UNIVERSIDADE FEDERAL DO PARÁ
NÚCLEO DE DESENVOLVIMENTO AMAZÔNICO EM ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

VANDER AUGUSTO OLIVEIRA DA SILVA

CLUSTERIZAÇÃO DE PADRÕES ESPAÇO-TEMPORAIS DE PRECIPITAÇÃO NA AMAZÔNIA VIA DEEP CONVOLUTIONAL AUTOENCODER

Tucuruí
2023

UNIVERSIDADE FEDERAL DO PARÁ
NÚCLEO DE DESENVOLVIMENTO AMAZÔNICO EM ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

VANDER AUGUSTO OLIVEIRA DA SILVA

CLUSTERIZAÇÃO DE PADRÕES ESPAÇO-TEMPORAIS DE PRECIPITAÇÃO NA AMAZÔNIA VIA DEEP CONVOLUTIONAL AUTOENCODER

Dissertação de Mestrado apresentado ao Programa de Pós-Graduação em Computação Aplicada. Núcleo de Desenvolvimento Amazônico em Engenharia - Universidade Federal do Pará, como requisito à obtenção do título de Mestre em Computação Aplicada.
Área de Concentração: Inteligência Computacional.
Orientador: Prof. Dr. Raphael Barros Teixeira

Tucuruí
2023

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

S586c Silva, Vander Augusto Oliveira da.
CLUSTERIZAÇÃO DE PADRÕES ESPAÇO-TEMPORAIS DE PRECIPITAÇÃO NA AMAZÔNIA VIA DEEP CONVOLUTIONAL AUTOENCODER / Vander Augusto Oliveira da Silva. — 2023.
82 f. : il. color.

Orientador(a): Prof. Dr. Raphael Barros Teixeira
Dissertação (Mestrado) - Universidade Federal do Pará, Núcleo de Desenvolvimento Amazônico em Engenharia, Mestrado Profissional em Computação Aplicada, Tucuruí, 2023.

1. Aprendizado de Máquina. 2. Autoencoder Convolucional Profundo. 3. Agrupamentos. 4. Reconhecimento de Padrões. 5. Séries Temporais de Precipitação. I. Título.

CDD 006.31

VANDER AUGUSTO OLIVEIRA DA SILVA

CLUSTERIZAÇÃO DE PADRÕES ESPAÇO-TEMPORAIS DE PRECIPITAÇÃO NA AMAZÔNIA VIA DEEP CONVOLUTIONAL AUTOENCODER

Dissertação de Mestrado apresentado ao Programa de Pós-Graduação em Computação Aplicada. Núcleo de Desenvolvimento Amazônico em Engenharia - Universidade Federal do Pará, como requisito à obtenção do título de Mestre em Computação Aplicada.
Área de Concentração: Inteligência Computacional.
Orientador: Prof. Dr. Raphael Barros Teixeira

Aprovada em 07 de julho de 2023.

BANCA EXAMINADORA:

Documento assinado digitalmente
gov.br RAPHAEL BARROS TEIXEIRA
Data: 25/08/2023 16:49:14-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Raphael Barros Teixeira - Orientador.

Documento assinado digitalmente
gov.br CLEISON DANIEL SILVA
Data: 25/08/2023 11:54:30-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Cleison Daniel Silva.

Documento assinado digitalmente
gov.br ALAN MARCEL FERNANDES DE SOUZA
Data: 16/08/2023 17:34:21-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Alan Marcel Fernandes de Souza.

Documento assinado digitalmente
gov.br THABATTA MOREIRA ALVES DE ARAUJO
Data: 18/08/2023 13:34:00-0300
Verifique em https://validar.iti.gov.br

Profª. Me. Thabatta Moreira Alves de Araújo.

AGRADECIMENTOS

Primeiramente, agradeço ao meu orientador, Prof. Dr Raphael Barros Teixeira, pela orientação, apoio e valiosas informações ao longo de todo o processo. Suas orientações críticas e conhecimento especializado foram fundamentais para o desenvolvimento deste trabalho.

Expresso minha gratidão aos membros da banca avaliadora, Prof. Dr. Cleison Silva, Prof. Dr. Alan Souza, e Prof. Me. Thabatta Araújo, por dedicarem seu tempo e expertise para avaliar este trabalho e fornecer sugestões construtivas que enriqueceram a qualidade final desta dissertação.

Agradeço também ao Núcleo de Desenvolvimento Amazônico em Engenharia da Universidade Federal do Pará e ao PPCA pela oportunidade de estudar e pesquisar nesta instituição renomada. A infraestrutura e o ambiente acadêmico proporcionaram as condições ideais para a realização deste trabalho.

Minha gratidão se estende aos colegas de curso e amigos que compartilharam suas ideias, experiências e apoio ao longo desta jornada. A troca de conhecimentos e as discussões enriquecedoras contribuíram significativamente para o desenvolvimento das ideias apresentadas neste trabalho.

Por fim, dedico um agradecimento especial à minha esposa Juliana Valandro e aos meus filhos João Davi e Miguel, cujo amor, paciência e encorajamento foram essenciais durante todo o percurso. Suas palavras de incentivo e apoio emocional foram a força motriz por trás da minha dedicação e perseverança na conclusão desta dissertação.

Cada contribuição, grande ou pequena, desempenhou um papel crucial na realização deste trabalho. Agradeço a todos por fazerem parte desta jornada acadêmica e por tornarem esta conquista possível.

Resumo

Estudos utilizando diferentes métodos de aprendizado de máquina para descoberta de conhecimento e reconhecimento de padrões em séries temporais de precipitação são cada vez mais frequentes na literatura. Identificar e analisar padrões em séries temporais de precipitação em uma determinada região é fundamental para seu desenvolvimento socioeconômico. Logo, pode-se afirmar que o conhecimento e compreensão das características pluviométricas das regiões são importantes para viabilizar o planejamento do uso, manejo e conservação dos recursos hídricos. O fenômeno natural da precipitação é um processo fundamental de impacto direto nas bacias hidrográficas e no desenvolvimento humano e ambiental. A variabilidade desse fenômeno produz implicações importantes na navegabilidade dos rios, sobre a abundância do indivíduo e a riqueza das espécies. Nos últimos anos muitos estudos com essa abordagem foram realizados no Brasil, principalmente na região amazônica. Esta pesquisa teve como objetivo **desenvolvimento de um método computacional** para análise de séries temporais de precipitação utilizando técnicas de *machine learning* com aprendizado não supervisionado, afim de **propor um método capaz de realizar a extração de características complexas dos dados, obtendo um mapa de atributos em baixa dimensionalidade para reconhecimento de padrões, descoberta de regiões homogêneas com relação à precipitação e reconstrução aproximada de séries temporais de precipitação da Amazônia Legal.** O modelo de rede neural de aprendizado profundo 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 os resultados se mostram promissores, as observações dos dados reconstruídos apresentaram um bom desempenho conforme avaliação da métrica de **RMSE** e **NRMSE** com valores resultantes iguais a **0.06610** e **0.3355** respectivamente. A análise da representação dos dados em baixa dimensão foi aplicada e analisada por uma estrutura de *clustering* usando aglomerativo hierárquico com método de Ward. Essa metodologia também apresentou bons resultados, pois realizou agrupamentos consistentes caracterizando regiões homogêneas com relação aos dados de precipitação. Desta forma, demonstrando que a representação em baixa dimensionalidade carregava as características principais das séries temporais dos dados analisados. Destaca-se que o método desenvolvido nesse estudo pode ser aplicado não apenas na região amazônica, mas também em outras áreas com desafios semelhantes relacionados à análise de séries temporais.

Palavras-Chave: 1. Aprendizado de Máquina. 2. Autoencoder Convolucional Profundo. 3. Agrupamentos. 4. Reconhecimento de Padrões. 5. Séries Temporais de Precipitação.

Abstract

Studies using different machine learning methods for knowledge discovery and pattern recognition in precipitation time series are increasingly frequent in the literature. Identify and analyze patterns in precipitation time series in a particular region is fundamental for its socioeconomic development. Therefore, it can be stated that 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 natural phenomenon of precipitation is a fundamental process with a direct impact on watersheds and on human and environmental development. The variability of this phenomenon has important implications for the navigability of rivers, individual abundance and species richness. In recent years, many studies with this approach have been carried out in Brazil, mainly in the Amazon region. This research aimed to **develop a computational method for analyzing time series of precipitation using machine learning techniques with unsupervised learning, in order to propose an method capable of extracting complex features from the data, obtaining a map of attributes at low dimensionality for pattern recognition, discovery of homogeneous regions with respect to precipitation and approximate reconstruction of precipitation time series in the Legal Amazon.** The proposed deep learning neural network model is trained to learn the main and most complex features of the original data and present them in low dimensionality in latent space. After the training, the results are promising, the observations of the reconstructed data showed a good performance as evaluated by the **RMSE** and **NRMSE** metric with resulting values equal to **0.06610** and **0.3355** respectively. The analysis of the representation of the data in low dimension was applied and analyzed by a clustering structure using hierarchical agglomerative with Ward's method. This methodology also showed good results, as it carried out consistent groupings characterizing homogeneous regions in relation to precipitation data. Thus, demonstrating that the representation in low dimensionality carried the main characteristics of the time series of the analyzed data. It is noteworthy that the method developed in this study can be applied not only in the Amazon region, but also in other areas with similar challenges related to time series analysis.

Keywords: 1. Machine Learning. 2. Deep Convolutional Autoencoder. 3. Clustering. 4. Pattern Recognition. 5. Precipitation Time Series.

LISTA DE ILUSTRAÇÕES

Figura 1 – Representação da estrutura básica da arquitetura de *deep learning*. 26
Figura 2 – Representação da arquitetura básica de um *autoencoder*. 28
Figura 3 – Representação da estrutura de AE completo. 29
Figura 4 – Representação da estrutura de AE subcompleto. 29
Figura 5 – Representação da arquitetura de uma *convolutional neural network*. 31
Figura 6 – Representação de uma estrutura de *deep convolutional autoencoder* padrão. 32
Figura 7 – Amazônia Legal: área de estudo. 34
Figura 8 – Fluxo de processamento para tratamento dos dados de entrada. 35
Figura 9 – Fluxo de processamento do *deep convolutional autoencoder* proposto. . . . 37
Figura 10 – Representação do funcionamento do *kernel size*. 40
Figura 11 – Representação do funcionamento dos *strides*. 40
Figura 12 – Gráfico de dispersão com os dados do espaço latente. 46
Figura 13 – Dendrograma para formação de *clusters* utilizando a distância euclidiana como medida. 48
Figura 14 – Gráfico de dispersão com os dados do espaço latente: Formação com 3 *clusters*. 49
Figura 15 – Gráficos de barras com estatísticas (desvio padrão, média e mediana) dos *clusters*: Formação com 3 *clusters*. 50
Figura 16 – Gráficos com estatísticas (desvio padrão, média e mediana) em coordenadas polares: Formação com 3 clusters. 52
Figura 17 – Gráfico comparativo das medidas estatísticas padrões identificadas em cada *cluster* - Coordenadas polares: Formação com 3 *clusters*. 53
Figura 18 – Gráfico comparativo das medidas estatísticas padrões identificadas em cada *cluster*: : Formação com 3 *clusters*. 53
Figura 19 – Mapa da Amazônia Legal com a disposição das estações pluviométricas formada por *cluster*: Formação com 3 *clusters*. 55
Figura 20 – Gráfico de dispersão com dados do espaço latente: Formação com 6 clusters. 56
Figura 21 – Gráfico de barras com estatísticas (desvio padrão, média e mediana) dos *clusters*: Formação com 6 *clusters*. 58
Figura 22 – Gráficos com estatísticas (desvio padrão, média e mediana) dos *clusters* formados em coordenadas polares: Formação com 6 *clusters*. 60
Figura 23 – Gráficos com coordenadas polares das estatísticas das observações dos *clusters* comparado entre *clusters* formados: Formação com 6 *clusters*. 61
Figura 24 – Gráficos das estatísticas das observações dos *clusters* comparado entre *clusters* formados: Formação com 6 *clusters*. 61
Figura 25 – Mapa da Amazônia Legal com a disposição das estações pluviométricas: Formação com 6 *clusters*. 62
Figura 26 – Gráfico de dispersão com dados do espaço latente: Formação com 9 *clusters*. 65

Figura 27 – Comparativo dos *clusters* das formações com 3, 6 e 9 grupos. 66
Figura 28 – Gráfico de barras com estatísticas (desvio padrão, média e mediana) dos *clusters* formados: Formação com 9 *clusters*. 68
Figura 29 – Gráficos em coordenadas polares com estatísticas (desvio padrão, média e mediana) dos *clusters*: Formação com 9 *clusters*. 72
Figura 30 – Gráficos em coordenadas polares comparando as estatísticas entre *clusters*: Formação com 9 *clusters*. 73
Figura 31 – Gráficos comparando as estatísticas entre *clusters*: Formação com 9 *clusters*. 73
Figura 32 – Mapa da Amazônia Legal com a disposição das estações pluviométricas: Formação com 9 *clusters*. 74

LISTA DE TABELAS

Tabela 1 – Valores adotados para os hiperparâmetros do modelo 43
Tabela 2 – RMSE e NRMSE gerados pelo modelo 47
Tabela 3 – Número de estações por *cluster* formado. 52
Tabela 4 – Número de estações por *cluster* – Formação com 6 clusters 57
Tabela 5 – Número de estações por *cluster* – Formação com 9 *clusters* 67

LISTA DE ABREVIATURAS E SIGLAS

ADAM – Adaptive Moment Estimation

AE – Autoencoder

CAE – Autoencoder Convolucional

CNN – Rede Neural Convolucional

DAE – Autoencoder Profundo

DCAE – Autoencoder Convolucional Profundo

DL – Deep Learning

LSTM – Long Short Term

ML – Machine Learning

MLP – Multilayer Perceptron

MSE – Erro Médio Quadrático

NRMSE – Raiz do erro Médio Quadrático Normalizado

RMSE – Raiz do erro Médio Quadrático

tfMRI – Task-based functional magnetic resonance imaging

SUMÁRIO

1 INTRODUÇÃO 12
1.1 Justificativa de Pesquisa 17
1.2 Objetivo Geral 19
1.3 Objetivos Específicos 20
1.4 Estrutura do Trabalho 20
2 FUNDAMENTAÇÃO TEÓRICA 21
2.1 Séries Temporais de Precipitação 21
2.2 Aprendizado de Máquina 22
2.2.1 Aprendizado de Máquina Supervisionado 23
2.2.2 Aprendizado de Máquina Não Supervisionado 23
2.2.2.1 *Clustering* 24
2.2.3 Redes Neurais Artificiais 24
2.2.4 Redes Neurais Profundas 25
2.2.4.1 Redes Neurais de Codificação Automática - *Autoencoders* 27
2.2.4.2 *Autoencoder* Convolucional Profundo 30
3 METODOLOGIA DE PESQUISA 34
3.1 Base de Dados para Treinamento do Modelo 34
3.1.1 Fase de Coleta dos Dados 34
3.1.2 Fase de Pré-processamento dos Dados 35
3.2 Arquitetura e Treinamento do Modelo Proposto 36
3.2.1 Configurações dos hiperparâmetros do modelo 39
3.2.2 Avaliação de desempenho do modelo proposto. 43
4 RESULTADOS E DISCUSSÃO 45
4.1 Resultados para análise dos dados pelo modelo DCAE e aplicação da técnica de *clustering*. 45
4.1.1 Resultado da aplicação da técnica de *clustering* para formação com 3 grupos. 49
4.1.2 Resultado da aplicação da técnica de *clustering* para formação com 6 grupos. 55
4.1.3 Resultado da aplicação da técnica de *clustering* para formação com 9 grupos. 64
5 CONCLUSÃO 75
REFERÊNCIAS 76

1 INTRODUÇÃO

Machine learning (ML), ou aprendizado de máquina, é um campo da inteligência artificial (IA) que busca modelar dados através da aplicação de algoritmos, visando aprender e realizar otimizações com objetivo de alcançar resultados que satisfaça a resolução de problemas, seja na regressão ou na classificação de dados.

Essa tecnologia tem sido amplamente utilizada em diversas áreas do conhecimento. A razão para o aumento de sua popularidade está diretamente ligada ao seu desempenho excepcional em resolver problemas complexos. Por exemplo, as técnicas de ML podem ser usadas para identificar padrões em dados diversos extraindo características e conhecimento profundo dos quais podem ser utilizados em tarefas de regressão, classificação, e no desenvolvimento de sistemas inteligentes (HUANG et al., 2017; BAIA; CASTRO, 2018; ESSIEN; GIANNETTI, 2020; YIN et al., 2020).

Atualmente, o uso de dados e ferramentas computacionais de modelagem desempenha um papel fundamental na tomada de decisões. Com uma necessidade crescente de informações e uma imensa quantidade de dados gerados, essas técnicas para extrair características complexas e identificar padrões têm sido muito úteis para processar esse grande volume de dados.

Essa explosão de dados engloba uma ampla quatidade de áreas de estudo, abrangendo desde registros financeiros e médicos até dados climáticos e comportamentais, entre outros. Essa diversidade de fontes de dados fornece oportunidades únicas para a aplicação de técnicas de análise e modelagem computacional avançada, permitindo uma compreensão mais profunda e abrangente dos fenômenos em estudo.

A extração de características complexas visa identificar e representar aspectos inspirados dos dados que podem não ser prontamente observáveis. Essas características podem ser quantitativas, como medidas estatísticas, ou qualitativas, como padrões específicos de comportamento. O objetivo é capturar a essência dos dados e fornecer percepções valiosas para apoiar a tomada de decisões.

Nesse contexto, as séries temporais têm recebido atenção especial. Uma série temporal é uma sequência de observações coletadas em intervalos regulares ao longo do tempo. A análise de séries temporais visa extrair informações e padrões relevantes nos dados, a fim de compreender melhor o comportamento temporal dos fenômenos estudados. Isso envolve a identificação de tendências, sazonalidades, ciclos e padrões irregulares que possam estar presentes nos dados ao longo do tempo.

Para lidar com a natureza dinâmica e sequencial das séries temporais, técnicas computacionais avançadas são aplicadas. Algoritmos de ML, como modelos de regressão, redes neurais e processos estocásticos, são utilizados para modelar e facilitar a compreensão do comportamento desses dados em série.

A análise de séries temporais pode envolver a aplicação de métodos estatísticos para decomposição, filtragem e detecção de padrões. Também são utilizadas técnicas de visualização de dados, como gráficos de linhas e diagramas de dispersão, para facilitar a interpretação dos resultados e auxiliar na comunicação das descobertas.

De acordo com Silva (2016), a identificação de padrões em séries temporais vem despertando, há algum tempo, interesse da comunidade científica, na resolução de problemas diversos relacionados a diferentes áreas de aplicação, sejam elas na área da saúde, finanças, industriais e ambientais.

A identificação de padrões pode ser definida como a ciência que versa sobre a classificação e descrição de objetos. Em séries temporais, os padrões são partes da sequência que ocorrem com frequência e não são conhecidas a priori. Para Bailão et al. (2020), reconhecimento de padrões é o estudo da forma de organizar os dados e pode ser realizada a partir de objetos representados em diversos tipos de dados.

A extração de recursos de séries temporais é importante porque esses dados podem ter volumes muito grandes e serem muito complexos, o que pode dificultar o processo de análise. Portanto, a utilização de técnicas de aprendizado de máquina para analisar e reduzir a dimensionalidade possibilitando encontrar padrões nesses dados, é fundamental para realizar análises mais precisas e eficientes.

Para Cruz et al. (2016), não é novidade na literatura que estuda ML, o esforço empregado na busca para encontrar representações em dados de toda natureza. Com o propósito de extrair recursos e na descoberta de conhecimento de padrões em séries temporais, vários métodos foram propostos e aplicados.

Essas aplicações utilizando esta abordagem, têm sido relevantes nesses cenários, levando-se em consideração que os dados utilizados para esse tipo de descoberta de conhecimento são dados brutos não rotulados, o tipo de aprendizado empregado nesses casos é o não supervisionado. Segundo Masci et al. (2011), um dos principais objetivos do aprendizado não supervisionado é a extração de recursos e a descoberta de padrões de dados não rotulados, detectando e removendo redundâncias e características consideradas fracas e mantendo as características mais fortes dos dados originais em boas representações.

Uma das áreas de conhecimento que utiliza técnicas de aprendizado de máquina para realizar projetos e pesquisas é a área do meio ambiente (DOURADO; OLIVEIRA; AVILA, 2013). Geralmente esses estudos são realizados utilizando como conjunto de dados algum tipo de série temporal. Todo conjunto de dados obtido através de medições frequentes e sequenciais ao longo do tempo, é considerado uma série temporal (ESLING; AGON, 2012).

Na Amazônia, a precipitação é um componente meteorológico muito importante para a realização de diversas atividades humanas, ambientais, industriais, agrícolas, científicas, etc., esse fenômeno natural possui uma variabilidade alta no tempo e no espaço (GONÇALVES

et al., 2017). Logo, pode-se afirmar que o conhecimento e compreensão das características pluviométricas das regiões são importantes para viabilizar o planejamento do uso, manejo e conservação dos recursos hídricos.

O fenômeno natural da precipitação é um processo fundamental de impacto direto nas bacias hidrográficas e no desenvolvimento humano e ambiental. Para Miranda et al. (2016), a variabilidade desse fenômeno produz implicações importantes na navegabilidade dos rios, sobre a abundância do indivíduo e a riqueza das espécies.

Os impactos socioeconômicos decorrentes da variação da precipitação tornam-na uma das variáveis climáticas de maior relevância. A escassez de chuva pode comprometer o abastecimento de alimentos, enquanto o excesso pode causar danos à infraestrutura das cidades por meio de alagamentos. Nesse sentido, um método que auxilia na análise e gestão da disponibilidade hídrica, e consequentemente no desenvolvimento socioeconômico, é a análise do comportamento das séries temporais de precipitação. (SEVERO; SILVA; TACHINI, 2019)

Considerando a relevância da análise de séries temporais de precipitação, especialmente no contexto amazônico, esse estudo concentrou-se no **desenvolvimento de um método computacional** para análise de séries temporais de precipitação utilizando técnicas de *machine learning* com aprendizado não supervisionado, afim de **propor um método capaz de realizar a extração de características complexas dos dados, obtendo um mapa de atributos em baixa dimensionalidade para reconhecimento de padrões, descoberta de regiões homogêneas com relação à precipitação e reconstrução aproximada de séries temporais em estudo.** O objetivo principal não foi apenas identificar padrões no ciclo anual de chuvas e caracterizar regiões homogêneas em relação à variabilidade de precipitação, mas fundamentalmente **desenvolver um método que fosse capaz de analisar, compreender e extrair conhecimento de séries temporais**.

É importante destacar que o método desenvolvido nesta pesquisa pode ser aplicado não apenas na região amazônica, mas também em outras áreas com desafios semelhantes relacionados à análise de séries temporais de precipitação. A combinação de técnicas de aprendizado de máquina e análise de dados climáticos oferece uma abordagem promissora para compreender a variabilidade e os padrões climáticos, auxiliando na tomada de decisões informadas e no planejamento adequado em diversas áreas, incluindo a gestão de recursos hídricos, agricultura, previsão de enchentes e gerenciamento de riscos ambientais.

O modelo adotado neste estudo para a análise dos dados de precipitação é baseado em uma estrutura de rede neural autocodificadora, também conhecida como *autoencoder* (AE), com camadas convolucionais e aprendizado profundo. Essa abordagem é altamente eficaz para aprender e adquirir conhecimento a partir dos dados de entrada, bem como para extrair características relevantes dos mesmos. Outro fator de destaque do modelo é sua capacidade de reduzir a dimensionalidade dos dados de forma não linear.

O AE é uma arquitetura de rede neural que visa reconstruir de forma aproximada os dados de entrada a partir de uma representação latente, ou seja, uma versão comprimida dos dados originais. A estrutura é composta por duas partes principais: o *encoder*, responsável por mapear os dados de entrada para a representação latente, e o *decoder*, que reconstrói os dados a partir dessa representação. Durante o treinamento, o modelo busca minimizar a diferença entre os dados de entrada e os dados reconstruídos, ajustando os pesos e hiperparâmetros da rede.

Ao utilizar camadas convolucionais, o modelo é capaz de capturar padrões espaciais nas séries temporais de precipitação, levando em consideração a correlação dos padrões expressos entre os dados das estações pluviométricas e a proximidade geográfica das mesmas. Essa abordagem é particularmente relevante para a análise de dados climáticos, uma vez que a distribuição espacial das chuvas desempenha um papel crucial na compreensão dos padrões climáticos e na identificação de regiões com comportamentos semelhantes.

No estudo em questão, além da utilização da estrutura de rede neural autocodificadora, também foi empregada a técnica de *clustering* ou agrupamento. Esta técnica de aprendizado de máquina tem como objetivo analisar os dados de precipitação em baixa dimensionalidade e agrupá-los levando em consideração suas características e padrões identificados. Essa abordagem permite a regionalização das estações pluviométricas com base na similaridade de suas características de precipitação.

A análise das características de precipitação e a identificação de regiões homogêneas por meio das técnicas de agrupamento fornecem conhecimentos valiosos para a compreensão da variabilidade espacial e temporal das chuvas. Essa abordagem é fundamental para estudos relacionados à gestão de recursos hídricos, planejamento de infraestrutura e tomada de decisões em situações que envolvem eventos climáticos, como enchentes e secas.

Diante dessas problemáticas, é notável o aumento significativo de estudos científicos que têm utilizado a análise de *clustering* como ferramenta para auxiliar na compreensão da distribuição espaço-temporal das chuvas, principalmente em regiões com ausência ou escassez de dados, como é o caso da região da Amazônia Legal. Dentre esses estudos, merecem destaque as contribuições de pesquisadores, como Amanajás e Braga (2012), Neves et al. (2017) e LIRA et al. (2019), entre outros.

Muitos desses estudos têm utilizado métodos baseados em hidrologia, geoestatística, testes estatísticos não paramétricos e até mesmo técnicas de aprendizado de máquina, como a Análise de Componente Principal (PCA), para a análise das características das séries temporais de precipitação na região amazônica.

Os dados de precipitação utilizados neste estudo foram obtidos no site da Agência Nacional de Águas e Saneamento Básico (ANA), disponível em https://www.gov.br/ana/pt-br. Esses dados representam os valores do ciclo diário de precipitação de 268 estações pluviométricas localizadas na região da Amazônia Legal, abrangendo o período de 1986 a 2015.

Durante a fase de pré-processamento dos dados, adotam-se duas etapas para garantir a qualidade e a adequação dos dados à aplicação do modelo proposto. Na primeira etapa, realiza-se o preenchimento das lacunas nos dados das séries temporais, a fim de lidar com os valores ausentes. Neste estudo, **optou-se por inserir o valor 0 (zero) nas tabelas de precipitação diária para representar essas lacunas, proporcionando uma abordagem mais completa para a análise subsequente**.

Na segunda etapa do pré-processamento, foi realizada a construção dos dados de entrada para o modelo proposto. Para o treinamento da rede DCAE proposta, são utilizados dois conjuntos de informações: os valores do acumulado mensal de cada ano observado para cada estação pluviométrica analisada e as coordenadas geográficas associadas a cada estação, incluindo a longitude e a latitude.

No desenvolvimento desta dissertação, todas as etapas de codificação foram implementadas utilizando a linguagem de programação Python, juntamente com suas principais bibliotecas e APIs voltadas para ciência de dados e aprendizado de máquina. Essa escolha se deve à flexibilidade, eficiência e vasta gama de recursos oferecidos por essa linguagem, tornando-a amplamente utilizada no campo da IA e ciência de dados.

Os resultados deste estudo demonstram-se altamente promissores, evidenciando a eficácia da abordagem adotada. A análise dos dados utilizando o DCAE revelou uma estrutura de baixa dimensionalidade que capturou as características essenciais dos dados originais de precipitação. A reconstrução aproximada dos dados de entrada apresentou valores RMSE e NRMSE consideráveis, indicando a capacidade do modelo em reproduzir com eficiência as informações importantes contidas nos dados.

A redução da dimensionalidade dos dados com a aplicação do DCAE permitiu a realização de uma análise de agrupamento dos dados de precipitação estudados. Os resultados obtidos revelaram a identificação de regiões homogêneas na Amazônia Legal com relação à essa variável. Isso significa que a metodologia proposta foi capaz de agrupar as estações pluviométricas com base nas características semelhantes de seus padrões de chuva, fornecendo informações valiosas sobre a variabilidade espacial da precipitação na região.

Esses achados são de grande valor, pois contribuem para a compreensão da distribuição e comportamento da precipitação na Amazônia Legal, auxiliando na gestão dos recursos hídricos, no planejamento de atividades agrícolas e no enfrentamento de eventos extremos, como secas e enchentes. A capacidade de identificar regiões homogêneas com base na variável de precipitação abre caminho para estudos mais aprofundados sobre os padrões climáticos da região, bem como para o desenvolvimento de estratégias de adaptação e mitigação dos impactos relacionados à variabilidade pluviométrica.

