

UNIVERSIDADE FEDERAL DE SANTA CATARINA  
CAMPUS REITOR JOÃO DAVID FERREIRA LIMA - CENTRO TECNOLÓGICO  
DEPARTAMENTO DE ENGENHARIA SANITÁRIA E AMBIENTAL  
CURSO DE ENGENHARIA SANITÁRIA E AMBIENTAL

Guilherme da Silva Izidoro

**Análise de Agrupamento sobre cenários de Energia Natural Afluente, em bacias do Sistema Nacional Interligado (SIN), gerados a partir da previsão de precipitação de médio prazo do Centro Europeu de Previsões Meteorológicas de Médio Prazo (ECMWF) aplicadas num modelo chuva-vazão.**

Florianópolis

2022

Guilherme da Silva Izidoro

**Análise de Agrupamento sobre cenários de Energia Natural Afluente, em bacias do Sistema Nacional Interligado (SIN), gerados a partir da previsão de precipitação de médio prazo do Centro Europeu de Previsões Meteorológicas de Médio Prazo (ECMWF) aplicadas num modelo chuva-vazão.**

Trabalho de Conclusão do Curso de Graduação em Engenharia Sanitária e Ambiental do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Bacharel em Engenharia Sanitária e Ambiental.  
Orientador: Prof. Dr<sup>a</sup>. Patrícia Kazue Uda  
Coorientador: MSc Filipe Viezzer da Silva

Florianópolis

2022

Ficha de identificação da obra elaborada pelo autor,  
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Izidoro, Guilherme da Silva

Análise de Agrupamento sobre cenários de Energia Natural Afluente, em bacias do Sistema Nacional Interligado (SIN), gerados a partir da previsão de precipitação de médio prazo do Centro Europeu de Previsões Meteorológicas de Médio Prazo (ECMWF) aplicadas num modelo chuva-vazão / Guilherme da Silva Izidoro ; orientador, Patrícia Kazue Uda, coorientador, Filipe Viezzer da Silva, 2022.

90 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia Sanitária e Ambiental, Florianópolis, 2022.

Inclui referências.

1. Engenharia Sanitária e Ambiental. 2. K-Means. 3. Sistema Interligado Nacional. 4. ECMWF-Extended-Range. 5. Análise de Agrupamento. I. Uda, Patrícia Kazue. II. da Silva, Filipe Viezzer. III. Universidade Federal de Santa Catarina. Graduação em Engenharia Sanitária e Ambiental. IV. Título.

Guilherme da Silva Izidoro

**Análise de Agrupamento sobre cenários de Energia Natural Afluente, em bacias do Sistema Nacional Interligado (SIN), gerados a partir da previsão de precipitação de médio prazo do Centro Europeu de Previsões Meteorológicas de Médio Prazo (ECMWF) aplicadas num modelo chuva-vazão.**

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Engenheiro Sanitarista e Ambiental e aprovado em sua forma final pelo Curso Engenharia Sanitária e Ambiental

Florianópolis, 13 de Dezembro de 2022.

---

Prof<sup>a</sup>. Maria Elisa Magri, Dr<sup>a</sup>.  
Coordenadora do Curso

**Banca Examinadora:**

---

Prof. Patrícia Kazue Uda, Dr<sup>a</sup>  
Orientadora  
Instituição UFSC

---

Prof. Davide Franco, Dr.  
Avaliador  
Instituição UFSC

---

Felipe Beltrán Rodríguez, Dr.  
Avaliador  
Norus

Dedico este trabalho aos meus pais, avós, tios e todos que vieram antes de mim.

## AGRADECIMENTOS

Agradeço primeiramente à toda minha família e às gerações de luta que travaram para eu poder estar escrevendo este trabalho. Tudo se torna relativamente fácil quando comparado ao esforço de vocês! Mais especificamente, agradeço à minha avó e meus dois avôs, que já não podem mais presenciar a conclusão desse trabalho.

Agradeço à minha companheira Júlia, por todo o amor e apoio que me oferece constantemente, durante meus altos e baixos.

Agradeço ao meu amigo, Tobias, que com sua facilidade de sempre, ofereceu ajuda com o código deste trabalho e tornou-o possível.

