observations analyzed by DCAE, corresponding to sixteen years of precipitation data of the studied stations. As mentioned earlier, the same station was able to present observations in different clusters. In the case of cluster 7, after the observations ranking step to determine which cluster each station would belong, it was found that this cluster did not receive any station, due to the low number of observations related to each station. Therefore, cluster 7 was removed from formation.

Nevertheless, the observations of this cluster shared characteristics with other clusters, especially with clusters 3 and 5, which received five stations each. These similarities indicated climatic patterns or specific geographic characteristics shared by these stations.

The cluster 8, represented by graph 16.H, exhibited as characteristic in its rain pattern a long dry phase. It was observed a period of low precipitation volume that extends for seven months, from July to December. In this period, the average volume of rain varied approximately between 10mm and 90mm, and August presented the lowest average volume. On the other hand, March stood out as the month with the highest average precipitation volume, with values close to 350mm. The rainy season covered the months from January to May.

These characteristics presented by cluster 8 in formation with nine clusters, resembled the corresponding characteristics of cluster 3 in formation with six groups, as illustrated in graph 11.C, and cluster 3 in formation with three groups, as shown in Chart 7.C. This similarity suggested the existence of similar weather patterns in these regions or the presence of common geographic factors that influenced rainfall patterns.

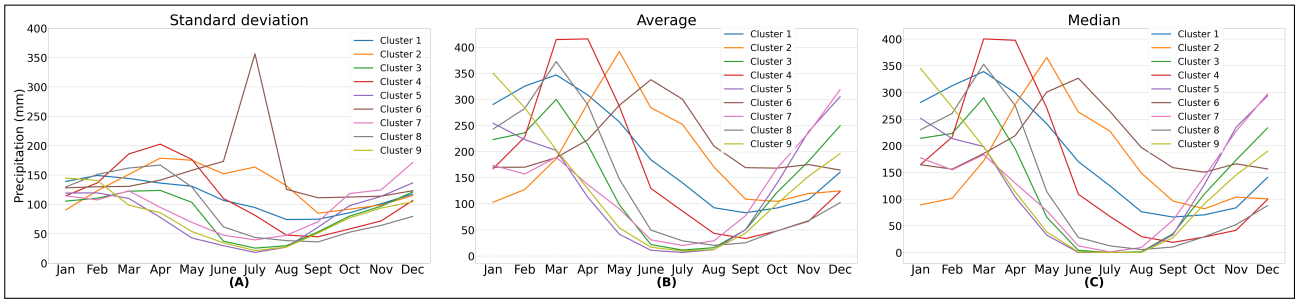
The graph 16.I, in figure 16, represented the pattern identified for cluster 9. This pattern revealed a more intense rainfall volume in the first months of the year, especially in January and February. On the other hand, the period from May to September was characterized by a lower average precipitation volume, with a maximum average of 50mm. It was possible to consider that the rainy season covered from November to March, with January being the month with the highest volume of rain. The drought, in turn, occurred from April to October. The characteristics presented by this cluster were similar to the characteristics of clusters 2 and 1 in the formations with six and three groups, respectively, as demonstrated in graphs 11.B and 7.A. This similarity suggested a relationship between the rain patterns of these clusters in different cluster configurations.

In the analysis of the formation with nine groups in question, the methodology previously established was followed, which included the application of statistical comparisons between the groups formed. Similarly, in this stage, a differentiated approach was adopted using line graphs for visualization, interpretation of data and comparison that represented the annual precipitation cycle of the stations that composed each cluster. Figure 17 represents this approach and provides valuable insights into patterns and relationships between groups.

The figure 18, presented the geographic layout of the 268 rainfall stations divided into groups that obtained similar patterns in relation to the precipitation characteristics. This spatial representation of the stations allowed to identify the homogeneous regions in terms of rainfall patterns, which provided valuable information on the spatial distribution of rainfall and the understanding of the factors that influenced rainfall variability.

When comparing the disposition of rainfall stations in relation to the strategies of configuration of the number of clusters used in the analyses, it was observed that the approach with nine clusters resulted in more heterogeneous regions in terms of the precipitation characteristics extracted by the DCAE model and analyzed by the clustering technique. This means that this configuration allowed to identify more detailed differences in precipitation characteristics between regions.