1.1 Justificativa de Pesquisa

De acordo com Dourado, Oliveira e Avila (2013), as mudanças e variações climáticas, são motivações de estudos, que buscam quantificar e monitorar com mais precisão as variáveis de influência. Um dos componentes climáticos utilizados em estudos de clima são as séries temporais de precipitação. A preocupação com a variabilidade temporal e espacial de precipitação foi tema recorrente em várias pesquisas nos últimos tempos, principalmente pelo impacto que pode causar as mudanças climáticas e a disponibilidade hídrica (PANDEY; KHARE, 2018)

Segundo Brubacher, Oliveira e Guasselli (2020), há um aumento na importância de estudos que visam reconhecer padrões e espacializar chuvas, uma vez que suas aplicações em pesquisas nos campos da climatologia, agricultura, hidrologia, gestão de desastres, planejamento e gestão ambiental são fundamentais para o crescimento socioeconômico de uma determinada região.

Nesse sentido, Dourado, Oliveira e Avila (2013) afirmam que a utilização de técnicas de mineração de dados que auxiliem no processo de descoberta de conhecimento, transformando dados climáticos em informações relevantes, despontam como opções promissoras. De acordo com Sousa, Guedes e Oliveira (2018), modelos computacionais de reconhecimento de padrões em séries temporais de precipitação, que utilizam métodos de ML, extraem características espaço-temporal de padrões históricos desse componente climático que podem ser utilizados em vários estudos, como identificação de regiões homogêneas tendo a chuva como fator principal e também em previsões de chuva.

Para Valente e Maldonado (2020), há algumas vantagens na utilização de ML para identificação e seleção de recursos em dados, destacando maior potencial na compreensão dos dados originais, generalização de novos objetos e redução de custos de coleta de dados. Determinar qual técnica desta abordagem utilizar é uma tarefa importante no processo de descoberta de conhecimento, pois requer análise e conhecimento profundo do problema.

Neste campo, existem várias metodologias promissoras que podem ser utilizadas no tópico de reconhecimento de padrões e descoberta de conhecimento. Muitos estudos que abordam *deep learning* (DL) vêm se destacando com resultados satisfatórios nesse contexto. De acordo com Li e Sano (2020), a utilização de técnicas computacionais que possam produzir tal conhecimento é importante para o desenvolvimento de vários estudos em diversas áreas, uma vez que, a extração de recursos de forma manual teria um custo humano e de tempo alto e requer dos agentes envolvidos um bom conhecimento sobre os domínios estudados.

No estudo de David e Netanyahu (2016), foi utilizado uma estrutura *convolutional autoencoder* (CAE) que demonstrou em seus resultados que o método empregado foi capaz de extrair conhecimento de dados de pinturas. Já na pesquisa de Baia e Castro (2018), duas estruturas de CAE foram utilizadas para reconhecer padrões em dados de sinal de eletrocardiogramas. Na primeira arquitetura o CAE foi treinado para reconhecimento de padrão de pacientes com

batimentos cardiacos normais, enquanto que a segunda aquitetura treinou com dados de pacientes com arritmia, obtendo uma boa extração de características dos sinais utilizados para treinamento. Em Buzuti e Thomaz (2019), onde arquiteturas de CAE também foram utilizadas, os resultados apresentaram recursos de alto e baixo nível, definindo os padrões dos dados de entrada.

A utilização de métodos de DL é muito comum no cénario atual da pesquisa científica que envolva estudos de ML, seja em trabalhos com aprendizado supervisionado como não supervisionado. Uma das arquiteturas de rede neural proposta é o *deep convolutional autoencoder*, muitos estudos vêm sendo realizados utilizando essa abordagem de apredizado e o campo de aplicação são os mais variados possíveis, como por exemplo, análise de sinais e imagens na área da saúde, análise de sinais sonoros no campo do processamento de linguagem natural, processamento de imagens nas áreas da arte e análise de variáveis climáticas no campo do meio ambiental, entre outras.

Huang et al. (2017) utilizam de forma eficaz técnicas de ML, especificamente uma estrutura DCAE, para analisar e descobrir conhecimento em séries temporais complexas de dados de *Task-based functional magnetic resonance imaging* (tfMRI).

A abordagem não supervisionada da pesquisa permite que o modelo aprenda automaticamente padrões nos dados, sem a necessidade de rótulos prévios ou supervisão externa. Isso pode ser especialmente útil em situações em que os dados são abundantes, mas os rótulos são escassos ou não estão disponíveis.

Os resultados promissores da pesquisa sugerem que a técnica pode ter aplicações valiosas em áreas como neurociência cognitiva e clínica. A combinação de técnicas de ML e análise de séries temporais é uma área de pesquisa em constante evolução, e espera-se que continue a produzir soluções avançadas para o processamento e análise de dados temporais.

No trabalho de Essien e Giannetti (2020), foi abordada uma técnica eficaz para prever a velocidade de máquinas em várias etapas de produção, utilizando uma arquitetura de aprendizado profundo do tipo DCAE em conjunto com *Long Short Term Memory* (LSTM).

A combinação de técnicas de AE e LSTM, é uma abordagem poderosa para processar e analisar dados temporais, especialmente em aplicações industriais. A capacidade de prever com precisão a velocidade de máquinas em várias etapas de produção pode levar a melhorias significativas na programação e planejamento da produção, o que pode ser valioso em processos de manufatura inteligentes. Além disso, a pesquisa ilustra como técnicas avançadas de DL, podem ser aplicadas com sucesso a séries temporais para obter resultados precisos e de alta qualidade.

A detecção de anomalias em séries temporais de *Internet of Things* (IoT) é um tema de grande interesse em muitas aplicações, como a monitoração de equipamentos, infraestrutura, saúde e segurança. O uso de técnicas como o *recurrent convolutional neural network*, proposto por Yin et al. (2020), pode ser uma forma eficaz de identificar padrões anômalos nessas séries

temporais de dados. Esse método pode fornecer informações importantes sobre as causas dessas anomalias e ajudar a tomar medidas para prevenir problemas futuros.

Chen et al. (2017a) propôs um estudo para reconstrução com eliminação de ruídos em imagens de tomografia computadorizada de baixas dosagens. Este estudo utilizou uma arquitetura de *deep autoencoder* com camadas convolucionais, e os resultados apresentados foram considerados competitivos em comparação com outros métodos presentes na literatura. A utilização de técnicas como *autoencoders*, tem se mostrado uma ferramenta eficaz para a eliminação de ruídos em imagens, o que pode levar a uma melhora na qualidade da imagem e a uma interpretação mais precisa dos dados.

Outro estudo que utilizou a arquitetura de um DCAE foi a pesquisa de Baia e Castro (2018) que elaborou uma proposta de sistema de classificação de sinais em eletrocardiogramas. Os resultados obtidos foram promissores uma vez atingida a acurácia de 88,9% considerando a base de dados de teste.

Desta forma, apresenta-se o modelo proposto desta pesquisa que dispõe de uma arquitetura baseada em DCAE que analisa as séries temporais de precipitação e extrai um mapa de atributos bidimensional que contém características complexas e relevantes que representam as informações espaço-temporal da região estudada. Os atributos obtidos possibilita a análise dessas características e a descoberta de padrões utilizando técnicas de *clustering*. Ao separar as informações em grupos que possuem similaridade em seus dados, torna-se possível a análise descritiva, identificação de padrões e o comportamento do seu ciclo.

Na jornada em preencher algumas das lacunas existentes com relação a este tema, este trabalho busca em seu conceito principal responder a alguns questionamentos que irão nortear o estudo. Desta forma, busca-se responder quatro perguntas de pesquisa:

1. **O modelo de *deep convolutional autoencoder* proposto é capaz de gerar um mapa de características no espaço latente que possa realizar a identificação de padrões em séries temporais de precipitação?**
2. **Qual a melhor configuração para esse modelo?**
3. **O modelo apresenta um bom nível na reconstrução aproximada dos dados originais?**
4. **O modelo apresentou bom nível de agrupamento das estações pluviométricas, destacando a existência de regiões homogêneas com relação a precipitação na Amazônia legal?**

1.2 Objetivo Geral

O objetivo geral deste estudo é o **desenvolvimento de um método computacional** para análise de séries temporais de precipitação utilizando técnicas de *machine learning* com

aprendizado não supervisionado, afim de **propor um método capaz de realizar a extração de características complexas dos dados, obtendo um mapa de atributos em baixa dimensionalidade para reconhecimento de padrões, descoberta de regiões homogêneas com relação à precipitação e reconstrução aproximada de séries temporais em estudo.**

1.3 Objetivos Específicos

São objetivos específicos almejados neste trabalho:

- Identificar de forma comprimida um mapa de características que represente os dados originais;
- Gerar novos conhecimentos dos dados estudados a partir da representação em baixa dimensionalidade encontrada;
- Implementar uma nova abordagem para reconhecimento de padrões de séries temporais com dados de precipitação utilizando técnicas de ML;
- Avaliar e comparar os resultados obtidos com outras abordagens de análise de séries temporais de precipitação.

1.4 Estrutura do Trabalho

As próximas seções da presente pesquisa estão estruturadas da seguinte forma: 2) Fundamentação Teórica; 3) Metodologia da Pesquisa; 4) Resultados e Discussão; 5) Conclusão e Referências Bibliográfica.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os principais conceitos e técnicas estudados no campo da identificação de padrões em séries temporais de precipitação utilizando métodos de ML não supervisionado com foco em DL, especificamente o método DCAE.

2.1 Séries Temporais de Precipitação

Com o advento do desenvolvimento tecnológico e o aumento do uso de técnicas computacionais mais avançadas em processos diferentes e em vários campos de trabalho e estudo, iniciou-se a escalada na geração e análise de grandes volumes de dados ordenados ao longo do tempo. De acordo com Hu et al. (2018), grandes volumes de observações durante o tempo são gerados através do uso de dispositivos e equipamentos com sensores, como celulares, navegadores GPS, monitores de saúde etc., isso faz parte do processo de digitalização da sociedade e indústria que está em curso.

No contexto atual, a busca por informações se tornou fundamental para o desenvolvimento de processos e a descoberta de conhecimento através de dados temporais é um método importante. Conhecimentos implícitos de processos que geram séries temporais são extraídos durante a análise destas séries temporais, permitindo-se conhecer e compreender melhor tais processos (YANG; GUO; JENSEN, 2013).

Devido a esta busca e necessidade de conhecimento, séries temporais passaram a ser extensivamente produzidas e utilizadas em várias áreas de aplicação (YANG et al., 2018), exemplo, biologia, finanças, saúde e meio ambiente. Dentro da extensão ambiental a descoberta de conhecimento através da análise das séries temporais de precipitação é uma das mais estudadas devido sua grande importância. A precipitação é o elemento que melhor define as mudanças climáticas da região tropical, além de influenciar no comportamento de outros elementos atmosféricos, tais como, temperatura do ar e umidade relativa do ar (SOUSA et al., 2015).

Segundo Sena et al. (2020), a precipitação é fundamental para manter o nível da água nas bacias hidrográficas, afinal, trata-se da principal fonte desse recurso natural. Diante de tal importância conhecer suas características e analisar as informações derivadas deste conhecimento é fato imprescindível para a gestão de recursos hídricos e o desenvolvimento humano. Corroboram com este pensamento Lyra, Oliveira-Júnior e Zeri (2014), ao destacar como ponto decisivo o conhecimento sobre informações de precipitação de uma região para determinar quais ações, em campos de áreas distintas, devem ser adotadas.

Para Delahaye et al. (2015), devido à grande mudança espacial e temporal influenciada pela dinâmica atmosférica da região amazônica, o processo de análise das precipitações acaba se tornado complexo. Sendo assim, para Almeida et al. (2015), entender a dinâmica pluviométrica da região amazônica não é apenas importante mas fundamental.

De acordo com Ishihara et al. (2013), há um grande problema de monitoramento pluviométrico na região Amazônica, devido a baixa densidade de estações instaladas, sendo que várias já se encontram desinstaladas, e outras apresentam muitos erros e descontinuidades nos dados, dificultando a análise de séries temporais na região.

A proposta deste trabalho, que utiliza técnicas de ML, como o modelo DCAE, é de extrema importância para o desenvolvimento de estudos e o crescimento socioeconômico de regiões onde os dados coletados de precipitação não são confiáveis ou mesmo onde não existem equipamentos de coleta adequados.

Por meio do DCAE, é possível estabelecer representações de padrões que permitam simular o comportamento das chuvas nesses locais. Isso significa que, mesmo na ausência de dados confiáveis ou estações de coleta, é possível obter estimativas e informações valiosas sobre os padrões de precipitação na região.

Ao fornecer uma ferramenta capaz de simular o comportamento da precipitação, mesmo em áreas com dados escassos, essa metodologia contribui diretamente para o desenvolvimento regional. Possibilitando uma melhor compreensão dos padrões climáticos, o que pode auxiliar na tomada de decisões informadas, no planejamento de atividades econômicas e na adoção de medidas de adaptação às mudanças climáticas.

Portanto, essa proposta de utilizar técnicas de ML, como o modelo DCAE, representa um avanço significativo no campo da análise de precipitação em regiões com dados limitados ou ausentes. Ela não apenas possibilita uma compreensão mais abrangente do clima, mas também promove o desenvolvimento socioeconômico sustentável da região, abrindo caminho para novas oportunidades de pesquisa e aplicação de soluções inovadoras.

2.2 Aprendizado de Máquina

ML é um campo da inteligência artificial que evoluiu do estudo de reconhecimento de padrões. Com a finalidade de desenvolver técnicas computacionais com a capacidade de aprender continuamente com os dados de forma supervisionada ou não supervisionada permitindo que as máquinas se adaptem a novos contextos de forma independente.

Para Heylman et al. (2015), o ML é geralmente definido como o processo de treinamento de um algoritmo para realizar previsões ou decisões com base em dados. Já para Carter et al. (2019), ML é uma técnica flexível e eficiente de identificação de padrões em grandes e complexos conjuntos de dados.

No ML são consideradas tarefas de previsão e descrição. Prever resultados, para entradas de dados anteriormente não conhecidos, através do modelo encontrado a partir dos dados de treinamento, é denominado previsão. A tarefa de descrição utiliza aprendizado não supervisionado para explorar ou descrever um conjunto de dados, detectando de forma espontânea os

padrões ou estruturas a partir dos próprios dados (FACELI et al., 2011). Dois dos principais tipos de aprendizado em ML acontecem de forma supervisionada e não supervisionada (ZOU et al., 2020).

2.2.1 Aprendizado de Máquina Supervisionado

O ML supervisionado tem como principal foco a classificação de objetos, que através do conjunto de dados já rotulados, realiza para novas instâncias uma predição com base no treinamento do modelo com os dados com rótulos conhecidos anteriormente. Para Moro et al. (2019), nesse tipo de aprendizado cada dado é mapeado realizando a previsão para uma classe cujo rótulo já é conhecido previamente.

De acordo com Carter et al. (2019), o aprendizado supervisionado envolve dados rotulados, enquanto o aprendizado não supervisionado permite que o algoritmo encontre padrões sem considerar a qual classe os dados pertencem originalmente. Já para Heylman et al. (2015), o aprendizado supervisionado é um subtipo de ML no qual um conjunto de dados de treinamento com classificações ou resultados conhecidos é usado para treinar o algoritmo criando um modelo estatístico que se ajusta aos dados de treinamento.

Geralmente problemas de ML supervisionado são classificados em problemas de regressão ou classificação. Problemas de regressão ocorrem quando tenta-se prever valores para uma determinada ocorrência futura com base em observações específicas apresentadas, esse valor pode ser o volume de chuva mensal, a nota de um aluno, sua altura etc. Já problemas de classificação ocorrem quando a proposta é encontrar uma classe, dentro de um escopo determinado, como se o aluno está aprovado ou reprovado.

2.2.2 Aprendizado de Máquina Não Supervisionado

No aprendizado não supervisionado, como ocorre nos métodos de agrupamento, não há a necessidade de informações a priori sobre o domínio avaliado, levando-se em consideração, apenas, a disposição dos dados e suas propriedades internas.

Neste modelo de aprendizado, que também é chamado de agrupamento ou análise exploratória de dados, os rótulos dos dados não estão disponíveis. De certa forma, pode-se dizer que os rótulos estão presentes na atividade de agrupamento, cada *cluster* (grupo) formado leva ao entendimento da origem de um rótulo, mas esses rótulos são oriundos da análise dos dados de entrada. O problema de *clustering* é o de separar um conjunto de dados em um número de *clusters* com base em alguma medida de similaridade.

Segundo Corcovia e Alves (2019), quando o rótulo da classe de cada amostra de dados de treinamento é desconhecido, e o número ou conjunto de classes a ser treinado pode não ser conhecido a priori, trata-se de uma abordagem de aprendizado não supervisionado. Além disso,

esta metodologia é também descritiva, pois descreve de forma concisa os dados disponíveis, fornecendo características das propriedades gerais dos dados.

Uma das técnicas de aprendizado não supervisionado bastante utilizada em diversos estudos é o *clustering*. Para Bouveyron e Brunet-Saumard (2014), *clustering* é uma ferramenta de análise de dados que visa agrupar dados em vários grupos homogêneos. O problema de agrupamento é estudado há anos e geralmente ocorre em aplicações para as quais é necessária uma partição dos dados.

2.2.2.1 *Clustering*

A análise de *clustering* é uma técnica que apresenta vários métodos, baseando-se principalmente em separar dados em grupos, levando-se em consideração as principais características e a similaridade entre elas, formando agrupamentos homogêneos e com dados semelhantes entre si, e ao mesmo tempo diferente entre os grupos (CRISPIM et al., 2020).

Segundo Yim e Ramdeen (2015), em vários campos de atuação, principalmente em muitas disciplinas científicas, também se faz necessário e fundamental identificar similaridade nos dados para construir grupos significativos. O objetivo central do método de *clustering* é descobrir um sistema de organização de observações onde membros do mesmo grupo compartilham propriedades específicas em comum, aumentando dentro do grupo a homogeneidade e a heterogeneidade entre grupos.

Para Crispim et al. (2020), existem várias técnicas de *clustering* que podem ser empregadas, que com base em conjuntos de dados, buscam traçar uma análise de dados, como também identificar as semelhanças e diferenças entre os mesmos. Dentre as técnicas existentes, pode-se citar o método aglomerativo hierárquico que em sua estrutura organiza os dados mais semelhantes, aglomerando-os de acordo com as características mais próximas ou em comum.

A análise de *cluster*, é encontrada na literatura como a técnica bem empregada na busca por estabelecer padrões espaciais para as variáveis meteorológicas, e dentre várias metodologias a de Ward (1963) tem sido bastante utilizada em diversos estudos apresentando bons resultados (OLIVEIRA-JÚNIOR et al., 2017).

2.2.3 Redes Neurais Artificiais

Redes neurais artificiais (RNA) fazem parte do conjunto de técnicas pertencentes ao ML, possuindo a nomenclatura “neural” em seu nome devido as inspirações no funcionamento do cérebro humano, modelando a forma como o cérebro responde a uma determinada tarefa. O alto desempenho das RNA está ligado à robusta interligação existente entre nódulos computacionais, conhecidos como "neurônios"ou células de processamento (HAYKIN, 2001).

De acordo com Karimpouli e Tahmasebi (2019), dentro do campo da ML existem vários algoritmos que podem ser utilizados para diversas tarefas, a RNA é uma classe desses algoritmos

que tem seu funcionamento espelhado nas funções sinápticas do cérebro humano. As RNAs podem realizar tarefas de predição ou classificação conforme os dados e o treinamento por elas utilizados.

As RNAs possuem características de aprendizado de regras de forma automática lidando bem com problemas de generalização, após treinar e aprender a solucionar um determinado problema essa rede poderá solucionar com bom desempenho problemas de natureza parecida.

Para Spörl, Castro e Luchiari (2011), é bastante comum a aplicação de RNA para solucionar problemas de alta complexidade, onde o conhecimento prévio do comportamento das variáveis não é algo estritamente necessário. Dentre várias características importantes destaca-se sua abordagem de aprendizado através de exemplos transformando e generalizando a informação aprendida. São utilizadas nos mais variados campos do conhecimento, destacando-se em aplicações que envolvam reconhecimento de padrões, análises de séries temporais, diagnósticos médicos entre outros.

As RNAs podem ser classificadas de acordo com com seu funcionamento e o modelo adotado em suas camadas, um dos modelos bastante utilizado atualmente é o de abordagem com estruturas convolucionais dentro da arquitetura da rede.

2.2.4 Redes Neurais Profundas

Quando abordado o ML, deve-se abrir um destaque especial para o *deep learning*, pois devido seu avanço significativo nos últimos anos, DL se tornou importante em vários campos de estudo e obteve sucesso em várias aplicações práticas (GOODFELLOW; BENGIO; COURVILLE, 2016).

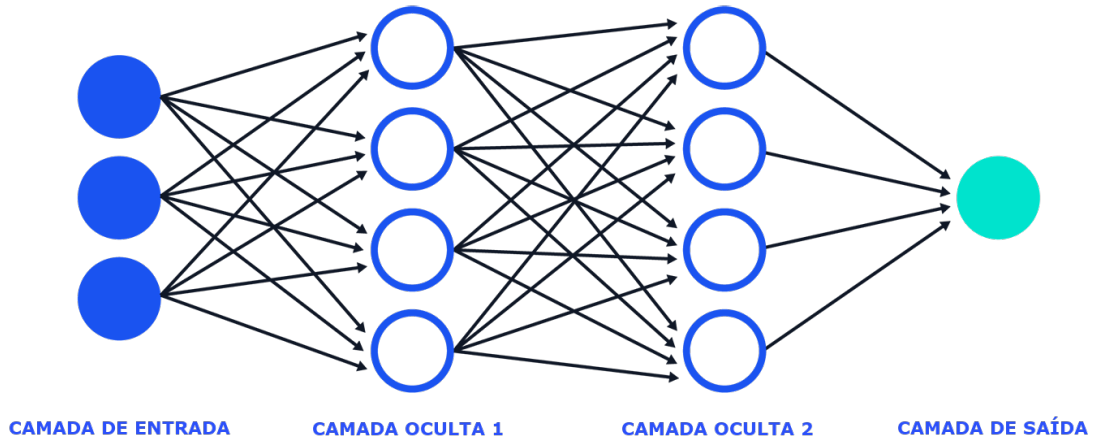
O conhecimento que modelos computacionais com estrutura de DL absorvem são, fundamentalmente, representações com vários níveis de abstração (LECUN; BENGIO; HINTON, 2015). É extenso o número de áreas de atuação que vem apresentando aplicações de sucesso com esses métodos, como por exemplo, processamento de linguagem natural, visão computacional, entre outros (MORAES; CASTRO, 2018). Os recursos complexos não lineares que são extraídos e aprendidos de forma hierárquica, são considerados as melhores características ou representações dos dados originais, e os modelos de DL são capazes de gerar tal representação (ZHANG et al., 2017).

A DL vem atraindo bastante atenção dos pesquisadores de ML e mineração de dados, seus resultados têm comprovado sua excelência no aprendizado hierárquico de características complexas de alto e médio nível de dados originais de baixo nível.

A estrutura básica da arquitetura de DL, como representado na figura 1, é normalmente formada por uma rede profunda de camadas conectadas com vários blocos de construções semelhantes amontoados de forma ordenada um após o outro. A base da estrutura ou a camada inicial recebe os dados brutos de entrada que processa e entrega seus dados de saída como dados

de entrada para a camada subsequente, e assim é realizado por toda a estrutura até chegar a última camada.

Figura 1 – Representação da estrutura básica da arquitetura de *deep learning*.



Fonte: Próprio Autor

De acordo com Ponti e Costa (2018), é uma realidade no contexto atual de estudos na área de ML, que a abordagem de DL reune métodos na tarefa de análise de dados de multimídias, como imagens e vídeos, análise de sinais como áudio e fala, bem como conteúdo textual. No entanto, vários são os modelos, algoritmos e componentes que compõem esses métodos. Segundo Buzuti e Thomaz (2019), na área de estudos relacionados ao aprendizado não supervisionado, como o de reconhecimento de padrões, pesquisas com abordagem em DL apresentaram resultados satisfatórios, superando outros resultados apresentados na literatura como, por exemplo, em Hinton e Salakhutdinov (2006) e em LeCun, Bengio e Hinton (2015).

No reconhecimento de padrões, a seleção de características é um processo fundamental para obter os melhores recursos extraídos para representação dos dados originais. Métodos utilizando DL têm apresentado números cada vez maiores em suas aplicações, e resultados melhores quando comparado aos métodos tradicionais (CHEN et al., 2017b).

Bengio, Courville e Vincent (2013), apresentam em seu estudo vantagens da utilização de DL para extração de conhecimento. Para Hwang e Chen (2017), com a DL, características simples são extraídas dos dados de entrada, e posteriormente recursos mais complexos são aprendidos através das outras camadas que fazem parte da arquiterura da DL.

De acordo com Moraes e Castro (2018), pesquisas utilizando DL apresentam bons resultados para problemas que envolvem aplicações de dados com características 2D. Isso não exclui a possibilidade da investigação de desempenho para problemas que apresentam dados 1D, como dados de séries temporais. A literatura já apresenta alguns autores que vêm pesquisando a aplicação das redes neurais profundas utilizando dados com essas características

(PENHA; CASTRO, 2017). Moraes e Castro (2018) destacam as unidades LSTM, as redes neurais autocodificadoras, as redes autocodificadoras empilhadas (*Stacked Autoencoders*) e as redes neurais convolucionais (CNN) como as principais redes de DL para resolução de problemas com dados 1D.

2.2.4.1 Redes Neurais de Codificação Automática - *Autoencoders*

De acordo com Lee e Carlberg (2020), existem vários tipos de redes neurais *feedforward*, e dentre elas consta a rede classificada como AE, que possui em seu escopo de características a busca por aprender o mapeamento de identidade, apresentando como resultado uma reprodução aproximada dos dados de entrada da rede neural.

Apenas aprender o padrão do mapa identidade e reproduzi-lo novamente de maneira aproximada é uma tarefa que não traz muitos efeitos práticos. Para que esse processo possa ser algo útil ele precisa vir atrelado a redução de dimensionalidade, compactando os dados de entrada e reproduzindo-o posteriormente.

A redes neurais de codificação automática são uma vertente da rede *multilayer perceptron* (MLP), onde a camada de entrada possui a mesma quantidade de neurônios da camada de saida, sendo que a característica principal do AE é a redução de dimensionalidade que ocorre nas camadas intermediárias até o espaço latente. De acordo com Essien e Giannetti (2020) e Xia et al. (2015), o AE clássico é uma rede neural *feedforward*, onde a camada de entrada e a camada de saída possuem a mesma quantidade de neurônios e as camadas ocultas intermediárias possuem menos neurônios.

o Funcionamento da arquitetura dos *autoencoders* consiste basicamente de dois processos que realizam, primeiramente, a codificação dos dados de entrada que encontram-se em alta dimensionalidade mapeando-os de forma não linear para uma espaço de baixa dimensionalidade, e em um segundo momento, acontece a decodificação que mapeia de forma não linear o espaço de baixa dimensionalidade reproduzindo de maneira aproximada os dados originais. Da mesma maneira (D'ANGELO; PALMIERI, 2021), define uma das especialidades dos AE a função de representar de forma aproximada os dados de entrada usando combinações não lineares dos mapas de recursos extraídos.

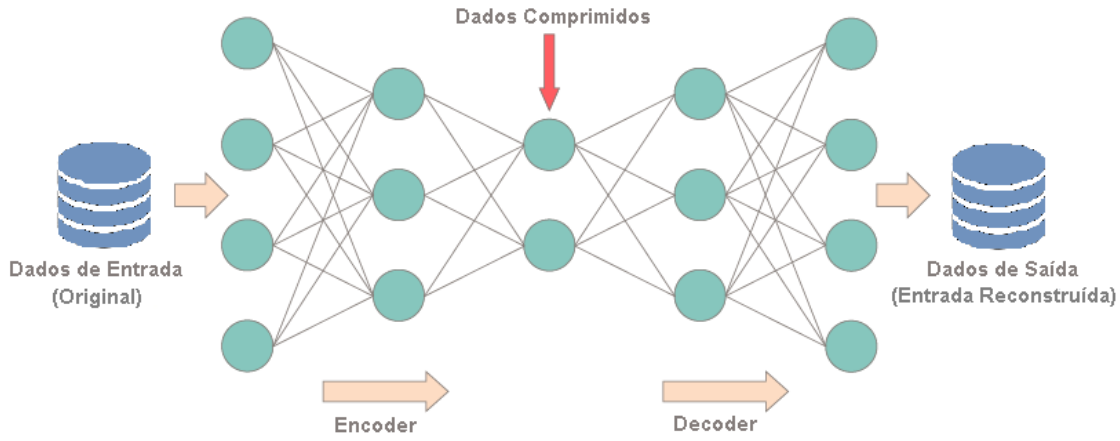
Para Wang, Yu e Wang (2022), AE além de possuírem a capacidade de extração de características em dimensionalidades menores dos que as apresentadas nos dados de entrada, também são capazes de detectar estruturas repetitivas, contribuindo significativamente para a redução e agrupamento da dimensionalidade dos dados em séries temporais.

Esses algoritmos são classificados, quanto ao seu aprendizado, como não supervisionados, que visam a descoberta de conhecimento dos dados brutos de entrada mapeando a função identidade entre a camada de entrada e saída, reconstruindo uma saída com valores aproximados aos da entrada da rede. Para Chen et al. (2017b), depois do processo de iterações o valor da

função de custo do algoritmo AE encontra seu ponto ótimo, fazendo com que os valores dos dados reconstruídos aproximem-se de forma maximizada dos valores dos dados da entrada original. De uma forma mais simples e objetiva Kieu et al. (2019), define a produção do AE como uma saída que reconstrói sua entrada removendo ao máximo os ruídos existentes.

Para conseguir realizar essa tarefa de reconstrução de dados de entrada os AE utilizam uma estrutura básica com duas partes, o codificador e o decodificador, envolvidos em três camadas de processamento. A figura 2 demonstra a estrutura básica do AE.

Figura 2 – Representação da arquitetura básica de um *autoencoder*.



Fonte: Próprio Autor

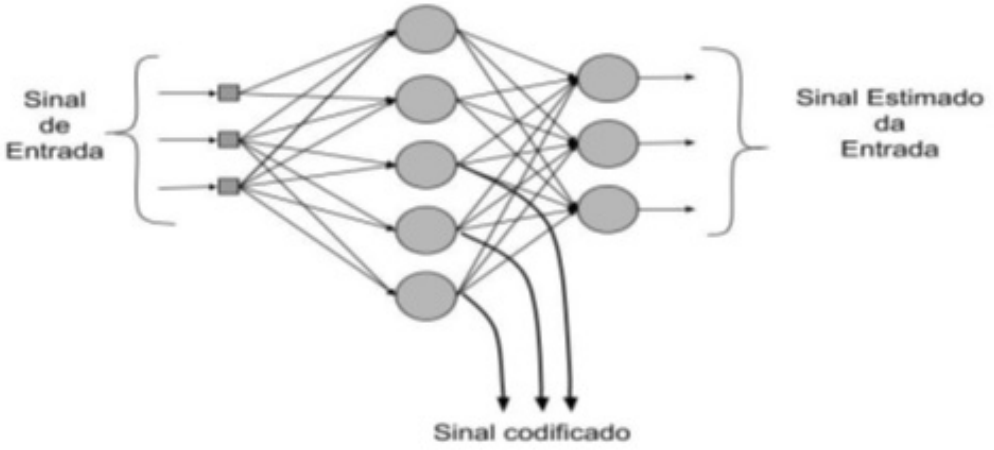
Segundo Wang, Zhao e Wang (2019), a estrutura básica do AE contém a camada do codificador, a camada oculta (espaço latente) e a camada do decodificador, onde a saída do codificador é a entrada da camada oculta e essa mesma camada oculta será a entrada da camada do decodificador. O codificador, como o próprio nome sugere, codifica os dados de entrada comprimindo-os para o espaço latente, essa camada oculta mapeada armazena as principais características extraídas da camada de entrada e posteriormente esse espaço latente será a entrada do decodificador, e por fim, o decodificador descomprime a camada oculta mapeada em uma saída com valores semelhantes aos da entrada menos o ruído. Devido a característica de compressão, o AE pode ser usado como um algoritmo estratégico para redução de dimensionalidade de séries temporais (BENGIO; GOODFELLOW; COURVILLE, 2015).

A parte oculta da camada codificadora tem por finalidade extrair os principais recursos dos neurônios de entrada, funcionando como uma peça de extração de características. Só serão mapeados na camada do espaço latente o conhecimento diretamente ligado às características fundamentais do conjunto de dados de entrada.

Outra parte importante que merece destaque na estrutura das camadas intermediárias dos AE é quanto ao número de neurônios existentes dentro dessas camadas. Para Moraes e Castro (2018), é chamado de AE subcompleto os que apresentam em sua estrutura o número de neurônios nas camadas intermediárias menores que os da camada de entrada e saída, e para os

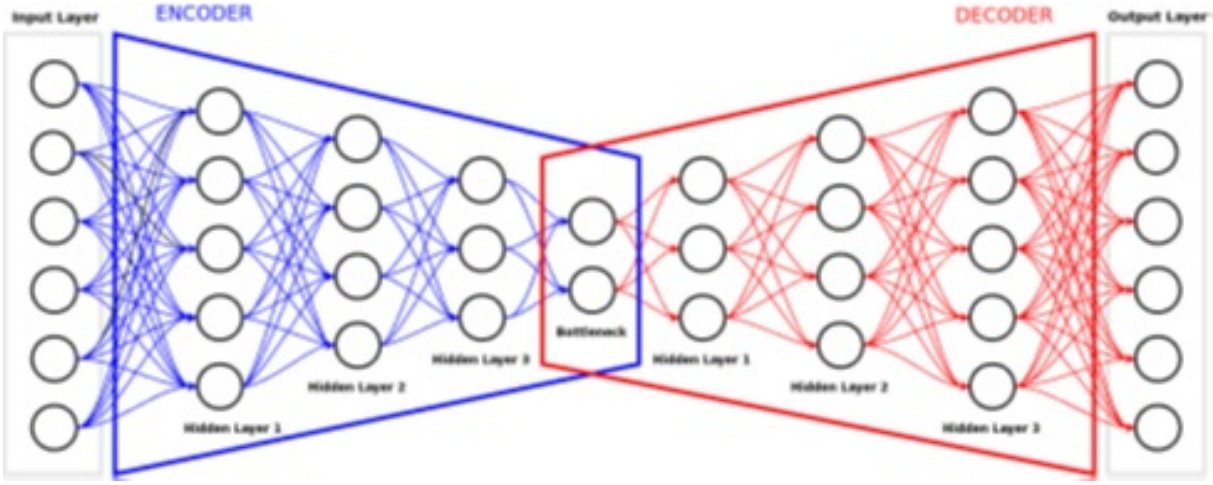
que possuem números maiores são classificados como AE completo. As figuras 3 e 4 apresentam representações das estruturas completa e subcompleta respectivamente.

Figura 3 – Representação da estrutura de AE completo.



Fonte: Moraes e Castro (2018)

Figura 4 – Representação da estrutura de AE subcompleto.



Fonte: Buzuti e Thomaz (2019)

As denotações matemáticas que representam o funcionamento do processamento das unidades envolvidas em uma rede neural de codificação automática são demonstradas a seguir. A unidade de codificação que recebe os dados de entrada é dada pela equação 2.1.

$$z = f(x) \tag{2.1}$$

Onde z é o espaço latente que denota a quantidade de recursos extraídos dos dados de entrada que foram codificados, x denota os dados de entrada, que neste estudo são séries temporais de precipitação e informações de localização geográficas. Já o processamento da unidade de decodificação da rede, que tem por finalidade a reconstrução aproximada dos dados originais a partir dos dados do espaço latente, é dada pela equação 2.2.

$$\hat{x} = g(z) \tag{2.2}$$

Onde $\hat{x}$ são os dados de saída reconstruídos a partir dos recursos extraídos dos dados originais, z denota a representação em baixa dimensionalidade dos dados originais que agora servem de entrada para a unidade de decodificação.

Trabalhos são apresentados na literatura utilizando diversas topologias de rede nas unidades de processamento do AE. Segundo (Castro et al., 2012), a rede MLP é frequentemente utilizada na metodologia de rede de condificação automática. Da mesma forma que as redes MLP são utilizadas na estrutura de processamento do AE as redes neurais convolucionais também podem ser utilizadas trazendo toda as especificidades do processamento típico desse tipo de rede para dentro da estrutura do AE (BAIA; CASTRO, 2018).

2.2.4.2 *Autoencoder* Convolucional Profundo

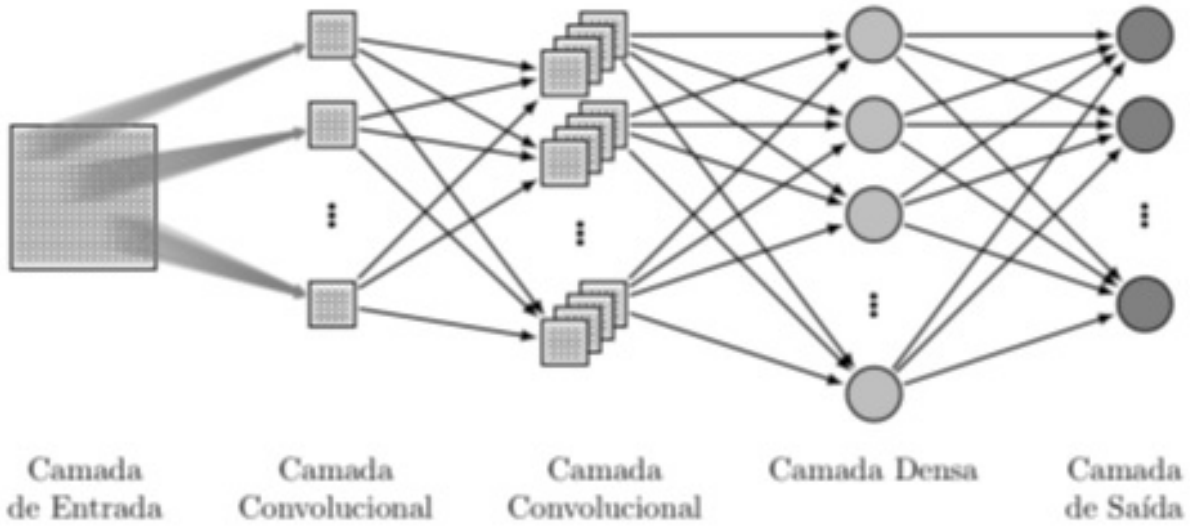
Dentro da abordagem dos algoritmos de codificação automática existem várias metodologias que podem ser utilizadas a depender do tipo de estrutura que será adotada nas duas unidades de processamento existentes (encoder e decoder). No estudo de Hinton e Salakhutdinov (2006), uma abordagem de DL utilizando *deep autoencoder* (DAE) é proposta para aprender e extrair conhecimentos em baixa dimensionalidade de características de um conjunto de imagens.

Uma outra abordagem que pode ser utilizada dentro da estrutura de um DAE é a *convolutional neural network* (CNN) nas camadas intermediárias. A característica principal da arquitetura de aprendizado profundo da rede neural convolucional é a utilização de filtros convolucionais e camadas de pooling para ajudar a extrair parâmetros relevantes dos dados de entrada (BAIA; CASTRO, 2018).

Segundo Barino e Santos (2020), a camada convolucional recebe uma matriz de entrada, onde aplica-se a convolução extraindo matrizes de menor ordem, conhecidas como *kernels*. Obtidos os *kernels*, seleciona-se alguns valores desta operação transformando-se em novas matrizes de menor ordem. Esta operação é conhecida como *pooling*. Após todo esse processo de filtragem e concentração das informações da matriz inicial, a matriz resultante é chamada de *features maps* ou mapas de características. Esses mapas podem ser utilizados como novas entradas de outras camadas convolucionais, em casos de DL. Para Guo et al. (2017), as CNN são redes neurais compostas de várias camadas, cujo seus neurônios realizam aprendizado individual

, e com os filtros empregados são capazes de extrair recursos que servem para o reconhecimento de padrões dos dados de entrada. Na figura 5 representa-se o funcionamento básico da CNN.

Figura 5 – Representação da arquitetura de uma *convolutional neural network*.



Fonte: Sakurai (2017)

Com os avanços no desenvolvimento de RNAs, especialmente devido a adoção da DL, novas alternativas foram desenvolvidas e apresentaram novas maneiras de lidar com problemas mais complexos, principalmente na análise de imagens. Dentro dessas novas alternativas destaca-se o algoritmo de CNN, que utiliza em sua estrutura funções convolucionais e agrupamentos para extração de características de alto nível e recursos para análise dos dados de entrada (KARIMPOULI; TAHMASEBI, 2019).

A arquitetura da rede CNN é capaz de assumir várias formas de configuração, isso é possível devido as características diferentes das camadas da rede e a variação de seus parâmetos que podem ser combinados (ALMEIDA et al., 2019). Convolução é combinação de duas instâncias para resultado de uma terceira. Essa combinação é realizada através da somatória do produto obtido durante o deslocamento que existe ao longo do campo onde há uma sobreposição entre as duas instâncias no processo.

Para D'Angelo e Palmieri (2021), as CNNs têm se destacado pela sua eficiência principalmente nos campos de processamento de imagem, reconhecimento de objetos, direção autônoma e processamento de linguagem natural. Mesmo que inicialmente as CNNs tenham sido utilizadas e obtendo destaque nas tarefas de processamento de imagem, na atualidade a utilização e destaque das CNNs estão sendo notadas em vários outros seguimentos.

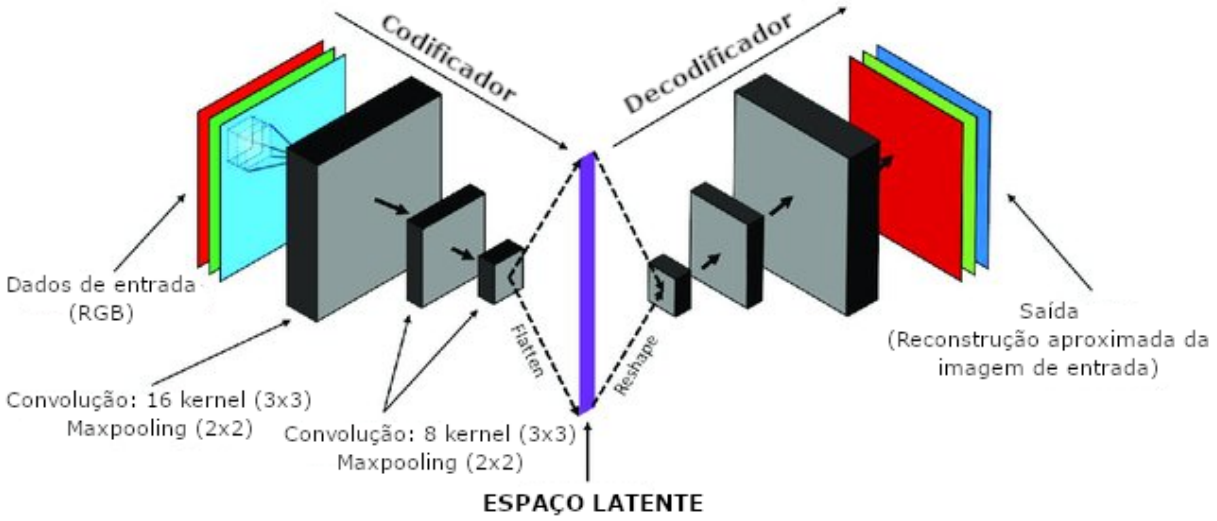
As redes convolucionais têm várias propriedades benéficas, como invariância de tradução, compartilhamento de peso e eficiência computacional. Essas propriedades os tornam especialmente adequados para tarefas de visão computacional, como reconhecimento de imagem, segmentação ou detecção de objetos. Suas propriedades também são úteis para o processamento

de séries temporais, onde tipicamente são empregadas convoluções 1D (WANG; YU; WANG, 2022).

Dois pontos fundamentais têm tornado comum o uso das CNNs para DL. As CNNs aprendem diretamente através de suas camadas convolucionais a extrair recursos, excluindo a necessidade de retirada manual. E estas CNNs podem ser retreinadas para novas tarefas de aprendizado (D'ANGELO; PALMIERI, 2021).

As DCAEs, são uma espécie de rede neural que possui em sua estrutura de codificação e decodificação estrutura do tipo convolucional. Na etapa de codificação a rede convolucional aprende a codificar os dados de entrada em um conjunto de recursos em dimensionalidade reduzida, e na etapa de decodificação a rede CNN tenta reconstruir de forma aproximada os dados de entrada do modelo. Desta forma, a CNN atua como um extrator principal de recursos complexos visando aprender de forma mais eficiente os recursos dos dados de entrada (WANG; YU; WANG, 2022). Na figura 6 o funcionamento da estrutura de DCAE é representada.

Figura 6 – Representação de uma estrutura de *deep convolutional autoencoder* padrão.



Fonte: Jiménez et al. (2020)

De acordo com (WANG; YU; WANG, 2022), quando trata-se de grandes quantidades de dados e sua variedade o DCAE utilizando o método de redução de dimensionalidade pode extrair características complexas e profundas desses dados massivos, garantindo que os mesmos possam ser projetados no espaço do mapa de características de onde possam ser visualizados.

AE com camadas convolucionais mantêm as informações espaciais dos dados de entrada da forma como são e extraem de forma branda os recursos das características mais relevantes nas chamadas camadas intermediárias, na etapa de convolução são criados vários pequenos mapas de recursos (BUZUTI; THOMAZ, 2019). Algumas das principais diferenças entre os AEs convencionais e os DCAEs estão na redução significativa de parâmetros, na melhora da eficiência usando campos receptivos locais e o compartilhamento dos pesos (TAN; LI, 2014) e (YUE et al., 2015).

Uma das principais características do DCAE é o compartilhamento dos pesos, logo sua representação matemática será semelhante a representação do AE convecional mas levando em consideração esse detalhe. Quando uma entrada x possuir apenas um canal a representação do espaço latente do n-ésimo mapa de características será dada pela equação 2.3.

$$z^n = f(x^n) \tag{2.3}$$

Tendo x como dados originais de entrada, z a representação dos dados no espaço latente. Logo temos o processamento de *decoder* dado pela equação 2.4.

$$\hat{x}^n = g(z^n) \tag{2.4}$$

Onde temos z^n que representa os dados originais comprimidos armazenados no espaço latente e $\hat{x}^n$ são os dados originais reconstruídos.

Segundo Moraes e Castro (2018), no CAE a operação de convolução existente entre as saídas de uma camada anterior e um conjunto de filtros da camada de convolução subsequente é dada de forma discreta. AEs de aprendizado profundo utilizam camadas totalmente conectadas em sua estrutura, em contra partida modelos de DCAE utilizam camadas convolucionais e deconvolucionais.

3 METODOLOGIA DE PESQUISA

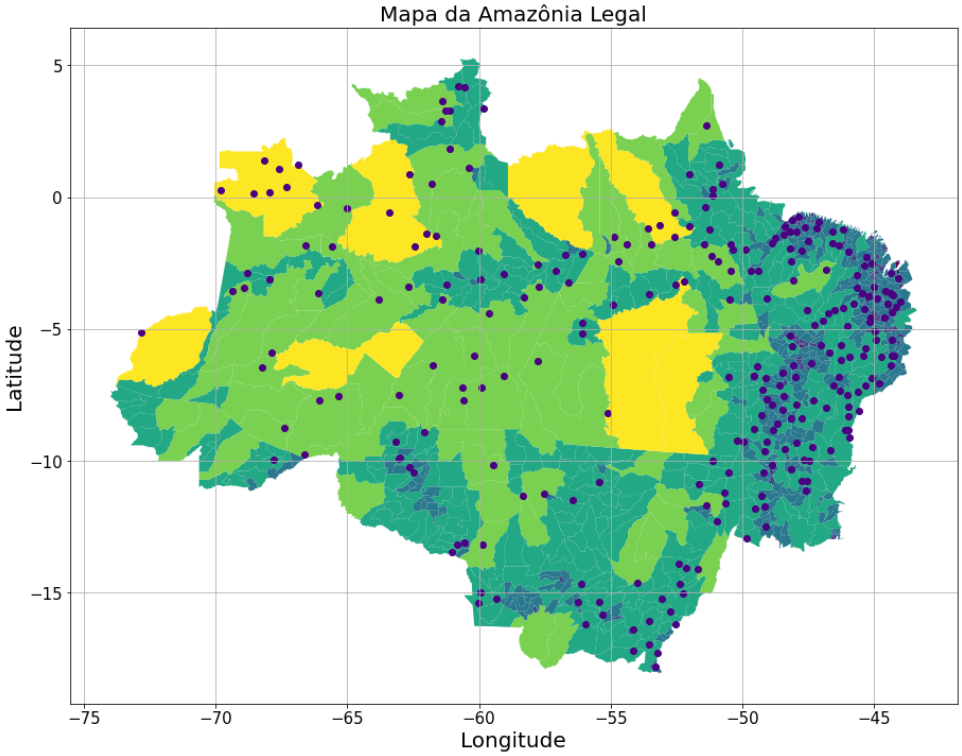
Neste capítulo, apresenta-se de forma detalhada os métodos que foram utilizados para realização deste trabalho. Os tópicos relacionados a coleta de dados, pré-processamento de dados, ambiente de desenvolvimento, configurações e arquitetura do modelo proposto, dinânmica dos fluxos de processos e o que mais foi necessário para melhor entendimento da metodologia abordada neste trabalho.

3.1 Base de Dados para Treinamento do Modelo

3.1.1 Fase de Coleta dos Dados

Os dados utilizados neste trabalho são valores correspondentes ao volume de precipitação de 268 estações pluviométricas localizadas na Amazônia Legal observadas em um período de 30 anos (1986 – 2015). De acordo com Dourado, Oliveira e Avila (2013), é uma padronização da Organização Meteorológica Mundial utilizar em estudos e projetos um universo amostral igual a 30 anos de observações visando representar o clima de uma determinada região. Os dados foram coletados em planilhas eletrônicas que podem ser encontradas no site (https://www.gov.br/ana/pt-br) da Agência Nacional de Águas e Saneamento Básico (ANA). Na figura 7 representa-se a área de estudo desta pesquisa.

Figura 7 – Amazônia Legal: área de estudo.



Fonte: Próprio Autor

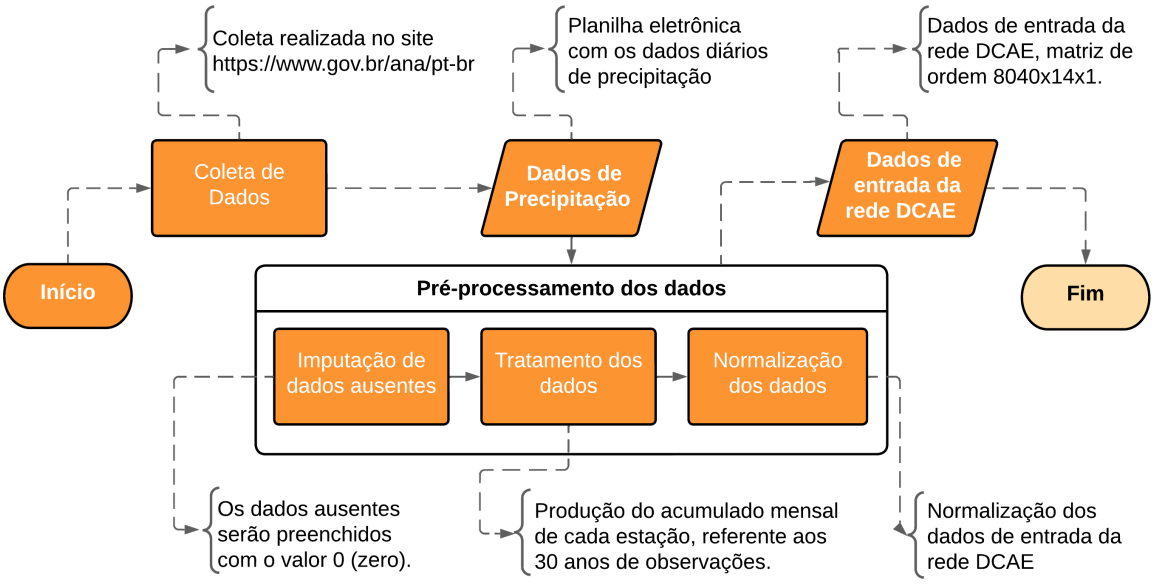
A Amazônia Legal dispõe atualmente em seus limites geográficos o quantitativo de 2.029 estações pluviométricas. A maioria das estações apresenta, dentro do período de estudo, dados com integridade comprometida, devido a não confiabilidade e inconsistência das informações ao longo das séries temporais disponíveis, esse comprometimento deriva do excesso de dados ausentes e de medições equivocadas (LIRA et al., 2019). Após análise e realização do tratamento dos dados ausentes, chegou-se ao quantitativo de 268 estações pluviométricas instaladas na região da Amazônia Legal, com séries temporais de dados diários, contínuos e completos, dentro do período de estudo, compreendido entre janeiro de 1986 a dezembro de 2015, totalizando os 30 anos de observações.

3.1.2 Fase de Pré-processamento dos Dados

Para Guarienti et al. (2015), observações pluviométricas utilizando instrumentos comuns como pluviômetro ou pluviógrafo são mais suscetíveis a falhas, que podem ocorrer por diversos motivos, seja por defeito no equipamento ou por negligência na observação, outro fator destacado é a representação local dos dados. Na etapa inicial de tratamento dos dados, devido às falhas apresentadas nas séries temporais, preenche-se os campos ausentes na tabela de precipitação diária com o valor 0 (zero), no processo seguinte, realiza-se o pré-processamento dos dados para a rede DCAE.

Na figura 8 segue uma ilustração do fluxo desse processo de tratamento dos dados brutos para transformação em dados de entrada das redes neurais.

Figura 8 – Fluxo de processamento para tratamento dos dados de entrada.



Fonte: Próprio Autor

No tratamento dos dados para obter-se as informações de entrada para o treinamento

da rede DCAE calcula-se os valores do acumulado mensal de cada ano observado para cada estação analisada e suas coordenadas geográficas (longitude e latitude), portanto, cada estação pluviométrica será representada por 30 observações com 14 atributos, uma amostra para cada ano analisado. Após finalizada essa etapa obtém-se uma matriz de 8040 x 14, sendo 8040 observações (30 anos para cada estação) com 14 atributos (acumulado mensal e coordenadas de localização geográficas).

3.2 Arquitetura e Treinamento do Modelo Proposto

O método de *deep convolutional autoencoder* é uma técnica de ML que utiliza redes neurais artificiais do tipo AE com camadas convolucionais para aprender e extrair uma representação de características de baixa dimensionalidade dos dados de entrada, e posterior reproduzi-los de forma aproximada na saída da rede. Essa técnica é amplamente utilizada em áreas como visão computacional, processamento de imagens, reconhecimento de fala, etc.

A arquitetura do DCAE é composta por duas partes principais: o codificador e o decodificador. O codificador é responsável por mapear os dados de entrada em uma representação latente de baixa dimensionalidade, enquanto o decodificador é responsável por reconstruir os dados de entrada a partir da representação latente.

Na etapa de codificação, várias camadas convolucionais e de *pooling* são utilizadas para extrair características relevantes dos dados de entrada e reduzir sua dimensionalidade. Essas camadas convolucionais são responsáveis por detectar padrões nos dados, enquanto as camadas de *pooling* são responsáveis por compactar os dados, preservando as características mais importantes.

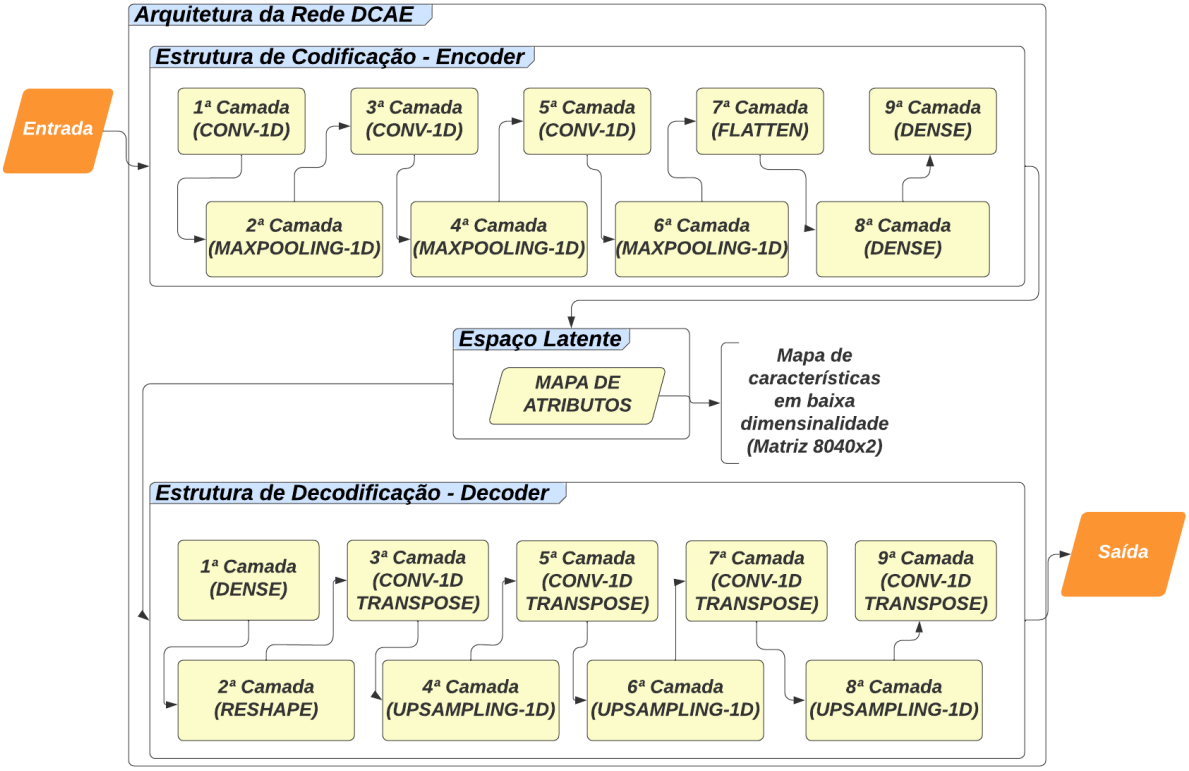
Na etapa de decodificação, as camadas convolucionais e de *pooling* são utilizadas de forma contrária para reconstruir os dados de entrada a partir da representação latente. As camadas de *pooling* são substituídas por camadas de *up-sampling*, que descompactam os dados, enquanto as camadas convolucionais são responsáveis por reconstruir os detalhes dos dados originais.

De acordo com Baia e Castro (2018), no processo de codificação esta etapa pode conter várias camadas de convolução e *pooling*, essas camadas também podem fazer parte do processo de decodificação mas realizando uma operação contrária, visando a reconstrução dos dados de entrada. A utilização de várias camadas convolucionais e de *pooling* no processo de codificação e decodificação permite que o DCAE aprenda uma representação mais robusta e compacta dos dados de entrada, tornando-o mais eficiente e preciso na reconstrução dos dados originais.