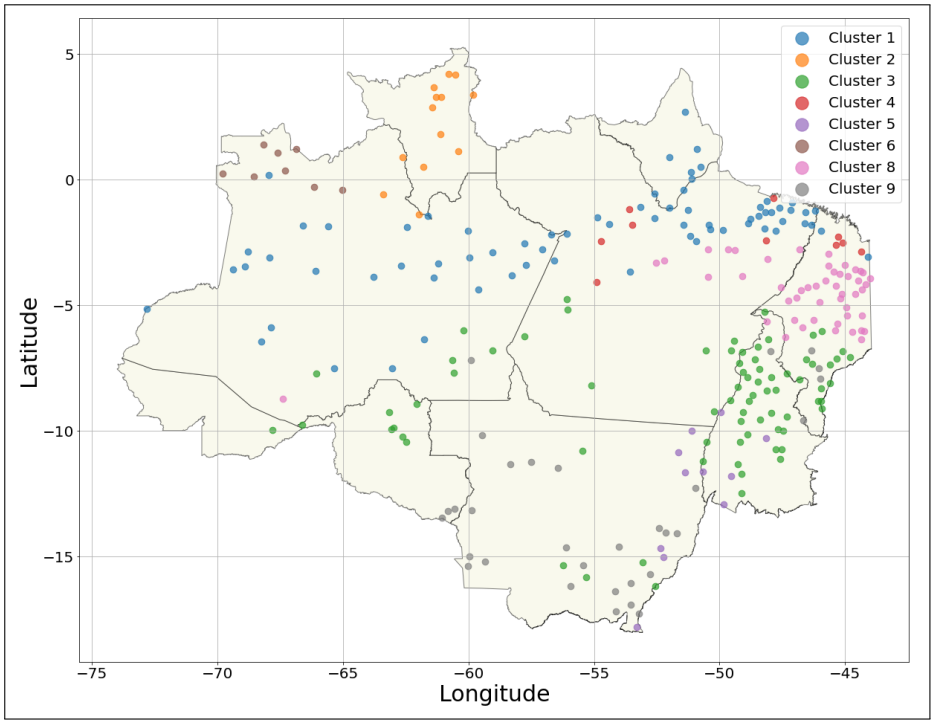
Figure 17 – Comparison chart of the standard statistical measures identified in each cluster: Formation with 9 clusters



Source: the authors (2023)

It is important to highlight that some regions were divided into subgroups, while others suffered few changes in their composition. This made it possible to evaluate the consistency of the characteristics extracted for these specific regions. Clusters 1, 2 and 8 were the least affected in the formation with nine clusters, compared to formations with three and six groups. This indicated that these clusters have distinct and stable characteristics along the different cluster configurations.

Figure 18 – Map of the Legal Amazon with the arrangement of pluviometric stations formed by cluster: Formation with 9 clusters



Source: the authors (2023)

4 CONCLUSION

The proposed model was trained with precipitation time series data collected in rainfall stations located in the Legal Amazon. This model can be approached by several studies that seek to use time series analysis of precipitation as a method to assist in rainfall prediction, classification of rainfall regions, among other results in the Legal Amazon region.

The methodological approach built in this study presents good results, since it was able to answer four questions that guided the realization of this research. The first question was: Is the proposed deep convolutional autoencoder model capable of generating a significant representation in the latent space that can make the discovery of new knowledge in precipitation time series? After treatment of the time series made by the model and application of the clustering technique, it was possible to answer in an affirmative representation of the data of the observations of the selected stations presented a good interpretation of the clusters to which they belonged.

The second and third questions were also answered by the model in a positive way, because empirically it was possible to apply the configuration of hyperparameters that best fit the model, and the level of reconstruction of the original data were also satisfactory as the errors in reconstruction were low, with RMSE and NRMSE values equal to 0.06610 and 0.3355, respectively.

The fourth question was also answered positively, since when applying the clustering technique in the data of the latent space, it was possible to establish homogeneous regions within the Legal Amazon in relation to the variable of precipitation.

Even with promising results the structure of the DCAE presented is open to improvements. Changes in the structure and configurations of the model, as well as the imputation of missing values in the time series using other methods, can generate improvement in performance both in the generation of patterns in low dimensionality and in the reconstruction of the data.

Future work related to this study may test new structures in the DCAE, seek and evaluate a methodology present in the literature for automatic configuration of the as well as implement a second DCAE structure equivalent to the first and apply as input data the output of the first DCAE network of the model.

REFERENCES

Amanajás, J. C. & Braga, C. C. (2012). Padrões espaço-temporal pluviométricos na amazônia oriental utilizando análise multivariada. *Revista Brasileira de Meteorologia*, 27:423–434.

Baia, A. F. & Castro, A. R. G. (2018). A Competitive Structure of Convolutional Autoencoder Networks for Electrocardiogram Signals Classification. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. (pp. 538-549). Porto Alegre: SBC.

Bailão, A. S. d. O. et al. (2020). *Reconhecimento de padrões por processos adaptativos de compressão* (Dissertação de Metsrado). Universidade Federal de Goiás, Goiânia, GO, Brasil.

Bhavsar, P., Safro, I., Bouaynaya, N., Polikar, R., & Dera, D. (2017). Machine learning in transportation data analytics. In *Data analytics for intelligent transportation systems*. (pp. 283-307). Elsevier.

Crispim, D. L., Fernandes, L. L., Ferreira Filho, D. F., & Lira, B. R. P. (2020). Comparação de métodos de agrupamentos hierárquicos aglomerativos em indicadores de sustentabilidade em municípios do estado do pará. *Research, Society and Development*, 9(2):e60922067–e60922067.

Cruz, E. B. et al. (2016). *Representação de séries temporais usando descritores de forma aplicados a recurrence plots* (Dissertação de Mestrado). Universidade Estadual de Campinas, Campinas, SP, Brasil.