Para a estrutura do modelo DCAE em estudo, implementa-se no processo de codificação uma camada de entrada, três camadas de convolução, três camadas de *pooling*, uma camada *flatten* e duas camadas *dense*. Do mesmo modo, no processo de decodificação dos dados implementa-se na arquitetura do modelo uma camada *dense*, uma camada *reshape*, quatro camadas de

convolução transposta e três camadas de *pooling*. Essa estrutura é representada na figura 9.

Figura 9 – Fluxo de processamento do *deep convolutional autoencoder* proposto.



Fonte: Próprio Autor

- Camada *Conv1D*

  A Camada *Conv1D* é uma camada usada em modelos de redes neurais para processar dados unidimensionais, como séries temporais, permitindo que a rede aprenda a representação dos dados de entrada de forma automática. A rede pode aprender a extrair características importantes dos dados de forma automática, sem depender de recursos humanos para tal tarefa. Isso torna os modelos mais flexíveis e adaptáveis a diferentes conjuntos de dados. Ela aplica um conjunto de filtros a janelas deslizantes do sinal de entrada, produzindo uma saída que é uma versão transformada do sinal original. Os filtros aplicados pela Camada *Conv1D* são aprendidos durante o treinamento da rede neural. Eles são responsáveis por extrair características importantes dos dados de entrada, como padrões temporais ou mudanças na amplitude do sinal. Cada filtro é aplicado a uma janela deslizante do sinal de entrada, produzindo uma saída que representa a presença ou ausência da característica correspondente naquela posição no sinal. .

- Camada *MaxPooling*

  A Camada *MaxPooling* é uma camada usada em modelos de redes neurais para reduzir a dimensionalidade dos dados, tornando o processamento mais eficiente. Ela funciona agrupando regiões adjacentes do sinal de entrada e mantendo apenas o valor máximo de

cada grupo, produzindo assim uma versão reduzida do sinal original. Por exemplo, se a entrada for uma matriz de 2x2 com os valores [1, 2, 3, 4], a Camada *MaxPooling* pode agrupar os valores [1, 2] e [3, 4] em duas regiões separadas e manter apenas os valores máximos de cada região, resultando em uma saída de [2, 4]. O processo de *MaxPooling* é usado para reduzir a resolução espacial dos dados e, ao mesmo tempo, manter as características mais importantes do sinal de entrada. Ele ajuda a evitar o overfitting (sobreajuste) do modelo, que ocorre quando o modelo é muito complexo e ajustado demais aos dados de treinamento, perdendo a capacidade de generalizar para novos dados.

- Camada *Flatten*

  A Camada *Flatten* é uma camada usada para transformar uma matriz multidimensional em um vetor unidimensional. Esta camada consegue "achatar"a matriz de entrada, transformando-a em um vetor único. Por exemplo, se a entrada for uma matriz de dimensões 3x3x2 (três dimensões com 3 colunas, 3 linhas e 2 canais), a Camada *Flatten* pode transformá-la em um vetor unidimensional de 18 elementos, simplesmente empilhando as colunas uma após a outra. A Camada *Flatten* é geralmente usada em modelos de redes neurais para conectar a saída de uma camada de processamento espacial, como a Camada *Conv2D*, a uma camada de processamento completamente conectada, como a Camada *Dense*. Isso permite que a rede neural processe tanto informações espaciais quanto informações em série, permitindo uma representação mais rica e mais precisa dos dados.

- Camada *Dense*

  A Camada *Dense* é uma camada que dentro de sua estrutura cria uma camada de processamento completamente conectada. Ela recebe uma entrada de um vetor unidimensional e aplica uma transformação linear a ela, produzindo uma saída de outro vetor unidimensional. Essa camada implementa a operação: saída = ativação (ponto (entrada, *kernel*) + polarização) onde ativação é a função de ativação elemento a elemento passada como argumento. O *kernel* e o *bias* são criados pela camada, o primeiro corresponde a uma matriz de pesos e o segundo a um vetor de polarização que só é aplicado quando o parâmetro "use-bias"for verdadeiro. A Camada *Dense* é usada para aprender uma representação mais compacta e significativa dos dados de entrada, permitindo que a rede neural faça previsões mais precisas. É uma camada importante em muitos modelos de redes neurais, como redes neurais profundas e redes neurais convolucionais.

- Camada *Reshape*

  A Camada *Reshape* é uma camada usada para alterar a forma de um tensor de entrada para uma nova forma especificada pelo usuário. Essa camada permite remodelar os dados de entrada em um formato diferente, sem alterar o número total de elementos. Por exemplo, se a entrada for um tensor de forma (8, 10), que contém 80 elementos, a camada *Reshape* pode transformá-lo em um tensor de forma (40, 2), também com 80 elementos. Essa camada é muito útil em modelos de redes neurais para ajustar as dimensões dos dados de

entrada à estrutura da rede neural, principalmente quando a entrada tem uma forma que não é adequada para o processamento desejado.

- Camada *Conv1DTranspose*
  A Camada *Conv1DTranspose* é uma camada usada para realizar a operação inversa da camada *Conv1D*. Enquanto a camada *Conv1D* aplica filtros para extrair características de uma sequência unidimensional de entrada, a camada *Conv1DTranspose* realiza a operação inversa, ou seja, usa filtros para sintetizar uma sequência a partir de um espaço latente ou de características de entrada. Essa camada também pode ser usada para reconstruir uma sequência unidimensional de entrada, como no caso de AE. A operação da camada *Conv1DTranspose* envolve o preenchimento dos valores de entrada com zeros, seguido pela convolução com filtros de transposição. Esses filtros de transposição são semelhantes aos filtros usados na camada *Conv1D*, mas são "invertidos"ou "transpostos"em relação a eles. A convolução com esses filtros de transposição produz uma saída que tem uma forma mais ampla do que a entrada, permitindo que a camada *Conv1DTranspose* sintetize uma sequência a partir de um espaço latente ou de características de entrada.

- Camada *UpSampling1D*
  A camada *UpSampling1D* é geralmente usada em modelos de redes neurais para ajudar na tarefa de reconstrução de dados de baixa resolução. Ela funciona aumentando o número de amostras de um sinal unidimensional, o que resulta em uma representação com mais detalhes e uma resolução mais alta. Ao aumentar a resolução espacial dos dados, a camada ajuda o modelo a aprender padrões mais finos e detalhados, o que pode levar a resultados mais precisos. O processo de aumento da resolução é realizado através da repetição das amostras do sinal original, de forma a preencher os espaços vazios entre elas. Isso é feito sem adicionar novas informações ao sinal original, mas apenas duplicando as informações já existentes.

Todo o processo de desenvolvimento e treinamento do modelo foi realizado em ambiente Python 3.8 (PYTHON, 2021), utilizando o módulo Keras de rede neural (KERAS, 2021) tendo como backend a biblioteca TensorFlow (TENSORFLOW, 2021) e o auxílio de uma GPU Geforce 8400 GS com suporte à tecnologia CUDA (NVIDEA, 2021). Outras características da plataforma de desenvolvimento e treinamento são processador i5-4590 3.30 GHz e 16GB de memória RAM.

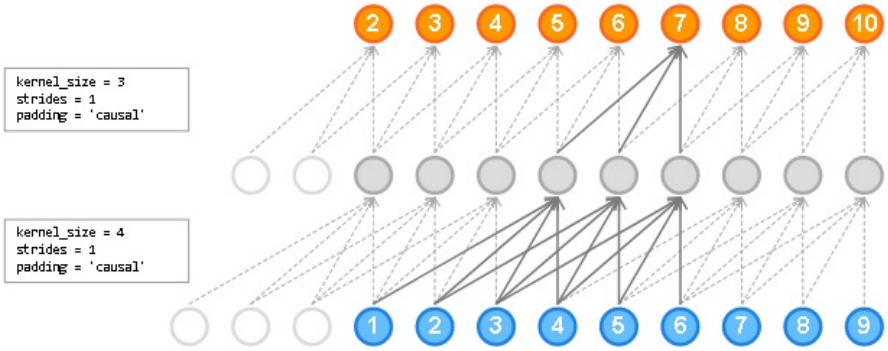
3.2.1 Configurações dos hiperparâmetros do modelo

A biblioteca Keras (KERAS, 2021) utilizada neste modelo possui alguns módulos de camadas que são necessários para a configuração da estrutura da rede, bem como seu pleno funcionamento. Esses módulos adotam o uso de hiperparâmetros que a cada iteração buscam ajustar o modelo para que a perda seja minimizada. Abaixo lista-se alguns dos hiperparâmetros dessas camadas.

- *Filters*
  Os filtros são pequenas peças de mapas de recursos que se originam na etapa de convolução, eles preservam a relação entre os dados de entrada, obtendo as principais características desses dados. Essas peças de mapas de recursos são como janelas que deslizam através dos dados para extrair as características.

- *Kernel size*
  Esse parâmetro específica o comprimento da janela de convolução, ou seja, a quantidades de dados que a janela irá comportar.

Figura 10 – Representação do funcionamento do *kernel size*.

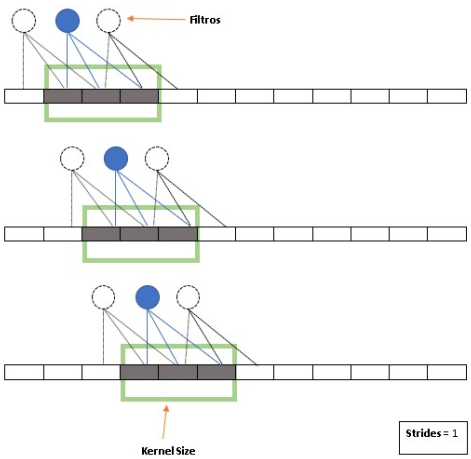


Fonte: Batzner (2019)

- *Activation*
  Esse parâmetro deverá ser preenchido com a função de ativação que deverá ser usada no modelo, visando garantir a transformação não linear ao longo do sinal de entrada.

- *Strides*
  É o número de passos que a janela de convolução se desloca na matriz de entrada. Quando esse valor é 1, a janela de convolução mudam 1 sinal por vez.

Figura 11 – Representação do funcionamento dos *strides*.



Fonte: Próprio Autor

- *Padding*
  Outro hiperparâmetro é o *padding* que oferece três opções: *valid, same e causal*. A opção *valid* quando escolhida não realiza nenhuma tarefa, ou seja, sem preenchimento. Quando utilizado o *same* resulta no preenchimento uniforme sentido esqueda, sentido direita ou para cima, para baixo da entrada, de forma que a saída tenha a dimensionalidade igual a da entrada. Utilizando *causal* teremos convoluções dilatadas, por exemplo a saída[n] , não depende da entrada[n+1:]. Muito útil quando trabalhamos com séries temporais em que o modelo deve manter a ordem dos dados.

Com relação a obtenção de valores ideais dos hiperparâmetros, buscou-se de forma empírica uma configuração que apresentasse resultados razoáveis à métrica de avaliação do modelo. Iniciou-se o processo com valores baixos, neste caso o valor dois, e iguais em todos os filtros e *kernels* das camadas convolucionais e convolucionais transpostas.

Após a primeira execução alterou-se apenas os valores dos filtros aumentando-os de forma gradual tendo como referência valores múltiplos da quantidade de meses em estudo. Após atribuir o valor 36, observou-se que a redução do erro da métrica de avaliação do modelo estabiliza, e comparado ao custo computacional gerado com o aumento dos valores deste parâmetro adotou-se este valor como referência.

Após encontrado um valor de referência para os filtros, buscou-se ajustar os valores de *kernel* das camadas convolucionais. A mesma metodologia adotada para determinar os valores dos filtros foi empregada para obtenção dos valores dos *kernels*. Desta forma, o valor de referência encontrado foi 18, onde da mesma forma, notou-se uma estabilidade na redução do erro de avaliação do modelo.

Com os valores de referência para os filtros e *kernels* definidos, realizou-se novos experimentos com os valores dos filtros e *kernels* variando entre as camadas do modelo, sendo os valores encontrados como os valores máximos. Desta forma, encontrou-se a melhor configuração para modelo proposto, de acordo com a métrica de avaliação.

Arquiteturas que utilizam camadas convolucionais recebem em sua camada de entrada dados em três dimensões, portanto, acrescentou-se o número de canais a matriz de dados em estudo, estabelecendo a matriz de ordem 8040x14x1.

Para treinamento do modelo utilizou-se um processo com 2000 iterações com lotes de tamanho igual a 30, fazendo com que todas as observações de cada estação fossem aplicadas em um único lote. Apresentou-se à camada de entrada do DCAE a matriz NONEx14x1, onde "NONE" representa o tamanho do lote de dados, portanto os dados de entrada do modelo são matrizes de ordem 30x14x1.

A arquitetura convolucional é amplamente utilizada em tarefas de processamento de imagens e reconhecimento de padrões, onde de acordo com Barino e Santos (2020), os dados de entrada são tipicamente matrizes com várias dimensões (como altura, largura e canais de cor).

No entanto, na análise de séries temporais, os dados de entrada são tipicamente vetores unidimensionais, onde cada elemento do vetor corresponde a uma medida em um determinado momento no tempo. Nesse caso, as camadas convolucionais precisam ser modificadas para trabalhar com dados de entrada unidimensionais.

As camadas convolucionais em arquiteturas para análise de séries temporais realizam convoluções unidimensionais. Nessa operação, um filtro é deslizado ao longo do vetor de entrada para extrair características relevantes em cada posição. A operação de *pooling* também pode ser realizada de forma unidimensional, reduzindo a dimensão do vetor sem comprometer a informação relevante.

Essas adaptações nas camadas convolucionais permitem que as arquiteturas convolucionais sejam utilizadas para análise de séries temporais, com aplicações em áreas como previsão de séries temporais, detecção de anomalias em séries temporais, processamento de sinais, entre outras.

Como otimizador do sistema de treinamento aplicou-se o processo que realiza a implementação do algoritmo *adaptive moment estimation* (ADAM), que é um otimizador de gradiente estocástico que combina a ideia de atualização de momento do *stochastic gradient descent* (SGD) com uma estimativa adaptativa de segundo momento das variações dos gradientes. Ele calcula uma taxa de aprendizagem adaptativa individual para cada parâmetro do modelo, o que ajuda a lidar com problemas de escala em diferentes direções.

ADAM é um método popular para o treinamento de redes neurais profundas devido à sua eficiência computacional e robustez em lidar com diferentes tipos de funções de custo e hiperparâmetros. Em comparação com outros otimizadores de gradiente estocástico, como SGD, RMSprop e Adagrad, o ADAM tende a convergir mais rapidamente para o mínimo global e é menos sensível à escolha de hiperparâmetros.

O algoritmo ADAM é computacionalmente eficiente porque mantém apenas uma pequena quantidade de memória para cada parâmetro do modelo. Além disso, o algoritmo utiliza uma média móvel do momento de primeira ordem e do momento de segunda ordem dos gradientes, o que ajuda a reduzir a variância dos gradientes e a lidar com problemas de escala em diferentes direções.

Em geral, o algoritmo ADAM é um método poderoso para a otimização de redes neurais profundas em problemas com grandes volumes de dados e configurações complexas de parâmetros. No entanto, é importante ajustar cuidadosamente os hiperparâmetros do algoritmo, como a taxa de aprendizagem, o coeficiente de *momentum* e os parâmetros de regularização, para obter os melhores resultados. Segundo Kingma e Ba (2014), trata-se de um método com custo computacional baixo, uma vez que necessita de pouco consumo de memória, ideal para solução de problemas com grandes volumes de dados e configurações complexas de parâmetros.

Na tabela 1 apresenta-se os valores adotados para os hiperparâmetros das camadas de

codificação e decodificação do modelo proposto.

Tabela 1 – Valores adotados para os hiperparâmetros do modelo

Descrição	***Filtros***	***Kernel Size***	***Activation***	***Strides***	***Padding***
Processo de Codificação					
Camada de Entrada	(8040, 14, 1)				
1ª Camada *Conv1D*	36	18	Selu	1	Causal
1ª Camada *Maxpooling*	6			1	
2ª Camada *Conv1D*	24	12	Selu	1	Causal
2ª Camada *Maxpooling*	5			1	
3ª Camada *Conv1D*	12	6	Selu	1	Causal
3ª Camada *Maxpooling*	4			1	
Camada *Flatten*					
1ª Camada *Dense*	6		selu		
2 ª Camada *Dense*	2				
Espaço latente	(8040, 2)				
Processo de Decodificação					
Camada de entrada (espaço latente)	(8040,2)				
1ª Camada *Dense*	168		Selu		
1ª Camada *Reshape*	(14,1)				
1ª Camada *Conv1DTranspose*	12	6	Selu	1	same
1ª Camada *Upsampling*	1				
2ª Camada *Conv1DTranspose*	24	12	Selu	1	same
2ª Camada *Upsampling*	1				
3ª Camada *Conv1DTranspose*	36	18	Selu	1	same
3ª Camada *Upsampling*	1				
4ª Camada *Conv1DTranspose*	1	1	Selu		same
Camada de saída	(8040, 14, 1)				

Fonte: Próprio Autor

3.2.2 Avaliação de desempenho do modelo proposto.

A avaliação do desempenho de um modelo é uma etapa crítica em qualquer processo de modelagem, pois permite avaliar a capacidade do modelo em realizar tarefas específicas. A abordagem comum para avaliação de modelos em muitas áreas é o cálculo do erro médio quadrático normalizado (NRMSE), do inglês *normalized root mean square error*, que é uma medida da diferença entre os valores previstos pelo modelo e os valores reais dos dados de teste.

Antes de avaliar o modelo com a métrica NRMSE, é necessário configurar o modelo para avaliar a perda na reconstrução dos dados de entrada, que é comumente chamada de *loss*. A *loss* é uma medida da diferença entre os valores previstos pelo modelo e os valores reais dos

dados de treinamento. O objetivo do processo de treinamento é minimizar essa *loss*, de modo que o modelo seja capaz de generalizar bem para novos dados.

Uma vez que o modelo foi treinado e a *loss* foi minimizada, é possível avaliar seu desempenho usando a métrica NRMSE. O NRMSE é calculado dividindo o *root mean square error* (RMSE) pela amplitude dos dados de teste. Ele é normalizado para permitir a comparação do desempenho de modelos em diferentes conjuntos de dados.

Desta forma, consegue-se avaliar a melhor configuração para o modelo proposto neste estudo. O *loss* é calculado utilizando o *mean square error* (MSE), de posse do MSE calculamos o RMSE que é dado pela equação 3.3.

$$e_i = \hat{x}_i - \bar{x}_i \tag{3.1}$$

$$\|e_i\| = \|\hat{x}_i - \bar{x}_i\|^2 \tag{3.2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|e_i\|^2} \tag{3.3}$$

Logo o NRMSE é dado pela equação 3.4.

$$NRMSE = \frac{RMSE}{\sqrt{var(x)}} \tag{3.4}$$

O cálculo do RMSE é baseado na média das diferenças entre os dados reconstruído e as observações originais ao quadrado. Com finalidade de avaliar o erro entre os valores preditos e as entradas originais, tendo seu resultado variando de a partir de zero. Já o NRMSE considera a escala dos valores observados, facilitando a comparação entre modelos que tenham escalas diferentes, desta maneira pode-se chegar a informação da proximidade entre o resultado e o dados original. A variação de seus valores parte de um.

4 RESULTADOS E DISCUSSÃO

No presente estudo, analisou-se dados de precipitação referente a 30 anos de 268 estações pluviométricas localizadas na Amazônia Legal. Utilizando técnicas de machine learning com aprendizado não supervisionado, realizou-se a descoberta de conhecimento e identificação de padrões dos dados estudados. Foram adotadas técnicas de redes neurais profundas do tipo *deep convolutional autoencoder* para extração de conhecimento e representação em baixa dimensionalidade dos dados, e técnicas de *clustering*, como o aglomerativo hierárquico com método de ligação de Ward, para a tarefa de agrupamento e melhor análise dos dados das estações.

4.1 Resultados para análise dos dados pelo modelo DCAE e aplicação da técnica de *clustering*.

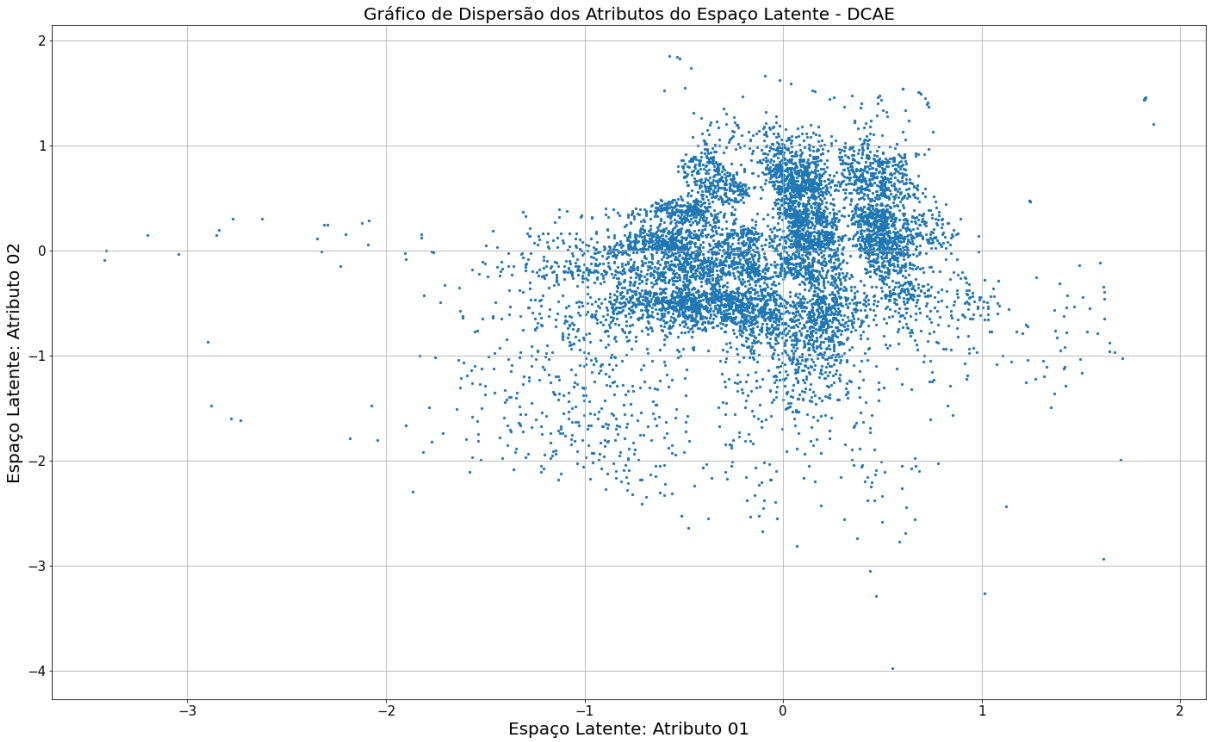
Utilizando técnicas de ML, realizou-se a descoberta de conhecimento e identificação de padrões dos dados estudados. Foram utilizadas técnicas de redes neurais profundas do tipo DCAE, que é uma técnica de rede neural com capacidade de aprender uma representação de baixa dimensionalidade dos dados de entrada, permitindo a extração de características relevantes e reduzindo a complexidade dos dados originais. Com a representação em baixa dimensionalidade dos dados, foi possível aplicar técnicas de *clustering* para identificar padrões e agrupar os dados em diferentes categorias ou grupos com características semelhantes.

De acordo com Tan, Steinbach e Kumar (2016), a utilização do *clustering* é comum para análise exploratória de dados, permitindo a identificação de grupos ou *clusters* de dados que apresentam comportamentos semelhantes. Essa técnica, pode ser útil para identificar padrões e relacionamentos entre os dados de precipitação das estações, fornecendo informações importantes para a tomada de decisão em várias áreas, como a agricultura, gestão de recursos hídricos, previsão de inundações, entre outras.

Após a etapa de codificação, realizada pelo modelo DCAE, os dados de precipitação das 268 estações selecionadas na Amazônia Legal são representados em baixa dimensionalidade no espaço latente, gerando os dados com estrutura bidimensional. Essa representação permite a visualização das 8040 observações das estações pluviométricas em um gráfico de dispersão, onde cada ponto representa o ciclo anual de precipitação, dos 30 anos estudados, para cada estação selecionada. Sua posição no gráfico é determinada pelos valores dos atributos das duas dimensões do espaço latente. A figura 12 apresenta o gráfico de dispersão, que é uma ferramenta importante para a análise exploratória dos dados e a identificação de padrões ou agrupamentos de estações pluviométricas com características semelhantes.

Após a análise dos atributos presentes no espaço latente, notou-se que a maioria das

Figura 12 – Gráfico de dispersão com os dados do espaço latente.



Fonte: Próprio Autor

observações estão concentradas, com relação aos eixos x e y, entre os valores -1.0 e 1.0. Esse resultado é evidenciado pelo gráfico de dispersão produzido, no qual foi possível observar a aglutinação considerável da parte majoritária das observações em baixa dimensionalidade. Essas observações representam o acumulado mensal de precipitação das estações pluviométricas selecionadas dentro do período de 30 anos em estudo.

Cada ponto no plano representa um ano de observação de cada estação, e os pontos mais distantes, podem ser considerados representação de anos com observações distintas dos padrões. Essas observações podem ter ocorrido por falhas na obtenção dos dados ou por características atípicas do fenômeno natural estudado. Portanto, é importante analisar esses outliers com cuidado e compreender suas causas.

Além da visualização dos dados, do espaço latente gerado pelo modelo, também utilizou-se como parâmetro de análise da rede a métrica do MSE para avaliar a perda de informações na reconstrução dos dados originais. Essa foi uma forma de avaliação do comportamento do modelo, levando-se em consideração que para reconstrução aproximada dos dados originais o processo de decodificação necessita de uma entrada, que são os dados do espaço latente. Essa entrada deve dispor de uma boa representação das características aprendidas dos dados originais. Neste sentido, pôde-se considerar que a métrica do MSE, utilizada na reconstrução aproximada dos dados originais, também seria uma opção de avaliação do desempenho do modelo proposto.

A fim de melhorar a métrica de erro adotada no modelo, tomou-se o MSE como raiz do

erro médio quadrático, RMSE, e em seguida como raiz do erro médio quadrático normalizada, NRMSE. Na tabela 2 estão dispostos os valores do RMSE e NRMSE gerados pelo DCAE proposto.

Tabela 2 – RMSE e NRMSE gerados pelo modelo

RMSE	NRMSE
0.06610	0.3355

Fonte: Próprio Autor

Após obtenção da representação em baixa dimensionalidade das séries temporais de precipitação das estações estudadas, utilizou-se técnicas de *clustering* com os dados gerados para análise e identificação de padrões e mapeamento de regiões homogêneas na Amazônia Legal com relação aos valores de precipitação mensal.

As técnicas de *clustering* têm uma ampla aplicação em diversas áreas do conhecimento, desempenhando um papel fundamental na compreensão dos dados por meio da formação de grupos com elementos apresentando características semelhantes. Essa abordagem é particularmente útil para a análise de diferentes situações, incluindo o tratamento de dados climáticos e a segmentação de conjuntos de estações pluviométricas com características homogêneas.

De acordo com Neves et al. (2017), Menezes, Fernandes e Rocha (2015) e Amanajás e Braga (2012), no contexto da análise de dados climáticos, a técnica de *clustering* permite agrupar estações pluviométricas com padrões semelhantes de precipitação, identificando regiões geográficas ou climáticas com características pluviométricas consistentes. Ao agrupar as estações pluviométricas com base em características semelhantes, pode-se identificar padrões climáticos regionais e compreender melhor as variações espaciais e temporais da precipitação. Essas informações são essenciais para conhecer os regimes de chuva, fornecendo informações valiosas para tomadas de decisão relacionadas a agricultura, gestão de recursos hídricos e outras áreas, bem como para a elaboração de estratégias de adaptação às condições climáticas locais.

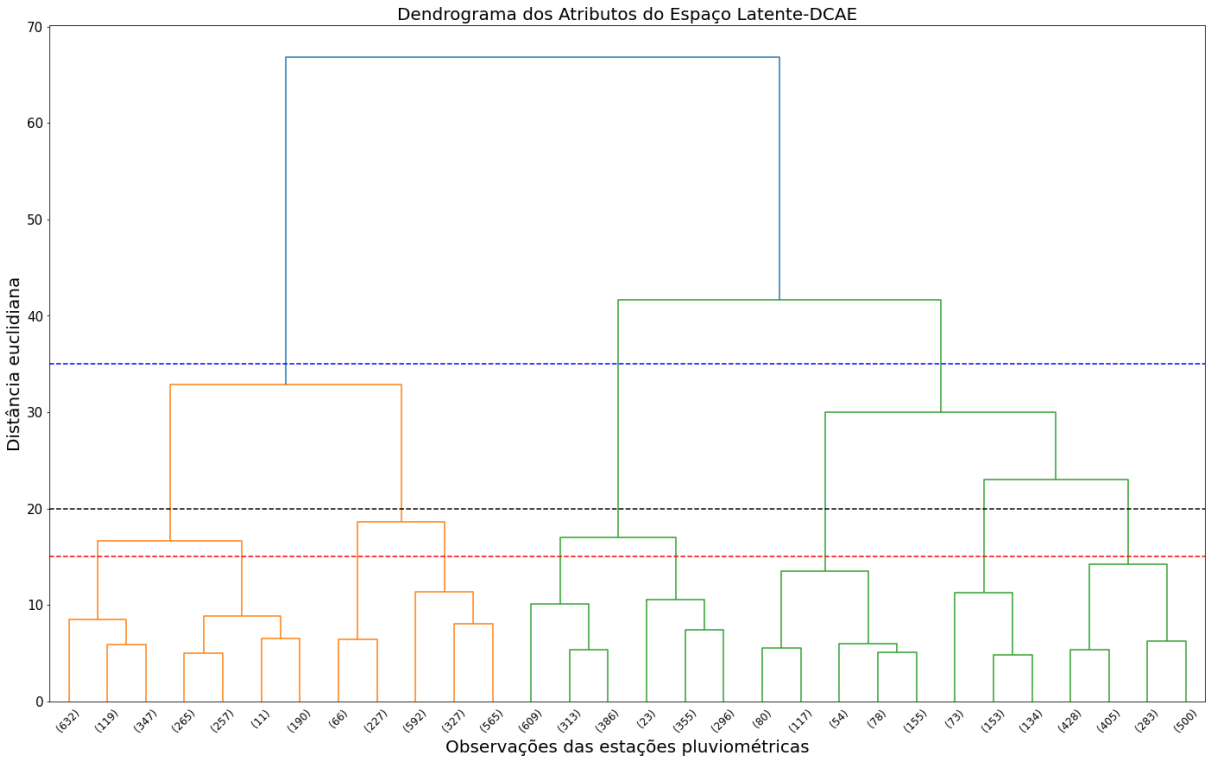
Existem vários métodos disponíveis na literatura para realização da análise de *clustering*, dos quais podem ser classificados como métodos hierárquicos e não hierárquicos. A principal característica que difere os métodos hierárquicos dos não hierárquicos, é a dispensa de conhecimento prévio de um número específico de *clusters*. Os métodos hierárquicos possuem duas subclasses que dividem seus algoritmos para realização de agrupamentos, que são os algoritmos de agrupamento aglomerativo (*agglomerative clustering*) e os agrupamentos por divisão (*divisive clustering*) (CRISPIM et al., 2020).

O *clustering* é uma técnica de aprendizado não supervisionado que tem como principal objetivo agrupar objetos com atributos semelhantes em grupos. Neste estudo, utilizou-se a técnica de aglomerativo hierárquico com o método de ligação de *Ward* para analisar e agrupar dados do espaço latente. É importante ressaltar que essa metodologia requer o conhecimento prévio do número de *clusters* que devem ser formados. Desta maneira, para obtenção da informação do

quantitativo de *clusters* necessários para análise dos atributos do espaço latente e realização do aprendizado não supervisionado, aplicou-se a técnica de dendrograma, que analisa as observações e, com base na métrica da distância euclidiana, separa os dados em *clusters*.

A aplicação do dendrograma, é uma técnica comumente utilizada em estudos que envolvem análise de dados e ML. Essa técnica, permite identificar o número adequado de *clusters* para a análise dos atributos do espaço latente e a realização do aprendizado não supervisionado. Além disso, a métrica da distância euclidiana é uma medida popular para calcular a similaridade entre dados, e assim, agrupá-los em *clusters*. A figura 13 apresenta o dendrograma formado.

Figura 13 – Dendrograma para formação de *clusters* utilizando a distância euclidiana como medida.



Fonte: Próprio Autor

Analisando o dendrograma gerado, observou-se dentre as formações possíveis, uma formação com três *clusters* diante dos dados de precipitação apresentados a técnica proposta para obtenção do número de *clusters*. Segundo Santos, Lucio e Silva (2015), três regiões homogêneas são suficientes para representar os diferentes padrões de precipitação na Amazônia. Pois as regiões são coerentes com a atuação dos principais sistemas atmosféricos responsáveis pela formação de chuva na Amazônia. Com relação a climatologia do estado do Pará, que pertence a Amazônia Legal, existem três subtipos climáticos: “Af”, “Am”, “Aw”, estes subtipos foram determinados com a classificação climática de Köppen e estão relacionados ao clima tropical chuvoso (CRISPIM et al., 2020).

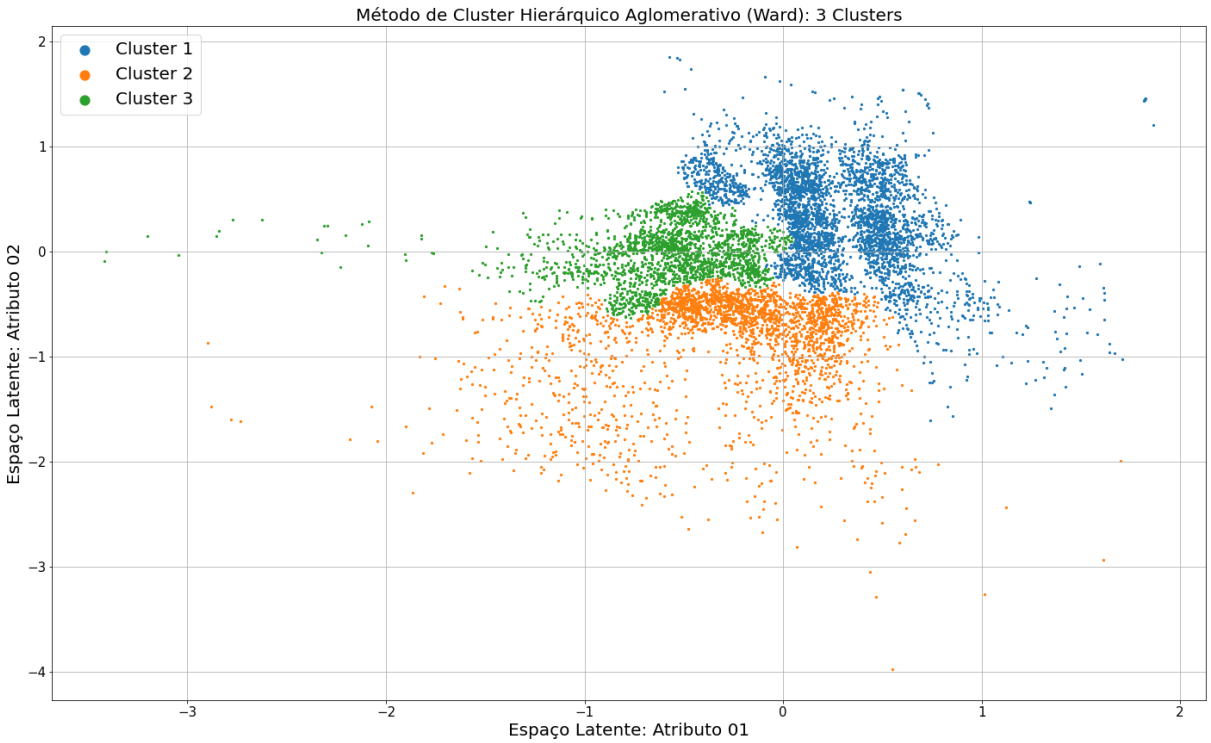
A linha pontilhada em azul corresponde a distância euclidiana adotada para separação em três grupos de observações. Para melhor análise, compreensão e comparação dos dados, outras

formações com diferentes números de *clusters* foram realizadas. Dessa maneira, as formações adotadas para o estudo foram as que apresentaram clusters com 3, 6 e 9 grupos distintos.

4.1.1 Resultado da aplicação da técnica de *clustering* para formação com 3 grupos.

Após a definição do número de *clusters* que foram empregados no estudo, aplicou-se as técnicas de *clustering* escolhidas no estudo para análise e comparação dos resultados. Iniciou-se, portanto, o processo de *clustering* com o método aglomerativo hierárquico com ligação de *Ward*. Na figura 14 o espaço latente é representado em gráfico de dispersão com a definição dos *clusters*, neste caso, com a formação de 3 grupos.

Figura 14 – Gráfico de dispersão com os dados do espaço latente: Formação com 3 *clusters*.



Fonte: Próprio Autor

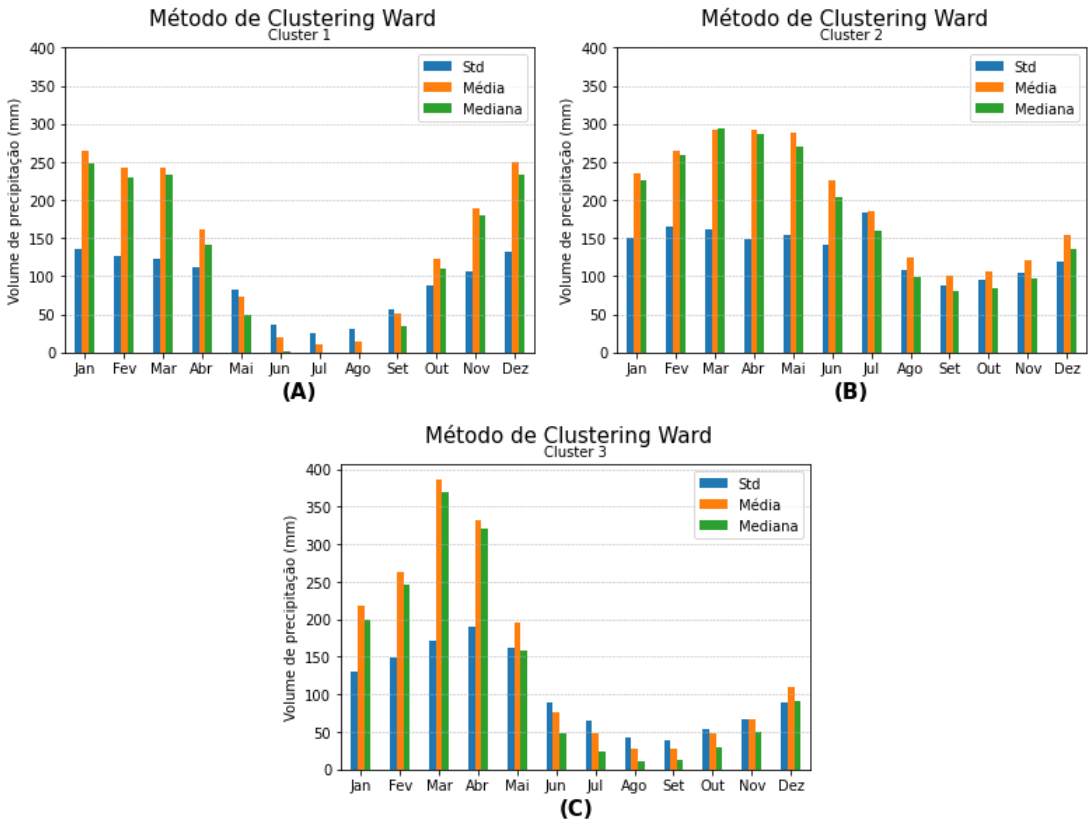
No estudo em questão, foram aplicadas técnicas de *clustering* para agrupar as observações relacionadas ao acumulado mensal de precipitação das estações pluviométricas selecionadas ao longo de 30 anos do estudo. Com base nessa análise, foi possível identificar a formação de três *clusters* distintos, sendo o cluster-1 o mais numeroso, com 3598 observações, seguido pelo cluster-2 com 2460 observações, e por fim, o cluster-3 com 1982 observações. No total, esses *clusters* representam um conjunto de 8040 observações, permitindo uma análise mais precisa dos dados de precipitação e suas variações ao longo dos anos. Essas informações são valiosas para a compreensão dos fenômenos naturais e contribuem para o desenvolvimento de modelos de previsão e análise de tendências climáticas.

O cluster-1, representado pelos pontos em azul, concentrou a maior parte de suas observações entre os valores 0 e 1 no eixo X do gráfico, e apresentou valores semelhantes para o eixo Y. Já o cluster-2, representado pelos pontos em laranja, apresentou a maior parte de suas observações concentradas entre -1 e 1 no eixo X, e entre -2 e 0 no eixo Y.

O cluster-3 é representado pelos pontos em verde dispersos no plano, e apresentou um nível de aglutinação maior do que os outros *clusters*. As observações deste *cluster* estão concentradas entre -1 e 0 no eixo X e entre -0.5 e 0.5 no eixo Y. A partir dessas informações, pôde-se identificar diferenças claras entre os três *clusters* e compreender melhor as características e padrões presentes nos dados.

Após a formação dos *clusters* e atribuição de cada observação a seu respectivo *cluster*, foi necessário realizar análises estatísticas dos dados para compreender melhor os padrões que determinaram a formação de cada grupo. Nesse sentido, foi fundamental calcular as medidas de desvio padrão, média e mediana das observações pertencentes a cada *cluster*. Dessa forma, foi possível obter uma visão geral da distribuição dos dados em cada grupo e realizar comparações entre os *clusters* representados nos gráficos das figuras 15, 16, 17 e 18. Essas análises estatísticas permitem identificar características distintas em cada *cluster* e podem ser úteis para a tomada de decisões em diversas áreas, como planejamento urbano, agronegócio e gestão de recursos hídricos, entre outras.

Figura 15 – Gráficos de barras com estatísticas (desvio padrão, média e mediana) dos *clusters*: Formação com 3 *clusters*.



Fonte: Próprio Autor

Essas informações, são importantes para o entendimento do comportamento dos *clusters* e auxiliando na identificação de padrões climáticos sazonais. Além disso, a análise do desvio padrão, média e mediana de cada *cluster* nos gráficos 15.A, 15.B e 15.C, também forneceu informações sobre as variações dos dados dentro de cada *cluster*. Esses indicadores estatísticos foram úteis para entender a consistência das observações dentro do mesmo *cluster* e a sua relação com os outros *clusters*. A comparação entre os diferentes *clusters* auxiliou na identificação de diferenças e similaridades nos padrões de precipitação, permitindo a realização de análises mais precisas e robustas.

O gráfico 15.A, que representa o cluster-1, mostrou que o volume de chuva é mais significativo nos meses de dezembro a março, enquanto que, nos meses de maio a setembro há uma diminuição no volume de chuva. Esse padrão é consistente com as características climáticas da região estudada, onde o período chuvoso começa em novembro, atinge o pico em janeiro e começa a diminuir em abril. A partir de maio, começa o período mais seco.

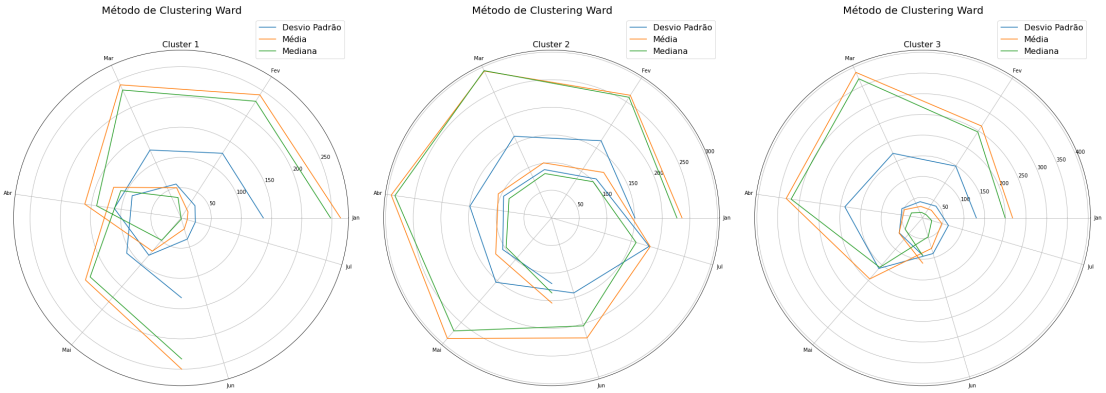
No gráfico 15.B que representa o cluster-2, observou-se que os meses de março, abril e maio apresentam os maiores volumes de chuva, enquanto que, agosto, setembro, outubro e novembro apresentam os menores volumes. Notou-se que o mês de dezembro pode ser considerado como um período de transição, com início das chuvas mais intensas que se estendem até julho, quando começa a transição para o período seco. Essa análise é importante para a compreensão dos padrões climáticos do cluster-2.

O estudo realizado por Lira et al. (2020b), analisou a série histórica da cidade de Belém e constatou que a distribuição das chuvas ao longo do ano apresenta variações significativas, formando um período muito chuvoso que compreende os meses de dezembro a maio e um período menos chuvoso que compreende os meses de agosto a novembro. Essa variação temporal pluviométrica é influenciada pelos principais sistemas atmosféricos que atuam na região. Esses resultados corroboram com os descobertos pelo presente estudo, que após análise dos dados da série temporal pelo sistema DCAE e aplicação do método de *clustering*, identificou que a estação pluviométrica que representa a cidade de Belém foi agrupada dentro do cluster-2. Esse *cluster*, apresentou as mesmas características da distribuição das chuvas tanto no período chuvoso quanto no período de seca, como identificado no Gráfico 15.B. Dessa forma, os resultados dos dois estudos se complementam e reforçam a importância de considerar a influência dos sistemas atmosféricos na variabilidade temporal das chuvas na região.

O cluster-3, representado pelo gráfico 15.C, apresentou características distintas dos demais *clusters* analisados. Nele, foi possível observar um período mais extenso com os volumes de chuvas mais baixos, que se estende de junho a dezembro, com destaque para os meses de agosto e setembro, que apresentam os menores valores de precipitação. Em contrapartida, o mês de março se destaca como o mais chuvoso. A análise estatística da série histórica demonstrou que o início do crescente volume de chuvas neste *cluster* ocorre em dezembro, atingindo o pico em março e começando a decrescer em maio. Em junho, inicia-se o período mais seco.

Os valores de desvio padrão, média e mediana também foram plotados em coordenadas polares como demonstrados nas figuras 16 A, 16 B e 16 C.

Figura 16 – Gráficos com estatísticas (desvio padrão, média e mediana) em coordenadas polares: Formação com 3 clusters.



Fonte: Próprio Autor

A análise dos *clusters* foi fundamental para compreender a variabilidade temporal e espacial das chuvas em uma determinada região. Entretanto, é importante destacar que uma mesma estação pluviométrica pôde apresentar observações em mais de um *cluster*, sendo explicado por diversos motivos, como por exemplo, um ano atípico ou falhas na captação do volume de precipitação.

Para solucionar essa questão, foi necessário criar um mecanismo de avaliação e ranking das observações de cada estação pluviométrica, a fim de definir a qual *cluster* uma determinada observação pertencia. Desse modo, uma estação **X** foi incluída no *cluster* **Y**, quando a maioria das observações dessa estação **X** pertencia ao *cluster* **Y**. Essa metodologia permitiu uma análise mais precisa da distribuição das chuvas em uma determinada região e a identificação das variações temporais e espaciais.

Após a realização do ranking as estações se distribuíram entre os *clusters* conforme a tabela 3.

Tabela 3 – Número de estações por *cluster* formado.

Identificação dos *clusters*	Número de estações por *cluster*
Cluster-1	132
Cluster-2	85
Cluster-3	51

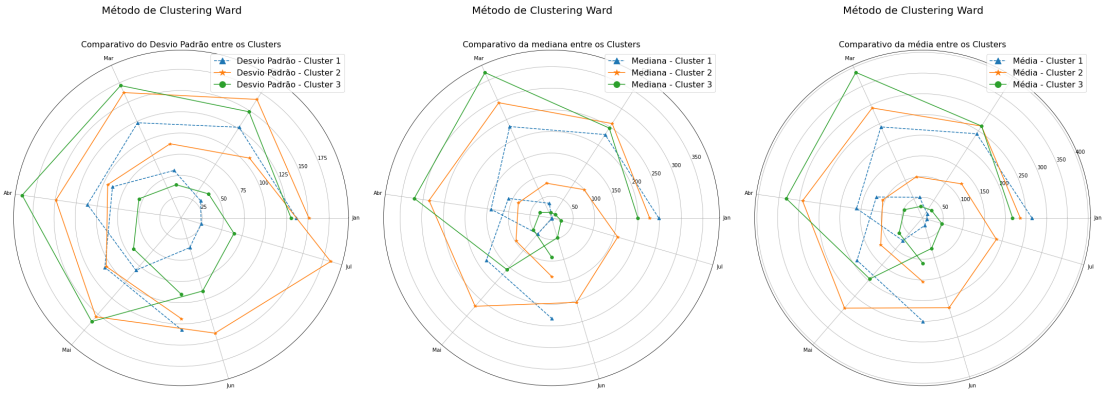
Fonte: Próprio Autor

Com a definição a qual *cluster* cada uma das estações pluviométricas pertencia, realizou-se uma análise estatística da mediana, da média e do desvio padrão dos *clusters* formados. As figuras 17 e 18, apresentam gráficos comparativos entre esses *clusters*, as medidas de desvio

padrão são mostradas nos gráficos 17.A e 18.A, a mediana nos gráficos 17.B e 18.B, e a média nos gráficos 17.C e 18.C.

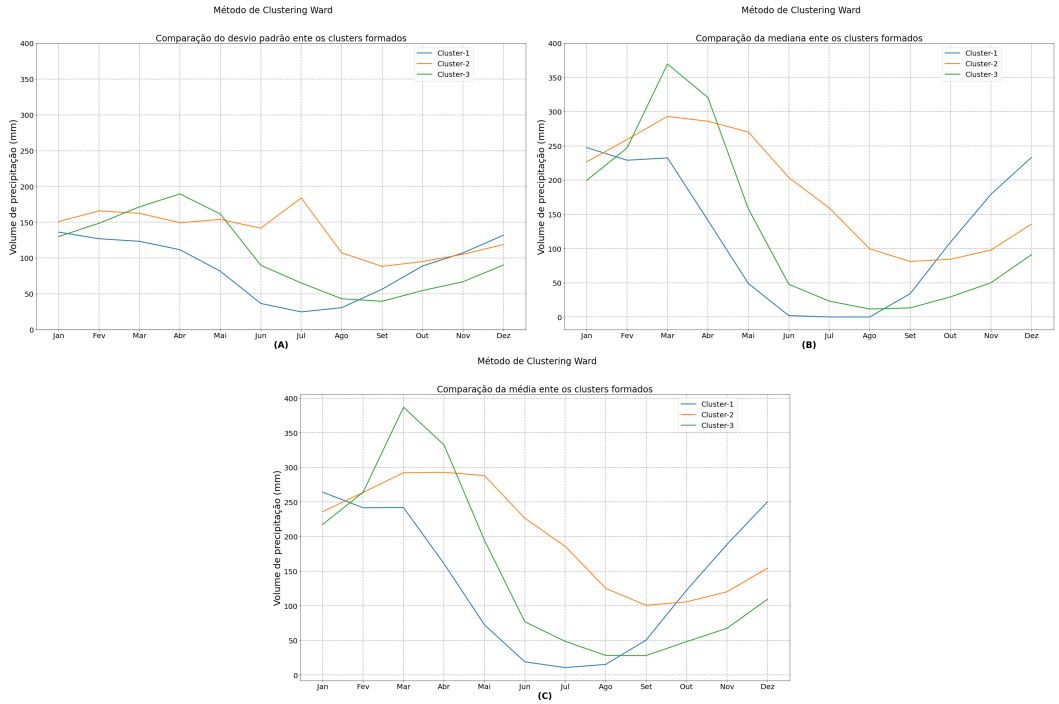
Os resultados apresentados nos gráficos demonstraram a distinção dos valores padrões para cada *cluster*, ressaltando as características próprias de cada grupo. O comportamento dos meses iniciais se mostraram semelhante entre os *clusters*, enquanto os meses entre abril e agosto apresentaram maior diferença nos valores de precipitação. Foi possível perceber que os meses que compõem o período de seca são os que apresentaram maior dissimilaridade entre os *clusters*, indicando que a identificação das características distintas entre os grupos é mais clara nesse período.

Figura 17 – Gráfico comparativo das medidas estatísticas padrões identificadas em cada *cluster* - Coordenadas polares: Formação com 3 *clusters*.



Fonte: Próprio Autor

Figura 18 – Gráfico comparativo das medidas estatísticas padrões identificadas em cada *cluster*: : Formação com 3 *clusters*.



Fonte: Próprio Autor

A análise de séries temporais de precipitação é uma aplicação importante no campo de estudo ambiental, pois permite entender as características de uma região em relação a esse fenômeno natural. Ao analisar essas séries, um dos objetivos é determinar se existem regiões homogêneas em relação à precipitação na área de estudo.

Na Amazônia Legal, por exemplo, essa análise pode ser crucial para entender como a precipitação afeta a vegetação, a fauna e a flora da região. Para realizar essa análise, um modelo DCAE foi utilizado para reduzir a dimensionalidade dos dados mensais das estações pluviométricas e gerar dados no espaço latente que representaram os valores mensais de precipitação.

Com os dados em dimensionalidade reduzida, foram aplicadas técnicas de *clustering* para avaliar os mesmos na busca por padrões e, consequentemente, separar as estações de acordo com os padrões encontrados. A utilização de técnicas de *clustering* permite a identificação de regiões homogêneas em relação à precipitação, o que pode ser extremamente útil para diversas áreas, incluindo a agricultura, a gestão de recursos hídricos e a previsão do clima.

A utilização de modelos de redução de dimensionalidade e técnicas de *clustering* é uma prática comum no campo de ML em geral, pois permitem a identificação de padrões e características que podem ser difíceis de detectar de outra forma. Na análise de séries temporais de precipitação, essa abordagem é especialmente importante, pois permite a identificação de regiões homogêneas e a compreensão das tendências e padrões da precipitação ao longo do tempo. Neste sentido, a figura 19 apresentou a disposição geográfica das estações pluviométricas na Amazônia Legal, bem como a configuração dos agrupamentos formados pelas estações.

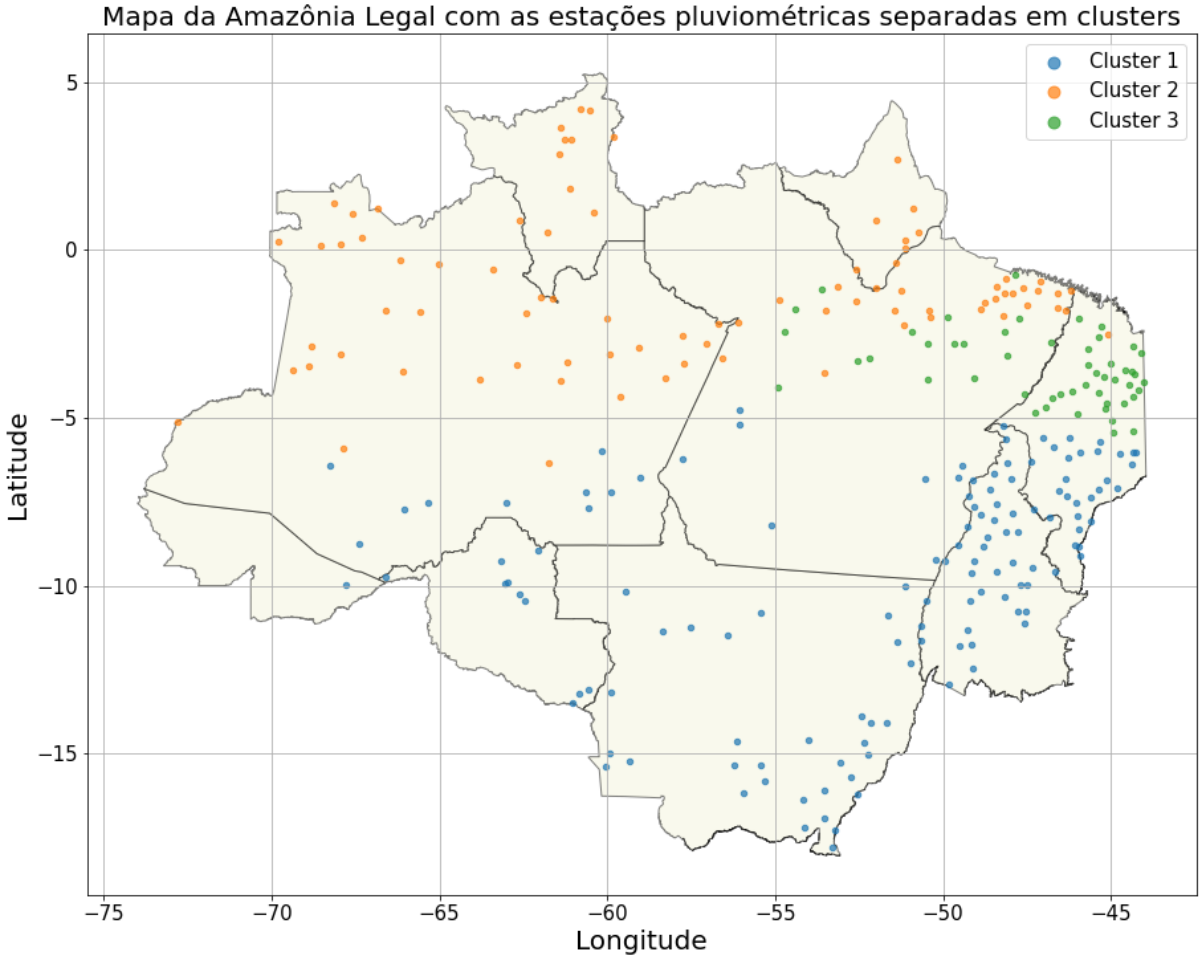
A análise de *clustering*, realizada com os dados das estações pluviométricas da Amazônia legal, permitiu-se observar uma divisão norte-sul com um grupamento menor em uma faixa ao leste entre os dois *clusters* maiores. Essa informação é valiosa para entender a distribuição da precipitação na região e pode ser utilizada para diversos fins práticos.