Dourado, C. d. S., Oliveira, S. R. d. M., & Avila, A. M. H. d. (2013). Análise de zonas homogêneas em séries temporais de precipitação no estado da bahia. *Bragantia*, 72(2):192–198.

Esling, P. & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):1–34.

Essien, A. & Giannetti, C. (2020). A deep learning model for smart manufacturing using convolutional lstm neural network autoencoders. *IEEE Transactions on Industrial Informatics*, 16(9):6069–6078.

Gonçalves, E. D., Pessoa, F. C. L., Neves, R. R., Rodrigues, R. S. S., & de Sousa, A. C. S. R. (2017). Identificação de regiões homogêneas e análise de regressão múltipla para regionalização de vazão na bacia hidrográfica do rio tapajós. *Revista Brasileira de Cartografia*, 69(9).

Granzotti, R. A. (2020). *Extração de características via autoencoders aplicada a interfaces cérebro-computador baseadas em potenciais evocados visualmente em regime estacionário*. (Tese de Doutorado). Faculdade de Engenharia Elétroca e de Computação, Universidade Estadual de Campinas, Campinas, SP, Brasil.

Guarienti, G. S. S. et al. (2015). *Desenvolvimento de uma técnica computacional de processamento espaço-temporal aplicada em séries de precipitaç (Dissertação de Metsrado). Universidade Federal de Mato Grosso, Cuiabá, MS, Brasilão*.

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., Guo, L., & Liu, T. (2017). Modeling task fmri data via deep convolutional autoencoder. *IEEE transactions on medical imaging*, 37(7):1551–1561.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lira, B. R. P., Crispim, D. L., Ferreira Filho, D. F., Fernandes, L. L., & Pessoa, F. C. L. (2020a). Agrupamento de Precipitação no estado do (Pará), Brasil. *Revista de Gestão de Águas da América Latina, 17*(19).

Lira, B. R. P. et al. (2019). *Avaliação do comportamento e da tendência pluviométrica na Amazônia Legal no Período de 1986 a 2015* (Dissertação de Mestrado). Universidade Federal do Pará, Belém, PA, Brasil.

Lira, B. R. P., Lopes, L. d. N. A., das Chaves, J. R., Santana, L. R., & Fernandes, L. L. (2020b). Identificação de Homogeneidade, Tendência e Magnitude da Precipitação em Belém (Pará) entre 1968 e 2018. *Anuário do Instituto de Geociências, 43*(4), 426–439.

Maggioni & Silva, A. (2016). *Classificação de séries temporais baseada em análise de recorrência e extração de características* (Dissertação de Mestrado). Universidade Federal de Mato Grosso do Sul, Campo Grande, MS, Brasil.

Masci, J., Meier, U., Cireşan, D., & Schmidhuber, J. (2018). Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks*. (pp. 52-59). Springer.

Menezes, F. P., Fernandes, L. L., & da Rocha, E. J. P. (2015). O uso da estatística para regionalização da precipitação no Estado do Pará, Brasil. *Revista Brasileira de Climatologia, 16*.

Neves, R. R., Gonçalves, E. D., Pessoa, F. C. L., Fernandes, L. L., Gómez, Y. D., & dos Santos, J. I. N. (2017). Identificação de regiões pluviometricamente homogêneas na sub bacia trombetas. *Revista AIDIS de Ingeniería y Ciencias Ambientales. Investigación, desarrollo y práctica*, 10(2):125–135.

Sá, J. E. F. C. S. d. (2023). *Aplicação de Técnicas de Ciência de Dados na Previsão de Consumos Energéticos* (Mestrado em Ciência de Dados). Instituto Politécnico de Leiria, Leiria, Portugal.

Santos, E. B., Lucio, P. S., & Silva, C. M. S. e. (2015). Precipitation regionalization of the brazilian amazon. *Atmospheric Science Letters*, 16(3):185–192.

Severo, D. L., dos Santos Silva, H., & Tachini, M. (2019). Flutuações climáticas da precipitação no vale do itajaí (sc). *Revista de Estudos Ambientais*, 20(2):37–48.

Yin, C., Zhang, S., Wang, J., & Xiong, N. N. (2020). Anomaly detection based on convolutional recurrent autoencoder for iot time series. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *52*(1), 112-122.

Author contributions

1 – Vander Augusto Oliveira da Silva
Master in Applied Computing
https://orcid.org/0009-0004-3374-3748 • vander.silva@ifpa.edu.br
Contribution: Conceptualization; Data curation; Formal Analysis; Investigation; Methodology; Visualization; Writing - original draft

2 – Raphael Barros Texeira
Doctorin Electrical Engineering
https://orcid.org/0000-0003-2993-802X • raphaelbt@ufpa.br
Contribution: Conceptualization; Project Administration; Supervision; Validation; Writing - review & Editing

How to cite this article

Silva, V. A. O., & Texeira, R. B. (2025). Clustering of spatio-temporal precipitation patterns in the Legal Amazon using deep convolutional autoencoder. *Ciência e Natura*, Santa Maria, v. 47, e85042. DOI: https://doi.org/10.5902/2179460X85042.