Um estudo anterior realizado por Lira et al. (2020a), destacou a importância de duas estações climáticas com maior volume de chuva em relação ao total de precipitação no estado do Pará. Quando a análise de *clustering* foi realizada com a formação em dois agrupamentos, foi possível observar a separação pluviométrica entre norte e sul no estado do Pará.

Além disso, a disposição dos *clusters* no mapa da Amazônia Legal pode fornecer informações valiosas sobre a distribuição da precipitação na região. Essas informações podem ser usadas para modelagem climática, previsão do tempo e outras aplicações práticas que exigem um conhecimento detalhado da distribuição da precipitação.

Como resultado da análise, foram encontrados três *clusters* que representaram diferentes regiões na Amazônia Legal: o cluster-1, representado por pontos azuis, ocupou a área mais ao sul; o cluster-2, representado por pontos laranja, concentrou-se na faixa ao norte; e o cluster-3, representado por pontos verdes, localizou-se em uma faixa mais ao leste, entre os dois outros *clusters*. Essa configuração espacial dos *clusters* sugere que a distribuição da precipitação na

Figura 19 – Mapa da Amazônia Legal com a disposição das estações pluviométricas formada por *cluster*: Formação com 3 *clusters*.



Fonte: Próprio Autor

região da Amazônia Legal apresenta uma clara divisão norte-sul, com uma zona de transição ao leste.

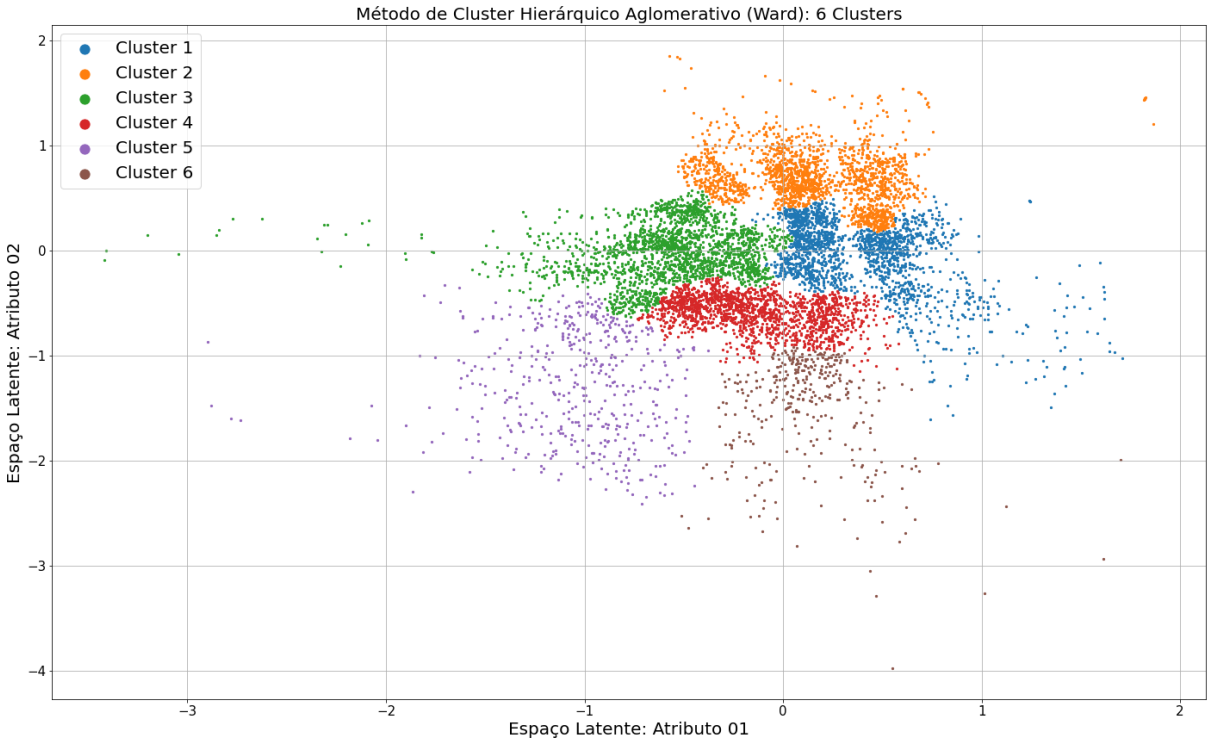
O fato da metodologia proposta ter encontrado padrões e agrupado as estações pluviométricas em regiões semelhantes, permitiu inferir que essas regiões apresentaram comportamentos climáticos similares no que diz respeito à precipitação. A identificação desses clusters é importante para a compreensão das variações da precipitação na região.

4.1.2 Resultado da aplicação da técnica de *clustering* para formação com 6 grupos.

Na figura 20 iniciou-se a análise dos resultados do processo de *clustering* com 6 grupos formados.

Ao analisar o gráfico fornecido, observou-se que foram identificados seis *clusters* de dados, cada um representado por uma cor diferente. Através da observação dos padrões de distribuição dos pontos de cada cluster, pôde-se inferir informações relevantes sobre a estrutura

Figura 20 – Gráfico de dispersão com dados do espaço latente: Formação com 6 clusters.



Fonte: Próprio Autor

dos dados.

O *cluster* 4, representado pela cor vermelha, apresentou uma melhor compactação entre suas observações. Isso indicou que as observações desse grupo estavam mais próximas umas das outras e, portanto, possuíam uma maior similaridade. Isso apontou para um indicativo de que as observações nesse *cluster* compartilharam características em comum.

Os *clusters* 1, 2 e 3, representados pelas cores azul, laranja e verde, respectivamente, apresentaram boa homogeneidade entre seus pontos, sugerindo que esses grupos possuíam uma maior consistência interna em relação às suas características. No entanto, houve uma certa dispersão entre algumas observações analisadas, indicando que esses *clusters* poderiam conter subgrupos com características distintas.

Já os *clusters* 5 e 6, representados pelas cores roxo e marrom, respectivamente, apresentaram grupos de observações com menor número de observações e mais dispersos comparados aos outros *clusters* formados. Indicando que esses grupos são mais heterogêneos e que poderiam estar compostos por observações que possuíam características distintas, o que poderia dificultar a identificação de padrões claros nessas áreas.

Ao observar o gráfico de dispersão com os 6 *clusters*, foi possível notar que houveram algumas mudanças em relação à metodologia utilizada para a formação de 3 grupos. Em particular, o *cluster* 3, representado pela cor verde, sofreu poucas alterações em sua composição, mantendo basicamente as mesmas observações. Indicando que as observações nesse *cluster*

possuíam características distintas das observações presentes nos outros *clusters*, e portanto, são mais robustas em relação às alterações metodológicas.

Ao comparar as metodologias com formações com 3 e 6 grupos, identificou-se que apenas os *clusters* 1 e 2, da primeira metodologia, sofreram alterações significativas. O *cluster* 1 da formação com 3 grupos foi dividido nos *clusters* 1 e 2 na formação com 6 grupos. Sugerindo a existência de duas subpopulações distintas nesse *cluster*. Já o *cluster* 2 da formação com 3 grupos foi dividido nos *clusters* 4, 5 e 6 na formação com 6 grupos, sendo que o *cluster* 4 obteve a maior fatia da divisão. Essa divisão demonstou que as observações presentes nesse *cluster* possuíam características distintas, e portanto, poderíam ser subdivididas em grupos mais homogêneos.

Após análise das observações e os *clusters* formados na configuração com seis grupos, realizou-se o ranking das observações, para determinar a qual grupo cada estação pluviométrica pertencia. Desta forma, os quantitativos das estações pluviométricas por *clusters* ficou definido conforme demonstrado na tabela 4.

Tabela 4 – Número de estações por *cluster* – Formação com 6 clusters

Identificação dos *clusters*	**Número de estações por *cluster***
Cluster-1	48
Cluster-2	69
Cluster-3	70
Cluster-4	60
Cluster-5	13
Cluster-6	8

Fonte: Próprio Autor

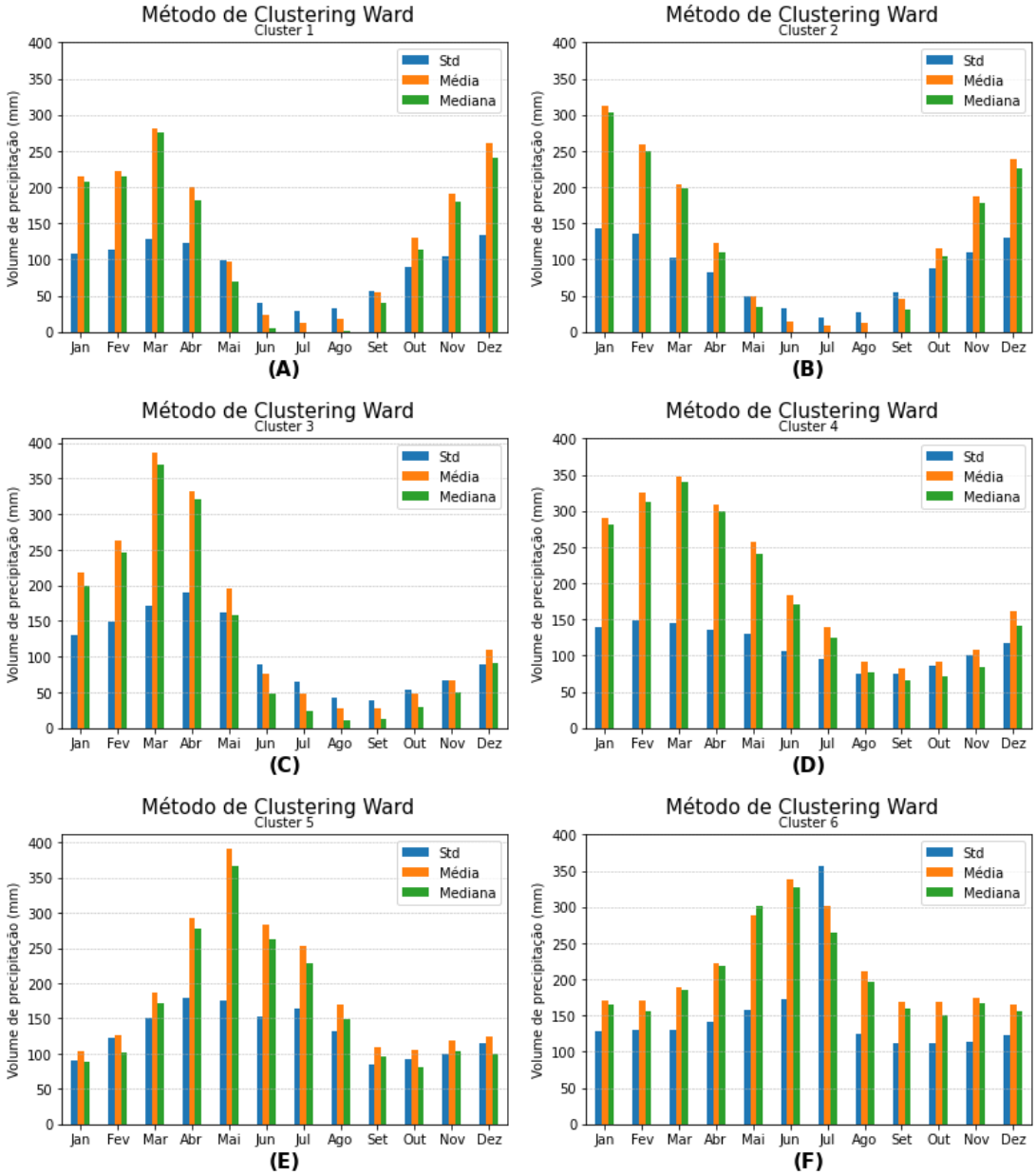
Apresentou-se nos gráficos da figura 21 as medidas estatísticas de mediana, desvio padrão e média dos *clusters* formados nessa configuração com seis agrupamentos.

Os gráficos apresentados na figura 21 mostram as medidas estatísticas de mediana, desvio padrão e média dos *clusters* formados na análise de *clustering* com seis agrupamentos. Ao analisar essas medidas, obteve-se informações valiosas sobre as características de cada grupo formado.

No gráfico 21.A, que representa o *cluster* 1, formado pelo método de *clustering* aglomerativo hierárquico com seis *clusters*, notou-se que o mês mais chuvoso nesse agrupamento corresponde ao mês de março. Além disso, é interessante notar que esse período de muitas chuvas iniciou-se em novembro e a queda no volume de chuvas começou em maio, estendendo-se até outubro, sendo julho o mês com o menor volume de precipitação.

Analisando as informações apresentadas no texto e os gráficos disponíveis, foi possível identificar a existência de um padrão sazonal no comportamento das chuvas na região representada pelo *cluster* 2, conforme indicado no Gráfico 21.B. Através da observação dos dados,

Figura 21 – Gráfico de barras com estatísticas (desvio padrão, média e mediana) dos *clusters*: Formação com 6 *clusters*.



Fonte: Próprio Autor

percebeu-se que os meses de maio a setembro são caracterizados como o período de maior seca, com julho apresentando o menor volume de chuva.

A partir do mês de outubro, no entanto, iniciou-se um processo de aumento gradativo no volume de chuvas na região, atingindo seu pico em janeiro, com volume médio superior a 300mm. Essa fase, caracteriza um período de maior pluviosidade na região, com um clima mais úmido e chuvas frequentes.

A partir de abril, por sua vez, as chuvas começaram a diminuir gradativamente, e a região iniciou uma transição para um período mais seco. Essa fase, caracteriza um período de menor pluviosidade, com um clima mais seco e chuvas menos frequentes.

Ao analisar a Figura 21.C, que representa o *cluster* 3, foi possível estabelecer que este é um dos *clusters* que apresentaram maior número de meses em um período de seca, com relação ao volume de precipitações. Durante o período seco, o volume médio de chuva no *cluster* 3 variou entre 25mm e 100mm, o que, certamente, indicou um desafio para atividades que dependiam de água, como a agricultura.

O período chuvoso no *cluster* 3, iniciou em janeiro e finalizou em maio, caracterizando uma fase de maior pluviosidade na região. Já o período com menor volume de chuva durou sete meses, iniciando em junho e percorrendo até dezembro. Entre os meses analisados, destacou-se o mês de março como o de maior volume de chuva e agosto como o de menor volume de precipitação.

O gráfico 21.D representa o *cluster* 4, que possui um padrão de precipitação bastante distinto dos demais *clusters*. Neste *cluster*, o mês de março se destacou com o maior volume de precipitação, com uma média aproximada de 350mm, enquanto o mês de setembro foi o de menor volume. O período mais seco abrange os meses de agosto a novembro. O mês de dezembro foi marcado pelo início das chuvas mais fortes neste *cluster*, caracterizando-se como um período de transição para o período de chuvas mais intensas, que se estenderam de janeiro a junho. Por sua vez, o mês de julho foi caracterizado pelo início da transição para os meses de menor volume de precipitação.

O *cluster* 5, exibido na figura 21.E, representa um padrão climático distinto, foi caracterizado por um período com chuvas intensas predominantes de abril a julho, tendo o mês de maio como o de maior volume médio de precipitação, e março como um mês de transição para o início das chuvas mais fortes. É importante destacar que janeiro foi um dos meses de menor volume médio de chuva nesse *cluster*. Além disso, agosto pôde ser considerado como o início da transição para o período mais seco, que se estendeu de setembro a fevereiro.

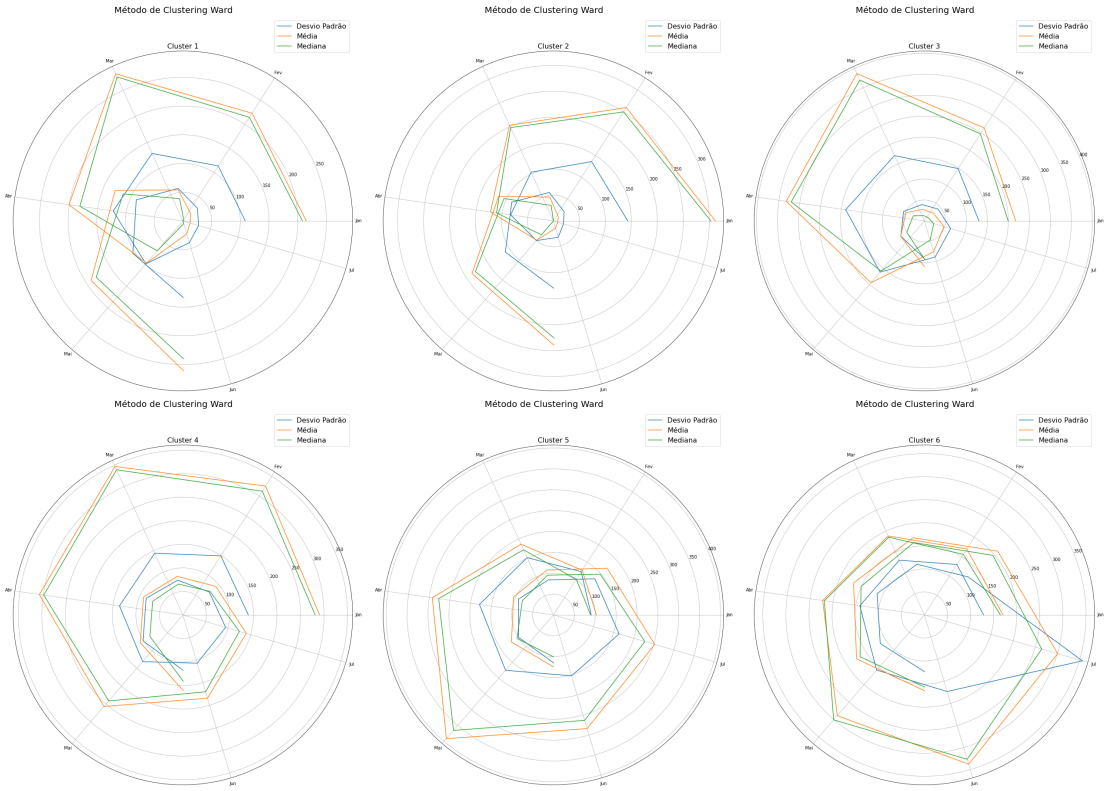
O *cluster* 6, descrito na figura 21.F, apresenta um período de seca bastante extenso, que se estendeu de setembro a março. Nesse *cluster*, o mês de abril pôde ser considerado a transição para o período de maior volume de precipitação. Esse *cluster* também foi caracterizado por um intervalo de chuvas mais fortes, que ocorreu entre os meses de maio a julho, tendo agosto como

o mês de transição para o início das secas.

Outra característica importante do *cluster* 6 foi o volume médio mensal de precipitação no período mais seco, que é superior aos dos demais *clusters*, medindo entre 150mm e 200mm. Sugerindo que mesmo durante os períodos considerados de seca, ainda foi possível o registro de chuvas fortes na abrangência do *cluster* 6.

As mesmas estatísticas foram geradas utilizando coordenadas polares e gráfico de linha para obtenção de uma abordagem diferente na análise destas informações. As figuras 22, 23 e 24 representam essas abordagens.

Figura 22 – Gráficos com estatísticas (desvio padrão, média e mediana) dos *clusters* formados em coordenadas polares: Formação com 6 *clusters*.



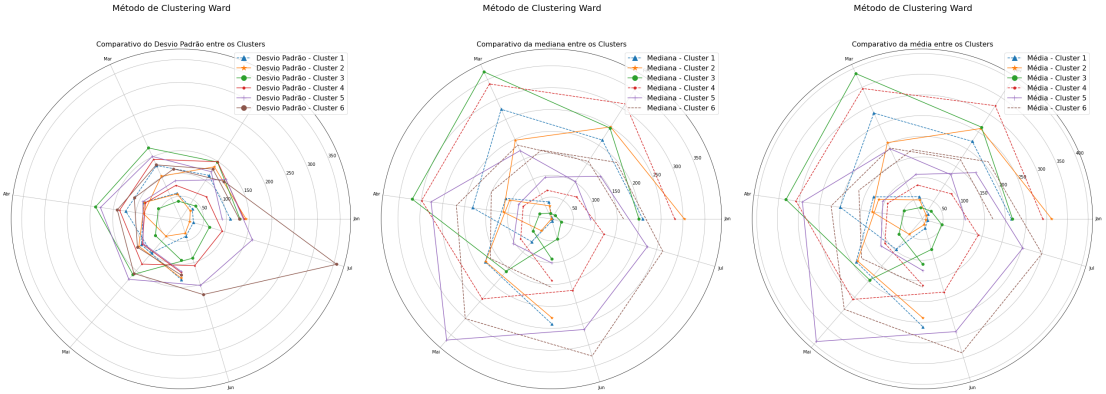
Fonte: Próprio Autor

A análise da distribuição geográfica das estações pluviométricas na Amazônia Legal é uma tarefa importante na busca por regiões homogêneas em relação ao componente climático de precipitação. Dessa forma, a configuração dos agrupamentos formados pelas estações, no formato com seis agrupamentos, é uma etapa relevante nesse processo. A partir dessa análise foi possível identificar regiões que possuíam comportamento similar em relação à precipitação.

Da mesma forma, como realizado na formação com três *clusters*, realizou-se a impressão da disposição geográfica das estações pluviométricas na Amazônia Legal, bem como a configuração dos agrupamentos formados pelas estações, no formato com seis agrupamentos. A figura 25 apresenta a disposição das estações para formação de seis grupos.

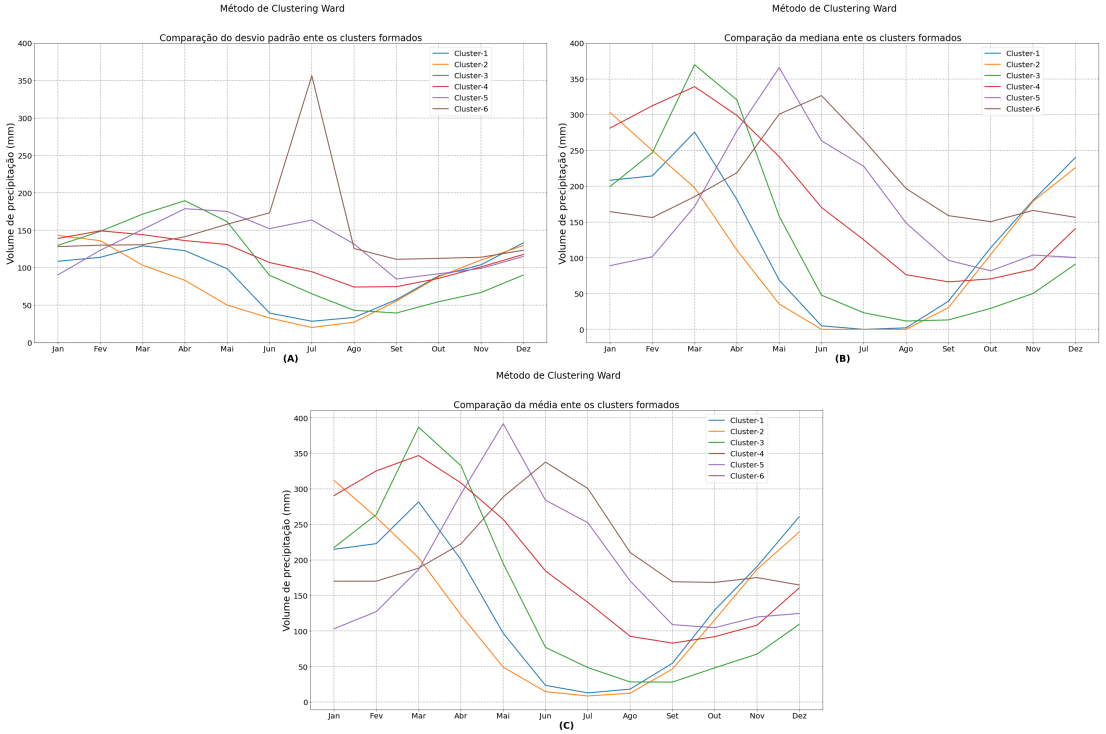
Realizar uma análise comparativa entre as diferentes configurações do número de *clusters*

Figura 23 – Gráficos com coordenadas polares das estatísticas das observações dos *clusters* comparado entre *clusters* formados: Formação com 6 *clusters*.



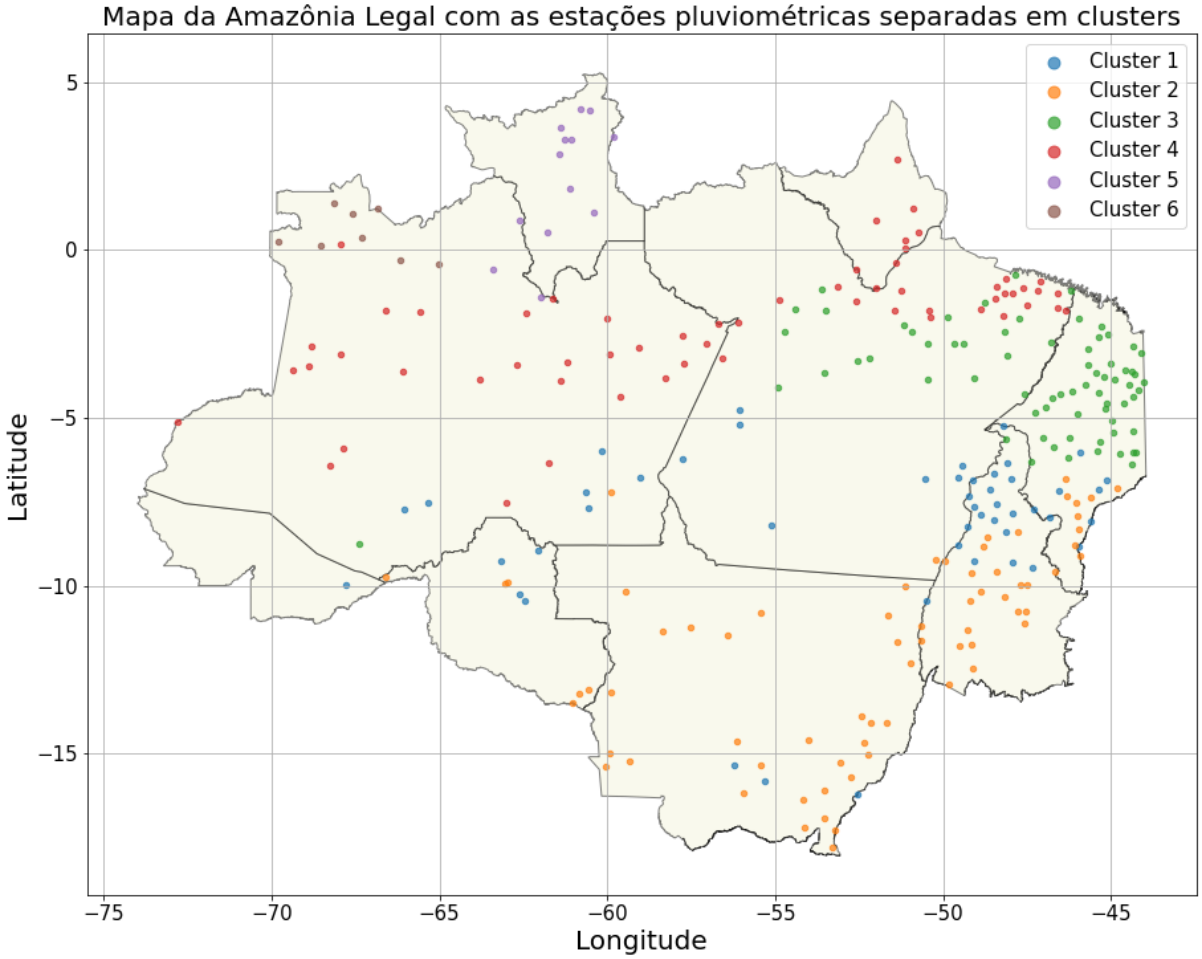
Fonte: Próprio Autor

Figura 24 – Gráficos das estatísticas das observações dos *clusters* comparado entre *clusters* formados: Formação com 6 *clusters*.



Fonte: Próprio Autor

Figura 25 – Mapa da Amazônia Legal com a disposição das estações pluviométricas: Formação com 6 *clusters*.



Fonte: Próprio Autor

é uma prática importante na análise de agrupamentos. No caso deste estudo, foram adotados 3, 6 e 9 grupos para avaliar a formação dos agrupamentos das estações pluviométricas na Amazônia Legal. Até o momento, foram avaliadas as formações com 3 e 6 grupos, o que permitiu comparar a disposição geográfica das estações pluviométricas nos diferentes agrupamentos. A escolha do número de *clusters* mais adequado deve considerar a complexidade dos dados e os objetivos do estudo, além de ser realizado com base em técnicas estatísticas e validação dos resultados. Com isso, é possível obter agrupamentos mais robustos e confiáveis, que possam fornecer informações importantes para a tomada de decisão

Dentre as configurações adotadas neste estudo, a formação com 6 agrupamentos apresentou os *clusters* 1 e 2, que poderiam ser considerados como uma subdivisão do *cluster* 1 da formação com 3 agrupamentos. Essa conclusão foi obtida a partir da análise das estações que foram alocadas em cada *cluster*, sendo que a maioria das estações do *cluster* 1 na formação com 3 agrupamentos foi alocada nos *clusters* 1 e 2 na formação com 6 agrupamentos. Essa informação foi relevante para entender a distribuição da precipitação nesses locais, permitindo uma análise mais detalhada das regiões que apresentaramm características similares em relação

à variável estudada.

O *cluster* 1 foi representado pelos pontos azuis que estão localizados em uma faixa central da Amazônia Legal, que se estende do sul do Maranhão até o Acre. Essa faixa passa pelo norte do Tocantins e sul do Amazonas, além de percorrer o sul e sudeste do Pará e o norte de Rondônia. É importante destacar que, mesmo sendo uma faixa central, houve algumas estações pluviométricas que se encontraram distantes dessa região, ficando ao sul do Mato Grosso. Esse *cluster* foi considerado um dos mais importantes, uma vez que apresentou uma grande variedade de estações e cobriu uma área geográfica extensa da Amazônia Legal.

Ao observar a distribuição das estações pluviométricas na Amazônia Legal, foi possível notar que o *cluster* 2, representado pelos pontos em laranja, apresentou certa uniformidade em sua disposição geográfica. A maioria das estações deste *cluster* foram localizadas no estado do Mato Grosso, com algumas estações no centro-sul do Tocantins e sul do Maranhão. Além disso, o *cluster* 2 possuiu outras cinco estações em Rondônia e uma no Amazonas, completando a sua formação.

Destaca-se nesta formação com 6 agrupamentos o *cluster* 3, que praticamente manteve a mesma disposição das estações do *cluster* 3 da formação com 3 agrupamentos. Suas estações se mantiveram numa região ao leste da Amazônia Legal, com maioria das estações localizadas no Maranhão, 43 estações no total. As outras estações ficaram localizadas da seguinte maneira: 24 estações ao norte do Pará; 2 estações na região norte do Tocantins; e 1 no sul do Amazonas, sendo essa a mais distante do grupo.

Outra informação importante, é o aumento no número de estações pertencentes a esse *cluster* com relação à formação anterior. Na primeira formação o *cluster* era composto por 51 estações pluviométricas, já nesse novo cenário o *cluster* dispõe de 70 estações. Sendo que quinze dessas novas estações vieram do *cluster* 1, e quatro dessas novas estações são provenientes do *cluster* 2, tendo como base a formação com 3 grupos.

Pôde-se destacar que na formação com 6 agrupamentos, o *cluster* 2 da formação anterior se dividiu em três *clusters* distintos, 4, 5 e 6. O *cluster* 4, representado pelos pontos em vermelho, se destacou por possuir estações pluviométricas distribuídas em regiões bastante distintas, como norte, nordeste e oeste do Pará, além de todas as estações do estado do Amapá e maioria das estações do Amazonas.

Essas informações são importantes para análises mais detalhadas sobre as características climáticas dessas regiões, permitindo uma melhor compreensão sobre as variações pluviométricas e suas possíveis implicações. A divisão do *cluster* 2 em três novos agrupamentos também pôde indicar mudanças significativas nas condições climáticas das regiões em questão, sugerindo a necessidade de estudos mais aprofundados sobre essas mudanças e suas implicações.

O *cluster* 5 abrangeu todas as dez estações pluviométricas pertencentes ao estado de Roraima, além de outras três estações localizadas ao norte do Amazonas, bem próximas à

fronteira com Roraima. As oitos estações pluviométricas que compõem o *cluster* 6, ficaram posicionadas ao norte do estado do Amazonas, próximas às fronteiras com a Colômbia e a Venezuela.

É importante ressaltar que a análise desses agrupamentos foi fundamental para entender a relação entre as variáveis climáticas na região da Amazônia Legal, permitindo a identificação de possíveis padrões e tendências que podem ser úteis em diversas áreas, desde o planejamento de ações de conservação ambiental até a gestão de recursos hídricos.

4.1.3 Resultado da aplicação da técnica de *clustering* para formação com 9 grupos.

Nesse tópico, iniciou-se o processo de análise dos dados de precipitação da Amazônia Legal, representados através dos atributos extraídos em baixa dimensionalidade pelo modelo DCAE proposto neste estudo, para a formação de agrupamentos divididos em nove *clusters*. Uma vez que os dados foram representados em baixa dimensionalidade pelo modelo DCAE, a técnica de *clustering* foi aplicada para identificar padrões nos dados e a segmentação dos dados em grupos homogêneos.

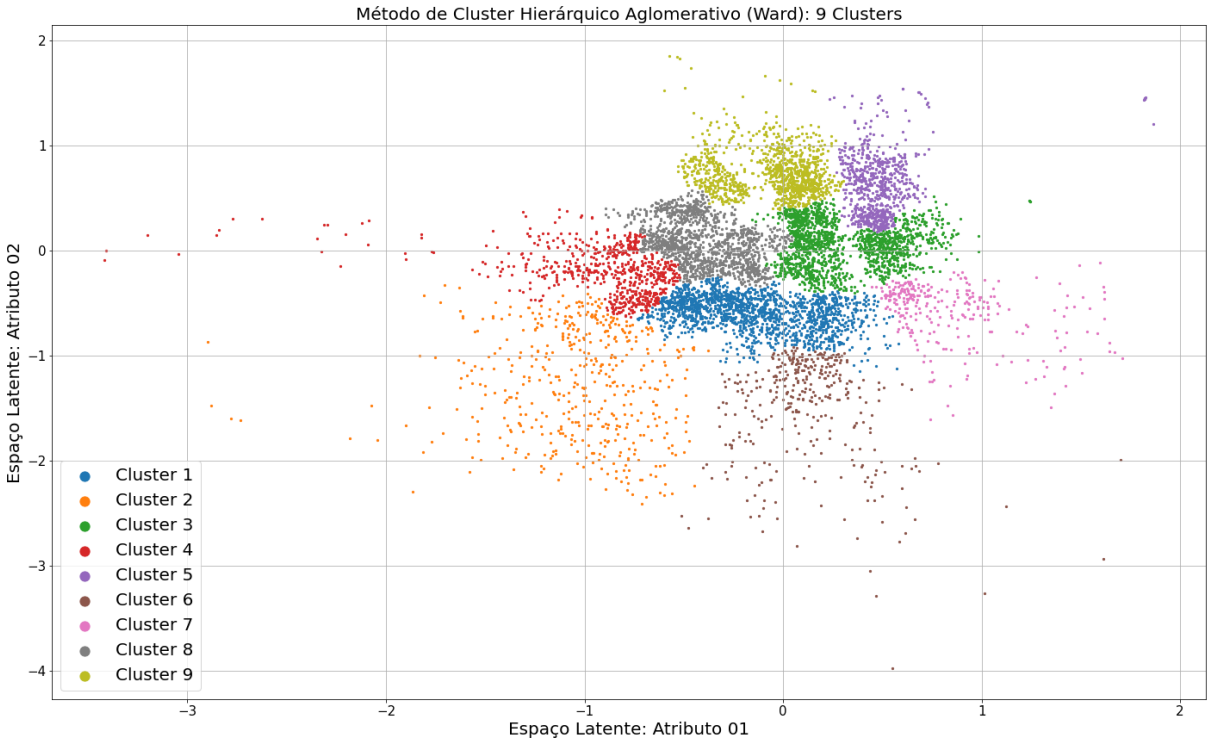
Ao analisar o gráfico de dispersão impresso na figura 26, pôde-se notar perfeitamente a divisão das observações em nove *clusters* distintos. Essa divisão é importante para a compreensão dos padrões de precipitação na Amazônia Legal. Ressalta-se, que nesse gráfico foram plotadas as 8040 observações correspondentes a cada ano, do período de 30 anos, das 268 estações estudadas.

Das 8040 observações avaliadas, o *cluster* 1 contou com a maior quantidade de amostras, com 1616 observações. Em seguida, o *cluster* 3 recebeu 1484 observações, e o *cluster* 8 recebeu 1308 observações. Por outro lado, os *clusters* com menor quantidade de observações foram o *cluster* 6, com apenas 360 amostras, e o *cluster* 7, com 293 amostras. O *cluster* 2 recebeu 484 observações, o *cluster* 4 recebeu 674 observações, o *cluster* 5 recebeu 723 observações, e o *cluster* 9 recebeu 1098 observações.

O estudo apresentou uma comparação entre diferentes configurações utilizadas na técnica de *clustering*, com o objetivo de avaliar o desempenho da formação de *clusters* a partir dos dados de precipitação da Amazônia Legal. A configuração em análise gerou nove *clusters* distintos, enquanto outras configurações foram utilizadas para gerar formações com três e seis agrupamentos.

Ao realizar a comparação entre essas diferentes configurações, foi possível notar que alguns grupos da formação com nove *clusters*, podem ser considerados subgrupos de *clusters* diferentes do atual, presentes em outras formações. Essa observação sugere que a utilização de nove *clusters* gerou uma segmentação excessiva dos dados, o que pode não ser ideal em determinados contextos de análise.

Figura 26 – Gráfico de dispersão com dados do espaço latente: Formação com 9 *clusters*.



Fonte: Próprio Autor

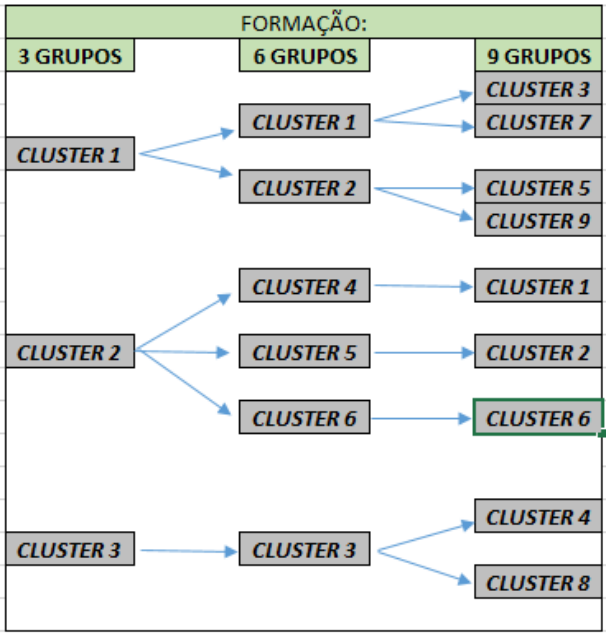
É importante destacar que a escolha do número de *clusters* é um aspecto crítico na técnica de *clustering*, e a determinação de um número ideal depende da natureza dos dados e dos objetivos da análise. A segmentação excessiva dos dados pode levar a agrupamentos pouco significativos, enquanto uma segmentação insuficiente pode não permitir a identificação de padrões importantes nos dados.

Analisando as formações dos agrupamentos obtidos através da técnica de *clustering* nesse estudo, foi possível identificar algumas relações interessantes entre as diferentes configurações utilizadas. Comparando a formação dos agrupamentos desse tópico com as demais formações que utilizaram três e seis agrupamentos, percebeu-se que o *cluster* 1 da formação com três grupos se dividiu em dois *clusters* na formação com seis grupos, os *clusters* 1 e 2.

Já na formação com nove agrupamentos, esses *clusters* 1 e 2 se dividiram em quatro *clusters*, sendo o *cluster* 3 formado a partir do *cluster* 1 da formação com seis grupos e o *cluster* 9 formado a partir do *cluster* 2 da formação com seis grupos. Além disso, o *cluster* 5 da formação com nove grupos foi formado a partir do *cluster* 2 da formação com seis grupos, enquanto o *cluster* 7 foi formado a partir do *cluster* 1 da formação com seis grupos.

Essas relações entre as diferentes formações de agrupamentos forneceram percepções valiosas para a compreensão dos dados de precipitação da Amazônia Legal, ajudando a identificar regiões com características pluviométricas semelhantes. Na figura 27, essa relação entre os agrupamentos foi demonstrada.

Figura 27 – Comparativo dos *clusters* das formações com 3, 6 e 9 grupos.



Fonte: Próprio Autor

Ao observar a disposição das observações no gráfico de dispersão e comparar com as formações com 3 e 6 grupos, pôde-se entender melhor a formação dos agrupamentos. De acordo com o quadro comparativo da figura 27, foi possível perceber a divisão dos *clusters* e suas origens relacionadas às demais formação. Esses subgrupos formados na configuração com divisão de nove *clusters* podem ser encaixados nas formações com 3 e 6 grupos, o que pode ser visto como uma vantagem dessa abordagem.

Dessa forma, a análise do gráfico de dispersão e do quadro comparativo pôde proporcionar uma compreensão mais aprofundada da formação dos agrupamentos e suas origens em relação às outras formações com menor número de grupos. Além disso, essa abordagem de divisão em nove *clusters* permitiu uma análise mais detalhada e específica dos dados de precipitação da Amazônia Legal.

A formação dos *clusters* por agrupamento com nove grupos, permitiu uma melhor compreensão da distribuição das observações no gráfico de dispersão. A partir daí, foi possível aplicar a metodologia de ranqueamento das observações para estabelecer a divisão das estações pluviométricas em grupos distintos.

Ao aplicar a metodologia de ranqueamento das observações na formação com nove agrupamentos distintos, foi possível estabelecer uma divisão das estações pluviométricas na Amazônia Legal em grupos com características semelhantes. Com a formação dos *clusters*, foi possível entender a relação entre as estações pluviométricas e a variável meteorológica, bem como, suas distribuições geográfica. O ranqueamento das observações também permitiu identificar as estações pluviométricas com as características mais semelhantes, facilitando a análise comparativa e a identificação de padrões climáticos na região.

O quantitativo de estações pluviométricas presentes nos *clusters* está descrito na tabela 5.

Tabela 5 – Número de estações por *cluster* – Formação com 9 *clusters*

Identificação dos *clusters*	**Número de estações por *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

Fonte: Próprio Autor

A análise de agrupamentos é uma técnica bastante utilizada em ciência de dados para identificar padrões e características em um conjunto de dados extraindo conhecimento dos mesmos. Nesse sentido, é importante apresentar medidas estatísticas que possam ajudar a visualizar e identificar esses padrões. No caso específico da formação de *clusters* com 9 grupos, assim como nas análises anteriores de 3 e 6 grupos, apresentou-se medidas estatísticas de média, mediana e desvio padrão de cada grupo formado.

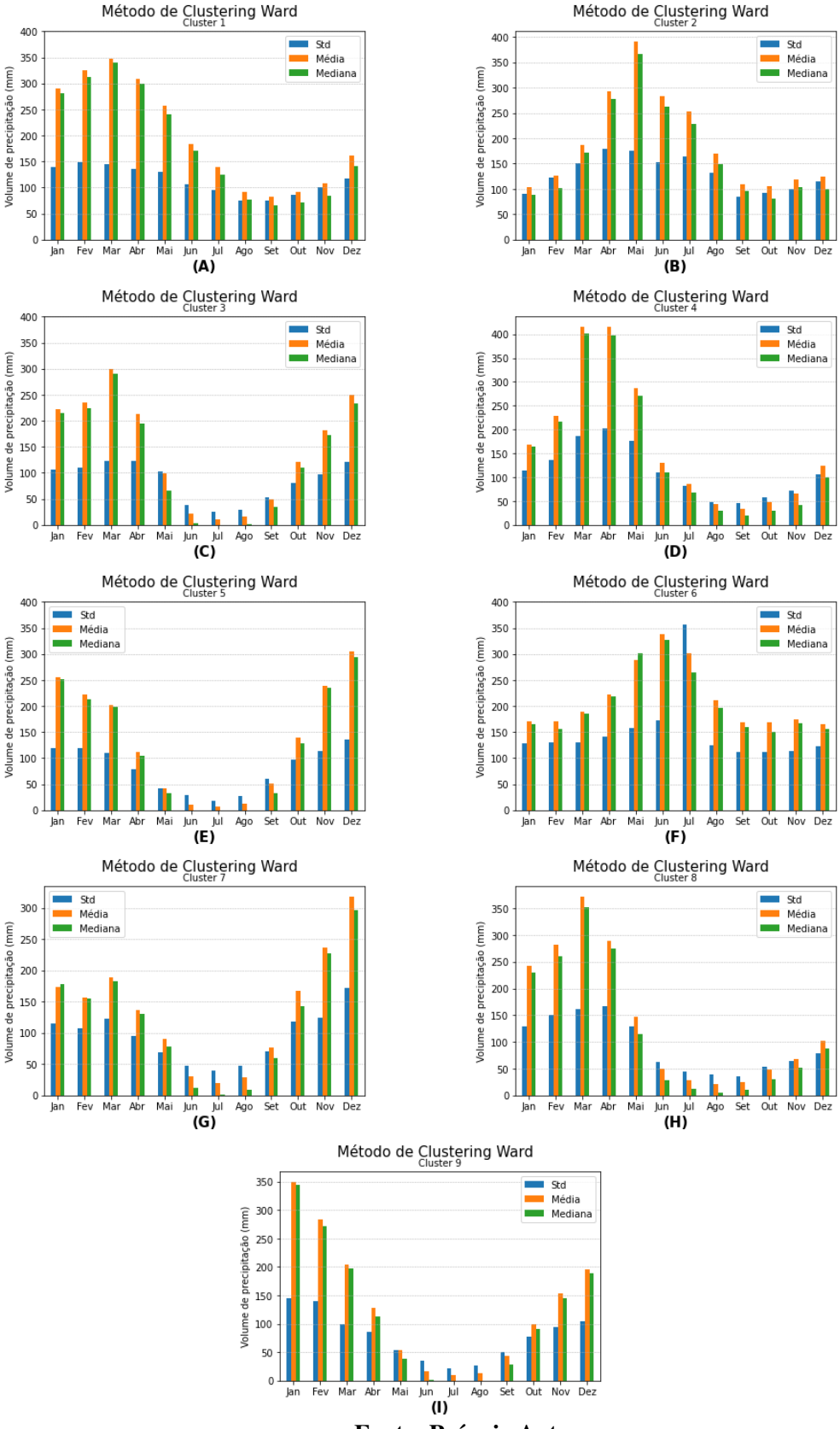
Essas medidas são importantes para resumir as informações contidas em cada *cluster* possibilitando que fossem comparadas entre si, proporcionando uma melhor compreensão das características de cada grupo, facilitando a interpretação dos resultados e a identificação de padrões e tendências nos dados. A média é uma medida que indica o valor central de um conjunto de dados, enquanto a mediana é uma medida que indica o valor que separa o conjunto de dados em duas partes iguais. Já o desvio padrão é uma medida de dispersão que indica o quanto os dados estão afastados da média.

Ao apresentar essas medidas estatísticas para cada *cluster*, foi possível identificar padrões pluviométricos que poderiam ser comuns a um determinado grupo de estações. Isso é importante para a compreensão do clima da região em estudo e para o planejamento de atividades que dependam das condições pluviométricas.

Na figura 28, apresentou-se, através de gráficos de barras, as estatísticas de cada *clusters* formado na configuração com nove grupos.

A análise do gráfico de dispersão da formação com nove agrupamentos, conforme apresentado na figura 26, revelou que alguns grupos desse conjunto são originados da divisão de grupos maiores encontrados em outras configurações. Esses subgrupos surgiram à medida que o número de *clusters* aumentou na técnica de agrupamento utilizada. Essa observação sugere que em diferentes formações com três, seis e nove agrupamentos, certos *clusters* compartilham características semelhantes.

Figura 28 – Gráfico de barras com estatísticas (desvio padrão, média e mediana) dos *clusters* formados: Formação com 9 *clusters*.



Fonte: Próprio Autor

Em particular, o *cluster* 1 da formação com nove grupos demonstrou semelhanças com os *clusters* 4 e 2 das formações com seis e três agrupamentos, respectivamente. Esses *clusters*, exibiram um padrão de chuvas mais intensas nos seis primeiros meses do ano, seguido por um período de estiagem nos seis meses seguintes. Vale ressaltar que o mês de março se destacou como o período de maior volume médio de precipitação, enquanto setembro foi o mês de menor volume médio. Essa descoberta, indicou a presença de uma tendência consistente nessas regiões, independentemente do número de agrupamentos considerados.

Ao analisar a formação em estudo, observou-se que o *cluster* 2 possuiu uma maior similaridade estatística com o *cluster* 5 da formação com seis agrupamentos. Ao comparar as estatísticas apresentadas nos gráficos de barra das figuras 15, 21 e 28, não foi possível identificar, na formação com três agrupamentos, um cluster que apresentasse padrões semelhantes ao *cluster* 2 em análise.

No entanto, é importante ressaltar que o *cluster* 5 da formação com seis grupos e o *cluster* 2 da formação com nove grupos são originados da divisão do *cluster* 2 da formação com três agrupamentos. Portanto, pôde-se considerar que o padrão encontrado nos *clusters* 2 e 5, que apresentaram características semelhantes, são subgrupos que estavam presentes no *cluster* 2 da formação com três grupos.

Observou-se que o *cluster* 3, representado pelo gráfico 28.C, apresentou padrões estatísticos aproximados aos demonstrados nos gráficos 15.A e 21.A, que representaram *clusters* da formação com três e seis agrupamentos respectivamente.

Esses *clusters*, compartilharam características semelhantes em relação aos padrões de precipitação. Nas regiões representadas por esses *clusters*, foi evidente um período de menor volume de precipitação, caracterizado por uma estiagem considerável nos meses de junho, julho e agosto. Por outro lado, as maiores médias de volume de chuva são registradas no período de dezembro a março.

O *cluster* 4, apresentou similaridade em suas características com *clusters* das demais formações analisadas. Seu padrão revelou que os meses de março e abril se destacaram com o maior volume médio de chuva, sendo o período de chuvas mais intensas compreendido entre janeiro e maio. Por outro lado, a temporada de maior estiagem ocorreu de julho a novembro, com os meses de agosto e setembro registrando os menores volumes de chuva.

Analisando as demais formações, identificou-se *clusters* com características próximas ao *cluster* 4, representados nos gráficos 15.C e 21.C. Esses *clusters*, também exibiram um padrão semelhante, com os meses de março e abril destacando-se pela maior média de precipitação e o período de janeiro a maio caracterizado por chuvas mais intensas. Similarmente, os meses de agosto e setembro são os de menor volume de chuvas nessas regiões.

O *cluster* 5, representado pelo gráfico 28.E, apresentou características estatísticas semelhantes ao *cluster* 1 da formação com três grupos e ao *cluster* 2 da formação com seis

agrupamentos, representados pelos gráficos 15.A e 21.B, respectivamente.

Uma diferença significativa entre esses *clusters* está no mês caracterizado como o de maior volume de chuva. No *cluster* do gráfico 28.E, o mês de dezembro registrou o maior volume médio de chuva, com uma média de 300 mm. Por outro lado, o *cluster* 2, representado pelo gráfico 21.B, também apresentou um volume médio de chuva em torno de 300 mm, mas o mês de ocorrência é janeiro. Já o *cluster* representado pelo gráfico 15.A indicou que janeiro é o mês mais chuvoso, embora com um volume médio de chuva um pouco superior a 250 mm.

O *cluster* 6, da formação com nove agrupamentos, representado pelo gráfico 28.F, apresentou características estatísticas interessantes. Ao analisar o padrão desse *cluster*, observou-se que ele possuía um período de estiagem prolongado, que vai de agosto a abril. Durante esses meses, o volume médio de precipitação varia entre 150 mm e 200 mm. Essa média de chuva é considerável, especialmente quando comparada a outros *clusters* que apresentaram médias semelhantes em seus períodos mais chuvosos.

No *cluster* 6, os meses de maio, junho e julho se destacaram como aqueles com o maior volume de precipitação. Especificamente, o mês de junho é caracterizado como o período de chuvas mais intensas nesse *cluster*.

Assim como ocorreu nos *clusters* 1, 2, 3, 4, 5 e 6, que apresentaram características semelhantes a outros *clusters* originários das demais formações, os *clusters* 8 e 9 também demonstraram alguma semelhança com *clusters* de outras formações. No entanto, foi identificado um caso em que não houve nenhuma semelhança, que ocorreu com o *cluster* 7.

O *cluster* 7 foi composto por dezesseis observações analisadas pelo DCAE, correspondendo a dezesseis anos de dados de precipitação das estações estudadas. Conforme mencionado anteriormente, uma mesma estação pôde apresentar observações em *clusters* distintos. No caso do *cluster* 7, após a etapa de ranqueamento das observações para determinar a qual *cluster* cada estação pertenceria, verificou-se que esse *cluster* não recebeu nenhuma estação, devido ao baixo número de observações relacionadas a cada estação. Portanto, o *cluster* 7 foi removido da formação.

Apesar disso, as observações desse *cluster* compartilharam características com outros *clusters*, especialmente com os *clusters* 3 e 5, que receberam cinco estações cada. Essas semelhanças indicaram padrões climáticos ou características geográficas específicas compartilhadas por essas estações.

O *cluster* 8, representado pelo gráfico 28.H, exibiu como característica em seu padrão de chuva uma fase de estiagem longa. Observou-se um período de baixo volume de precipitação que se estende por sete meses, de julho a dezembro. Nesse período, o volume médio de chuva variou aproximadamente entre 10mm e 90mm, sendo que agosto apresentou o menor volume médio. Por outro lado, março se destacou como o mês de maior volume médio de precipitação, com valores próximos a 350mm. O período chuvoso abrangeu os meses de janeiro a maio.

Essas características apresentadas pelo *cluster* 8 na formação com nove agrupamentos, assemelharam-se às características correspondentes do *cluster* 3 na formação com seis grupos, conforme ilustrado no gráfico 21.C, e ao *cluster* 3 na formação com três grupos, conforme demonstrado no gráfico 15.C. Essa similaridade sugeriu a existência de padrões climáticos semelhantes nessas regiões ou a presença de fatores geográficos comuns que influenciaram os padrões de chuva.

O gráfico 28.I, presente na figura 28, representou o padrão identificado para o *cluster* 9. Esse padrão, revelou um volume de chuva mais intenso nos primeiros meses do ano, especialmente em janeiro e fevereiro. Por outro lado, o período de maio a setembro foi caracterizado por um menor volume médio de precipitação, com uma média máxima de 50mm. Pôde-se considerar que o período chuvoso abrangeu de novembro a março, sendo janeiro o mês de maior volume de chuva. A estiagem, por sua vez, ocorreu de abril a outubro.

As características apresentadas por esse *cluster* assemelharam-se às características dos *clusters* 2 e 1 nas formações com seis e três grupos, respectivamente, conforme demonstrou-se nos gráficos 21.B e 15.A. Essa similaridade sugeriu uma relação entre os padrões de chuva desses *clusters* em diferentes configurações de agrupamentos.

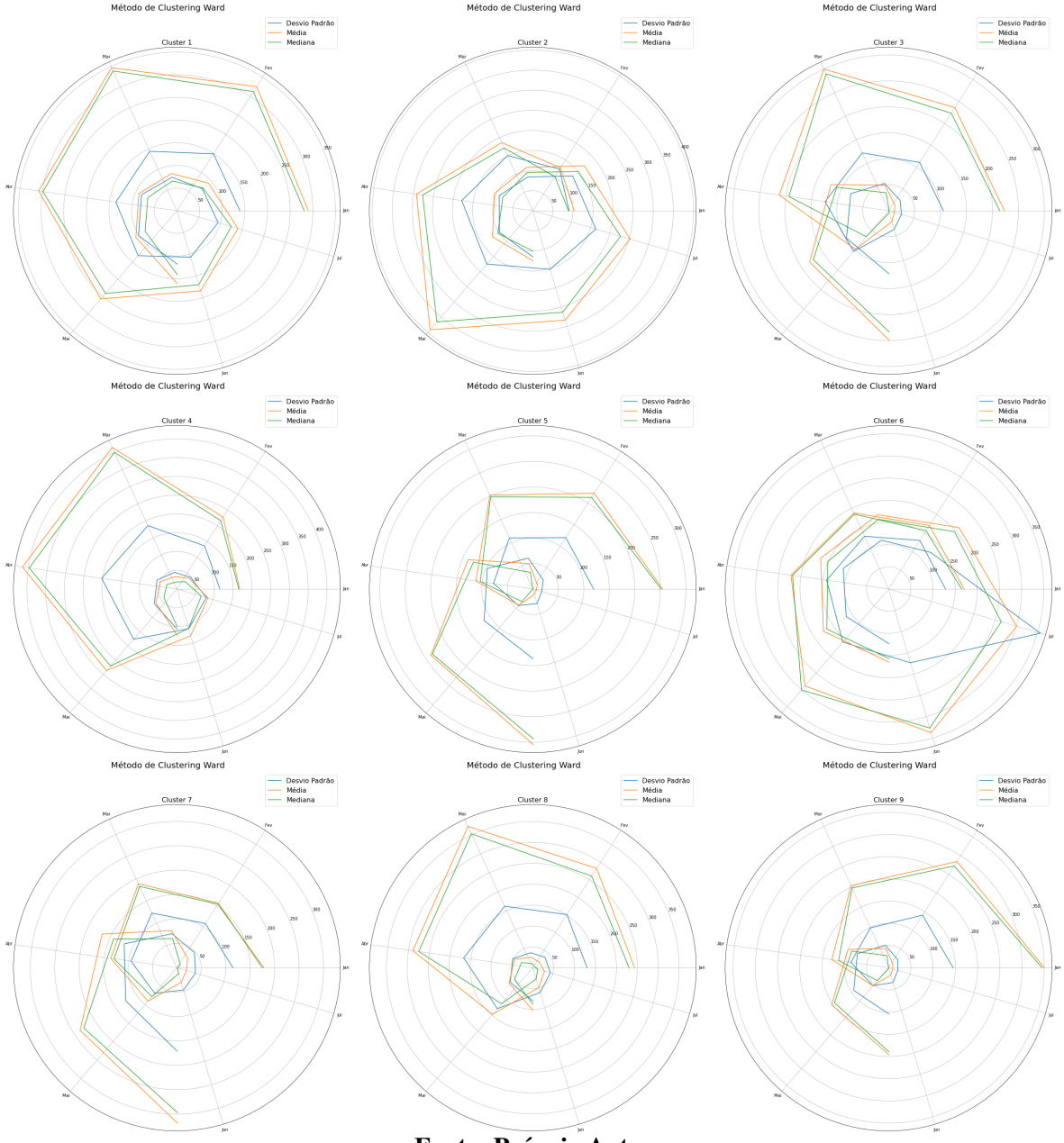
Na análise da formação com nove grupos em questão, seguiu-se a metodologia estabelecida anteriormente, que incluiu a aplicação de comparações estatísticas entre os grupos formados. Da mesma forma, nesta etapa, adotou-se uma abordagem diferenciada utilizando coordenadas polares e gráficos de linha para visualização e interpretação dos dados que representaram o ciclo anual de precipitação das estações que compuseram cada *cluster*. As figuras 29, 30 e 31 representam essas abordagens e fornecem percepções valiosas sobre os padrões e relações entre os grupos.

A figura 32, apresentou a disposição geográfica das 268 estações pluviométricas divididas em grupos que obtiveram padrões semelhantes em relação às características de precipitação. Essa representação espacial das estações permitiu identificar as regiões homogêneas em termos de padrões de chuva, o que proporcionou informações valiosas sobre a distribuição espacial das chuvas e a compreensão dos fatores que influenciaram a variabilidade pluviométrica.

Conforme ressaltado por Amanajás e Braga (2012), a aplicação de técnicas de agrupamento contribui para a delimitação de regiões pluviométricas homogêneas, ou seja, áreas onde os padrões de chuva ao longo do tempo são semelhantes. Isso possibilita uma compreensão mais precisa dos padrões de distribuição da precipitação e da variabilidade temporal das chuvas dentro da bacia hidrográfica.

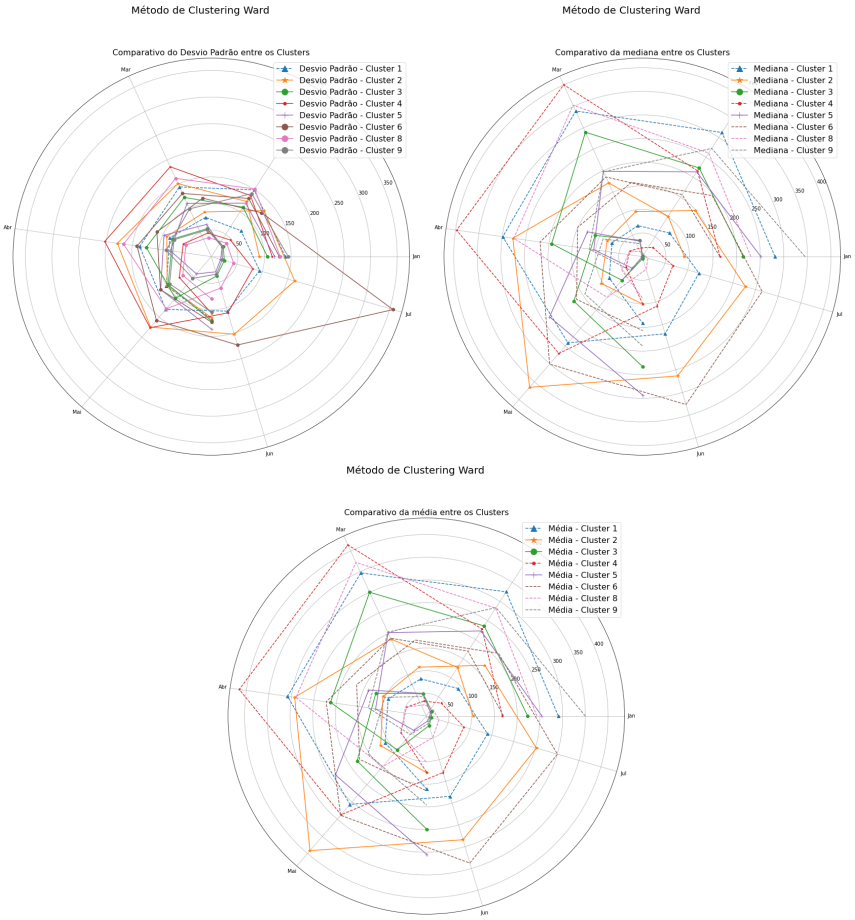
Ao comparar a disposição das estações pluviométricas em relação às estratégias de configuração do número de *clusters* utilizados nas análises, observou-se que a abordagem com nove agrupamentos resultou em regiões mais heterogêneas em termos das características de precipitação extraídas pelo modelo DCAE e analisadas pela técnica de agrupamento. Isso

Figura 29 – Gráficos em coordenadas polares com estatísticas (desvio padrão, média e mediana) dos *clusters*: Formação com 9 *clusters*.



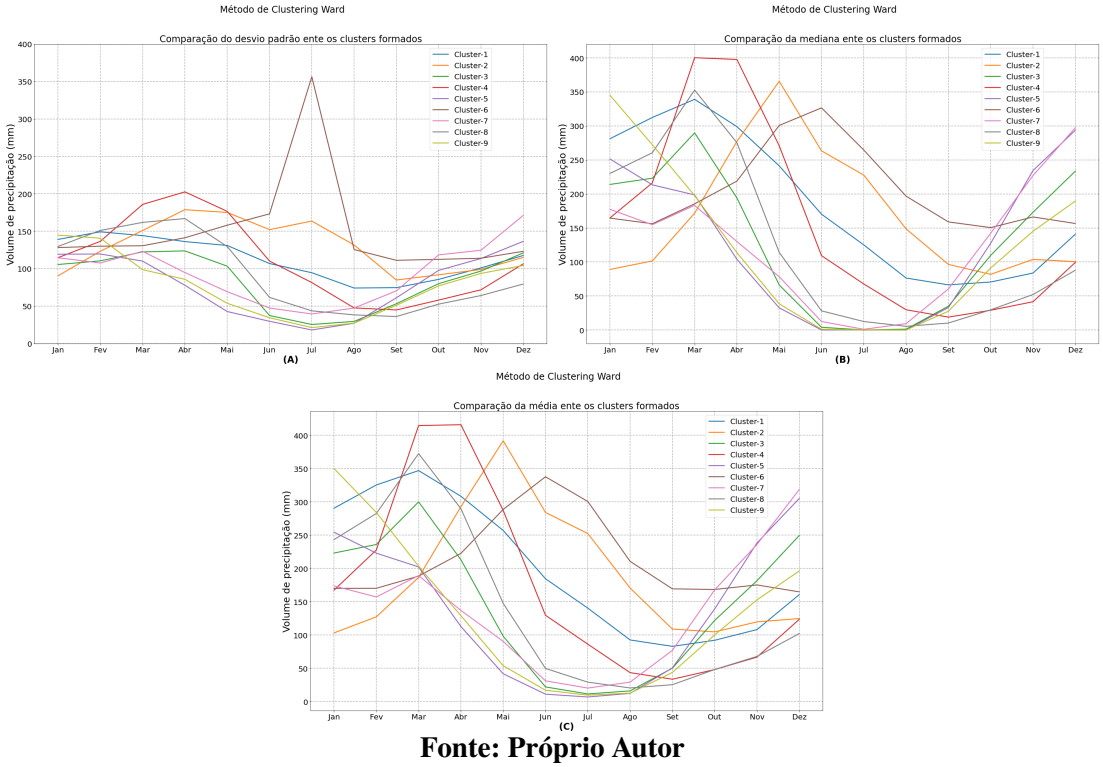
Fonte: Próprio Autor

Figura 30 – Gráficos em coordenadas polares comparando as estatísticas entre *clusters*: Formação com 9 *clusters*.



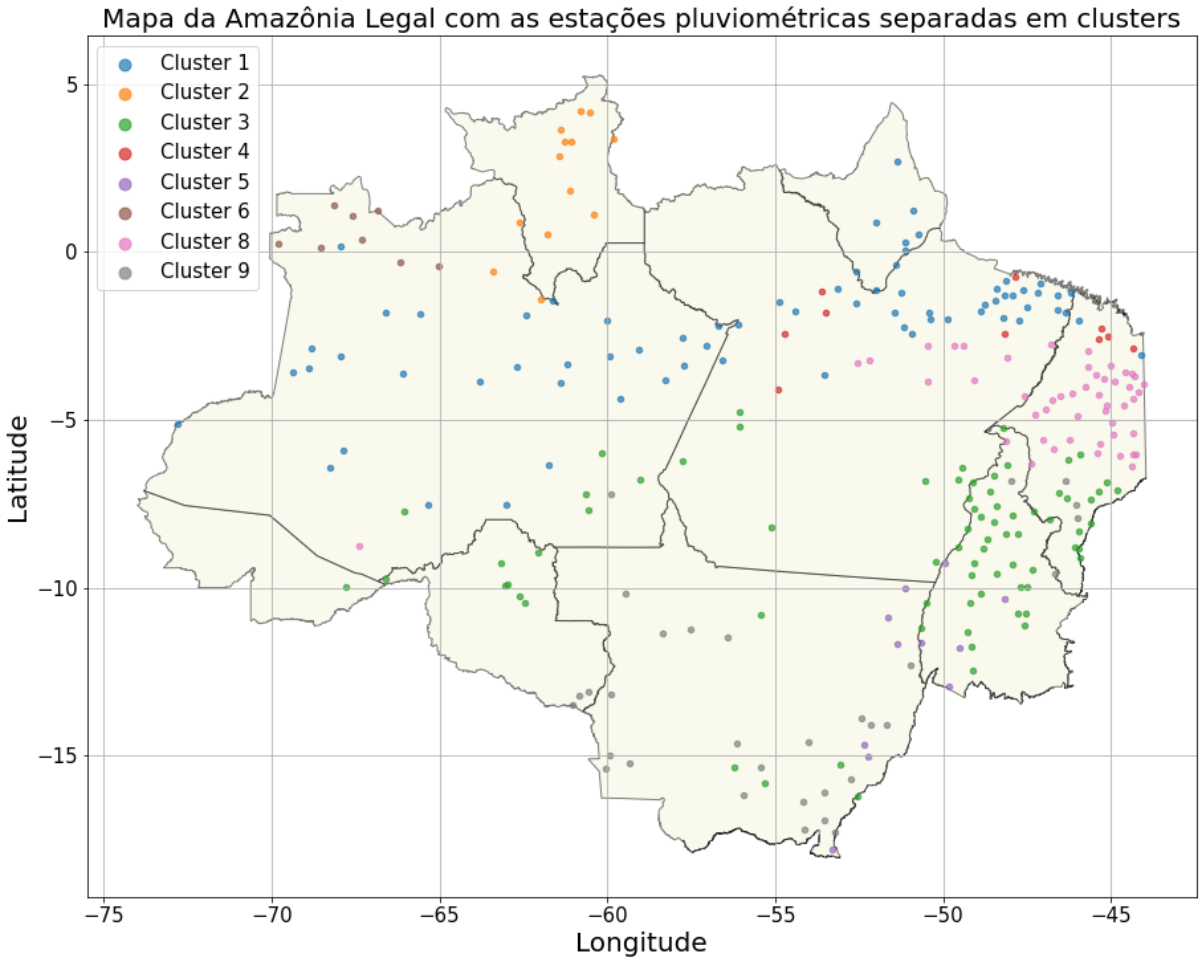
Fonte: Próprio Autor

Figura 31 – Gráficos comparando as estatísticas entre *clusters*: Formação com 9 *clusters*.



Fonte: Próprio Autor

Figura 32 – Mapa da Amazônia Legal com a disposição das estações pluviométricas: Formação com 9 *clusters*.



Fonte: Próprio Autor

significa que essa configuração permitiu identificar diferenças mais detalhadas nas características de precipitação entre as regiões.

É importante destacar que algumas regiões se dividiram em subgrupos, enquanto outras sofreram poucas alterações em sua composição. Isso possibilitou avaliar a consistência das características extraídas para essas regiões específicas. Os *clusters* 1, 2 e 8 foram os menos afetados na formação com nove agrupamentos, em comparação com as formações com três e seis grupos. Isso indicou que esses *clusters* possuem características distintas e estáveis ao longo das diferentes configurações de agrupamento.

5 CONCLUSÃO

Esta dissertação apresenta uma proposta de aplicação de redes neurais com aprendizado profundo do tipo *deep convolutional autoencoder* associada a técnica de *clustering* para análise de séries temporais de precipitação, objetivando a descoberta de conhecimento em baixa dimensionalidade dos dados, identificação de padrões e regiões homogêneas com relação à precipitação e reconstrução aproximada de séries temporais.

O modelo proposto foi treinado com dados de séries temporais de precipitação coletados em estações pluviométricas localizadas na Amazônia Legal. Este modelo pode ser abordado por diversos estudos que busquem utilizar análises de séries temporais de precipitação como método para auxiliar na previsão de chuva, classificação de regiões pluviométricas, entre outros resultados na região da Amazônia Legal.

A implementação construída neste estudo apresenta bons resultados, uma vez que conseguiu responder as três perguntas que nortearam a realização desta pesquisa. O primeiro questionamento foi: O modelo de deep convolutional autoencoder proposto é capaz de gerar uma representação significativa no espaço latente que possa realizar a descoberta de novos conhecimentos em séries temporais de precipitação? Após tratamento das séries temporais feita pelo modelo e aplicação da técnica de clustering, pôde-se responder de forma afirmativa o questionamento, pois as representações em baixa dimensionalidade dos dados das observações das estações selecionadas apresentaram uma boa interpretação dos *clusters* as quais pertenciam.

O segundo e terceiro questionamentos também foram respondidos pelo modelo de maneira positiva, pois de forma empírica pôde-se aplicar a configuração de hiperparâmetros que melhor se adequou ao modelo, e o nível de reconstrução dos dados originais também foram satisfatórios na medida que os erros na reconstrução foram baixos, com valores de RMSE e NRMSE iguais a 0.06610 e 0.3355, respectivamente.

O quarto questinamento também foi respondido positivamente, uma vez que ao aplicar a técnica de *clustering* nos dados do espaço latente, foi possível estabelecer regiões homogêneas, dentro da Amazônia Legal, com relação a variável de precipitação.

Mesmo com resultados promissores a estrutura do DCAE apresentada está aberta a melhorias. Mudanças na estrutura e configurações do modelo, bem como a imputação dos valores ausentes nas séries temporais utilizando outros métodos, podem gerar melhora no desempenho tanto na geração de padrões em baixa dimensionalidade com na reconstrução dos dados.

Trabalho futuros relacionados a esse estudo podem testar novas estruturas no DCAE, buscar e avaliar uma metodologia presente na literatura para configuração automática dos hiperparâmetros da rede, como também implementar uma segunda estrutura DCAE equivalente a primeira e aplicar como dados de entrada a saída da primeira rede DCAE do modelo.

REFERÊNCIAS

ALMEIDA, C. C. d. et al. Identificação e classificação de imagens usando rede neural convolucional e"machine learning": implementação em sistema embarcado. [sn], 2019.

ALMEIDA, C. T. d. et al. Avaliação das estimativas de precipitação do produto 3b43-trmm do estado do amazonas. **Floresta e Ambiente**, SciELO Brasil, v. 22, n. 3, p. 279–286, 2015.

AMANAJÁS, J. C.; BRAGA, C. C. Padrões espaço-temporal pluviométricos na amazônia oriental utilizando análise multivariada. **Revista Brasileira de Meteorologia**, SciELO Brasil, v. 27, p. 423–434, 2012.

BAIA, A. F.; CASTRO, A. R. G. A competitive structure of convolutional autoencoder networks for electrocardiogram signals classification. In: SBC. **Anais do XV Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2018. p. 538–549.

BAILÃO, A. S. d. O. et al. Reconhecimento de padrões por processos adaptativos de compressão. Universidade Federal de Goiás, 2020.

BARINO, F. O.; SANTOS, A. B. dos. Rede neural convolucional 1d aplicada à previsão da vazão no rio madeira. **XXXVIII Simpoósio brasileiro de telecomunicoes e processamento de sinais**, 2020.

BATZNER, K. **Convolutions in Autoregressive Neural Networks**. 2019. Disponível em: <https://theblog.github.io/post/convolution-in-autoregressive-neural-networks/>.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 35, n. 8, p. 1798–1828, 2013.

BENGIO, Y.; GOODFELLOW, I. J.; COURVILLE, A. Deep learning. Book in preparation for MIT Press. 2015. Disponível em: <http://www.iro.umontreal.ca/~bengioy/dlbook>.

BOUVEYRON, C.; BRUNET-SAUMARD, C. Model-based clustering of high-dimensional data: A review. **Computational Statistics & Data Analysis**, Elsevier, v. 71, p. 52–78, 2014.

BRUBACHER, J. P.; OLIVEIRA, G. G. d.; GUASSELLI, L. A. Preenchimento de falhas em séries temporais de precipitação diária no rio grande do sul. **Revista Brasileira de Meteorologia**, SciELO Brasil, v. 35, n. 2, p. 335–344, 2020.

BUZUTI, L. F.; THOMAZ, C. E. Understanding fully-connected and convolution allayers in unsupervised learning using face images. In: SBC. **Anais do XV Workshop de Visão Computacional**. [S.l.], 2019. p. 13–18.

CARTER, J. A. et al. Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes. **Expert Systems with Applications**, Elsevier, v. 115, p. 245–255, 2019.

CHEN, H. et al. Low-dose ct with a residual encoder-decoder convolutional neural network. **IEEE transactions on medical imaging**, IEEE, v. 36, n. 12, p. 2524–2535, 2017.

CHEN, M. et al. Deep features learning for medical image analysis with convolutional autoencoder neural network. **IEEE Transactions on Big Data**, IEEE, 2017.

CORCOVIA, L. O.; ALVES, R. S. Aprendizagem de máquina e mineração de dados: avaliação de métodos de aprendizagem. **Revista Interface Tecnológica**, v. 16, n. 1, p. 90–101, 2019.

CRISPIM, D. L. et al. Comparação de métodos de agrupamentos hierárquicos aglomerativos em indicadores de sustentabilidade em municípios do estado do pará. **Research, Society and Development**, v. 9, n. 2, p. e60922067–e60922067, 2020.

CRUZ, E. B. et al. Representação de séries temporais usando descritores de forma aplicados a recurrence plots. [sn], 2016.

DAVID, O. E.; NETANYAHU, N. S. Deeppainter: Painter classification using deep convolutional autoencoders. In: SPRINGER. **International conference on artificial neural networks**. [S.l.], 2016. p. 20–28.

DELAHAYE, F. et al. A consistent gauge database for daily rainfall analysis over the legal brazilian amazon. **Journal of hydrology**, Elsevier, v. 527, p. 292–304, 2015.

DOURADO, C. d. S.; OLIVEIRA, S. R. d. M.; AVILA, A. M. H. d. Análise de zonas homogêneas em séries temporais de precipitação no estado da bahia. **Bragantia**, SciELO Brasil, v. 72, n. 2, p. 192–198, 2013.

D'ANGELO, G.; PALMIERI, F. Network traffic classification using deep convolutional recurrent autoencoder neural networks for spatial–temporal features extraction. **Journal of Network and Computer Applications**, Elsevier, v. 173, p. 102890, 2021.

ESLING, P.; AGON, C. Time-series data mining. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 45, n. 1, p. 1–34, 2012.

ESSIEN, A.; GIANNETTI, C. A deep learning model for smart manufacturing using convolutional lstm neural network autoencoders. **IEEE Transactions on Industrial Informatics**, IEEE, v. 16, n. 9, p. 6069–6078, 2020.

FACELI, K. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. **Rio de Janeiro: LTC**, v. 2, p. 192, 2011.

GONÇALVES, E. D. et al. Identificação de regiões homogêneas e análise de regressão múltipla para regionalização de vazão na bacia hidrográfica do rio tapajós. **Revista Brasileira de Cartografia**, v. 69, n. 9, 2017.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

GUARIENTI, G. S. S. et al. Desenvolvimento de uma técnica computacional de processamento espaço-temporal aplicada em séries de precipitação. Universidade Federal de Mato Grosso, 2015.

GUO, K. et al. Angel-eye: A complete design flow for mapping cnn onto embedded fpga. **IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems**, IEEE, v. 37, n. 1, p. 35–47, 2017.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2001.

HEYLMAN, C. et al. Supervised machine learning for classification of the electrophysiological effects of chronotropic drugs on human induced pluripotent stem cell-derived cardiomyocytes. **PloS one**, Public Library of Science San Francisco, CA USA, v. 10, n. 12, p. e0144572, 2015.

HINTON, G. E.; SALAKHUTDINOV, R. R. Reducing the dimensionality of data with neural networks. **science**, American Association for the Advancement of Science, v. 313, n. 5786, p. 504–507, 2006.

HU, J. et al. Risk-aware path selection with time-varying, uncertain travel costs: a time series approach. **The VLDB Journal**, Springer, v. 27, n. 2, p. 179–200, 2018.

HUANG, H. et al. Modeling task fmri data via deep convolutional autoencoder. **IEEE transactions on medical imaging**, IEEE, v. 37, n. 7, p. 1551–1561, 2017.

HWANG, K.; CHEN, M. **Big-data analytics for cloud, IoT and cognitive computing**. [S.l.]: John Wiley & Sons, 2017.

ISHIHARA, J. H. et al. Avaliação do monitoramento pluviométrico da amazônia legal. **Engenharia Ambiental**, v. 10, n. 3, p. 132–144, 2013.

JIMÉNEZ, M. et al. Galaxy image classification based on citizen science data: a comparative study. **IEEE Access**, IEEE, v. 8, p. 47232–47246, 2020.

KARIMPOULI, S.; TAHMASEBI, P. Segmentation of digital rock images using deep convolutional autoencoder networks. **Computers & geosciences**, Elsevier, v. 126, p. 142–150, 2019.

KERAS. **Keras: Deep learning API written in Python**. 2021. Disponível em: <https://keras.io/about/>.

KIEU, T. et al. Outlier detection for time series with recurrent autoencoder ensembles. In: **IJCAI**. [S.l.: s.n.], 2019. p. 2725–2732.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LEE, K.; CARLBERG, K. T. Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders. **Journal of Computational Physics**, Elsevier, v. 404, p. 108973, 2020.

LI, B.; SANO, A. Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. **Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies**, ACM New York, NY, USA, v. 4, n. 2, p. 1–26, 2020.

LIRA, B. R. P. et al. Agrupamento de precipitação no estado do (pará), brasil. **Revista de Gestão de Águas da América Latina**, v. 17, n. 19, 2020.

LIRA, B. R. P. et al. Identificação de homogeneidade, tendência e magnitude da precipitação em belém (pará) entre 1968 e 2018. **Anuário do Instituto de Geociências**, v. 43, n. 4, p. 426–439, 2020.

LIRA, B. R. P. et al. Avaliação do comportamento e da tendência pluviométrica na amazônia legal no período de 1986 a 2015. Universidade Federal do Pará, 2019.

LYRA, G. B.; OLIVEIRA-JÚNIOR, J. F.; ZERI, M. Cluster analysis applied to the spatial and temporal variability of monthly rainfall in alagoas state, northeast of brazil. **International Journal of Climatology**, Wiley Online Library, v. 34, n. 13, p. 3546–3558, 2014.

MASCI, J. et al. Stacked convolutional auto-encoders for hierarchical feature extraction. In: SPRINGER. **International conference on artificial neural networks**. [S.l.], 2011. p. 52–59.

MENEZES, F. P.; FERNANDES, L. L.; ROCHA, E. J. P. da. O uso da estatística para regionalização da precipitação no estado do pará, brasil. **Revista Brasileira de Climatologia**, v. 16, 2015.

MIRANDA, A. C. R. et al. Regiões hidrologicamente homogêneas na amazônia com base nas precipitações mensais. Universidade Federal de Viçosa, 2016.

MORAES, H. R. S.; CASTRO, A. R. G. Deep neural networks with application to transformer failure diagnosis. In: SBC. **Anais do XV Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2018. p. 287–298.

MORO, F. L. et al. Análise do impacto da agregação dos fluxos ip nos algoritmos de aprendizado de máquina supervisionado voltados para a detecção de intrusão. In: SBC. **Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**. [S.l.], 2019. p. 946–957.

NEVES, R. R. et al. Identificação de regiões pluviometricamente homogêneas na sub bacia trombetas. **Revista AIDIS de Ingeniería y Ciencias Ambientales. Investigación, desarrollo y práctica**, v. 10, n. 2, p. 125–135, 2017.

NVIDEA. **Nvidea: "CUDA Zone"**. 2021. Disponível em: <https://developer.nvidia.com/cuda-zone>.

OLIVEIRA-JÚNIOR, J. F. d. et al. Cluster analysis identified rainfall homogeneous regions in tocantins state, brazil. **Biosci. j.(Online)**, p. 333–340, 2017.

PANDEY, B. K.; KHARE, D. Identification of trend in long term precipitation and reference evapotranspiration over narmada river basin (india). **Global and planetary change**, Elsevier, v. 161, p. 172–182, 2018.

PENHA, D. de P.; CASTRO, A. R. G. Convolutional neural network applied to the identification of residential equipment in non-intrusive load monitoring systems. In: **3rd International Conference on Artificial Intelligence and Applications**. [S.l.: s.n.], 2017. p. 11–21.

PONTI, M. A.; COSTA, G. B. P. D. Como funciona o deep learning. **arXiv preprint arXiv:1806.07908**, 2018.

PYTHON. **"Python"**. 2021. Disponível em: <https://www.python.org/>.

SAKURAI, R. **Implementando a estrutura de uma Rede Neural Convolucional utilizando o MapReduce do Spark**. 2017. Disponível em: <http://www.sakurai.dev.br/cnn-mapreduce/>.

SANTOS, E. B.; LUCIO, P. S.; SILVA, C. M. S. e. Precipitation regionalization of the brazilian amazon. **Atmospheric Science Letters**, Wiley Online Library, v. 16, n. 3, p. 185–192, 2015.

SENA, I. B. D. de C. et al. Determinação de regiões pluviometricamente homogêneas na bacia do rio doce/mg. **Revista Mineira de Recursos Hídricos**, v. 1, n. 1, 2020.

SEVERO, D. L.; SILVA, H. dos S.; TACHINI, M. Flutuações climáticas da precipitação no vale do itajaí (sc). **Revista de Estudos Ambientais**, v. 20, n. 2, p. 37–48, 2019.

SILVA, A. Maggioni e. Classificação de séries temporais baseada em análise de recorrência e extração de características. 2016.

SOUSA, M. L. d. S. et al. Variabilidade espaço-temporal da precipitação na amazônia durante eventos enos. Revista Brasileira de Geografia Física, 2015.

SOUSA, R. dos S.; GUEDES, E. B.; OLIVEIRA, M. B. L. de. Previsão anual de precipitações em manaus, amazonas: Um comparativo de técnicas de aprendizado de máquina. In: SBC. **Anais do IX Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais**. [S.l.], 2018.

SPÖRL, C.; CASTRO, E.; LUCHIARI, A. Aplicação de redes neurais artificiais na construção de modelos de fragilidade ambiental. **Revista do Departamento de Geografia**, v. 21, p. 113–135, 2011.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. [S.l.]: Pearson Education India, 2016.

TAN, S.; LI, B. Stacked convolutional auto-encoders for steganalysis of digital images. In: IEEE. **Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific**. [S.l.], 2014. p. 1–4.

TENSORFLOW. **"TensorFlow"**. 2021. Disponível em: <https://www.tensorflow.org/?hl=pt-br>.

VALENTE, J. M.; MALDONADO, S. Svr-ffs: A novel forward feature selection approach for high-frequency time series forecasting using support vector regression. **Expert Systems with Applications**, Elsevier, v. 160, p. 113729, 2020.

WANG, W.; ZHAO, M.; WANG, J. Effective android malware detection with a hybrid model based on deep autoencoder and convolutional neural network. **Journal of Ambient Intelligence and Humanized Computing**, Springer, v. 10, n. 8, p. 3035–3043, 2019.

WANG, Y.; YU, Z.; WANG, Z. A temporal clustering method fusing deep convolutional autoencoders and dimensionality reduction methods and its application in air quality visualization. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 227, p. 104607, 2022.

XIA, Y. et al. Learning discriminative reconstructions for unsupervised outlier removal. In: **Proceedings of the IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2015. p. 1511–1519.

YANG, B. et al. Pace: a pa th-ce ntric paradigm for stochastic path finding. **The VLDB Journal**, Springer, v. 27, n. 2, p. 153–178, 2018.

YANG, B.; GUO, C.; JENSEN, C. S. Travel cost inference from sparse, spatio temporally correlated time series using markov models. **Proceedings of the VLDB Endowment**, VLDB Endowment, v. 6, n. 9, p. 769–780, 2013.

YIM, O.; RAMDEEN, K. T. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. **The quantitative methods for psychology**, v. 11, n. 1, p. 8–21, 2015.

YIN, C. et al. Anomaly detection based on convolutional recurrent autoencoder for iot time series. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, IEEE, 2020.

YUE, J. et al. Spectral–spatial classification of hyperspectral images using deep convolutional neural networks. **Remote Sensing Letters**, Taylor & Francis, v. 6, n. 6, p. 468–477, 2015.

ZHANG, C. et al. An up-to-date comparison of state-of-the-art classification algorithms. **Expert Systems with Applications**, Elsevier, v. 82, p. 128–150, 2017.

ZOU, Q. et al. Sequence clustering in bioinformatics: an empirical study. **Briefings in bioinformatics**, Oxford University Press, v. 21, n. 1, p. 1–10, 2020.