Agradeço aos amigos e orientadores deste trabalho, Patrícia e Filipe, pela paciência e tempo dedicado na formulação deste trabalho, e pelo acolhimento e preocupação que recebi durante a graduação.

Agradeço aos meus amigos de longa data, João, Duda e aos amigos de graduação, Ferreira, Marcelo, Jean, José, Antunes, Eduardo e Thiago, pelo companheirismo durante os momentos mais difíceis dessa graduação.

Agradeço aos colegas do LAHIMAR, Leo, Priscilla, Xande, Emanuel, Matheus e Davide, por me acolherem no melhor laboratório do departamento, por toda a conversa, troca de experiências e por me mostrarem o caminho que decidi trilhar dentro da graduação.

Agradeço à Norus pelos dados fornecidos e pela oportunidade de estar desenvolvendo este trabalho.

Agradeço à UFSC por toda a infraestrutura fornecida ao longo de cinco anos e meio juntos.

## RESUMO

O presente trabalho tem como objetivo avaliar a assertividade de previsões de Energia Natural Afluente (ENA) geradas com base em uma análise de agrupamento dos 51 membros da previsão meteorológica *ECMWF-Extended-Range*. Para isso, aplicou-se o algoritmo K-Means sobre cenários de ENA gerados a partir do produto Extended-Range do ECMWF, do dia 23 de Setembro de 2021. Esta data está dentro do período de transição de seco para úmido para diversas bacias do Sistema Interligado Nacional (SIN). ENAs das sub-bacias que compõem o SIN foram agregadas espacialmente, em duas regiões distintas, e temporalmente, através de médias semanais, conforme metodologia proposta pelo Operador Nacional do Sistema Interligado (ONS) em estudos de análise de agrupamentos sobre cenários de ENA. Foram testados agrupamentos com K variando de 0 a 10. Utilizou-se como função objetivo do algoritmo a distância de Frobenius e, através do Elbow Method, determinou-se como 10 o número ideal de grupos. Uma única série de ENA prevista foi gerada a partir da ponderação de membros representativos dos agrupamentos, com base na quantidade de membros associados à cada grupo, sendo esta série denominada de ENA ponderada. A ENA ponderada conseguiu prever crescimentos de ENA não previstos pela previsão oficial do ONS no Planejamento Mensal de Operação Energética (PMO) de Outubro de 2021, em todos os submercados do SIN. Os resultados sugerem que a ENA ponderada é mais assertiva que a previsão oficial do ONS nos submercados Sul, Sudeste e Norte, com coeficientes de Nash-Sutcliffe de 0.23, 0.46 e 0.12, e erro médio absoluto percentual de 15.8, 19.9 e 11.0, respectivamente. Apesar da maior assertividade, os valores das métricas calculadas para a ENA ponderada no estudo de caso não são satisfatórios, portanto, recomenda-se a extensão da metodologia para outras datas previstas em diferentes cenários hidrológicos, além da variação da agregação temporal e espacial utilizada.

**Palavras-chave:** Sistema Interligado Nacional, SMAP, *ECMWF-Extended-Range*, análise de agrupamento; K-Means.

## ABSTRACT

This work evaluates the assertiveness of Affluent Natural Energy (ENA) forecasts, which are generated from a cluster analysis considering 51 members of the meteorological forecast ECMWF-Extended-Range. The K-Means algorithm is applied over the multiple ENA scenarios which generated through the Extended-Range product, from September 23, 2021. This date is in between the transition period from dry to wet season for many basins from the Sistema Interligado Nacional (SIN). The hydrological basins that make-up the SIN, were merged spacially, in two distinct regions, and temporally, by weekly means, according to methodology proposed by the Operador Nacional do Sistema Interligado (ONS) in clustering studies of ENA. The number of clusters tested ranged from 2 to 10. The Frobenius distance is considered in the objective function. The ideal number of clusters found through the Elbow Method is 10. A single series of ENA forecast was generated from the weighting of representative members and the number of members clustered in each group, and this forecast series is called weighted ENA. The weighted ENA achieved to predict the rise of observed ENA which was noted predicted by the National Operator of the Interligated System (ONS) on the Monthly Planning of Energy Operation (PMO) of October 2021, in all submarkets of the Interligated National System (SIN). This showed that the weighted ENA was more accurate than the ONS forecast, on the Sul, Sudeste and Norte submarkets, with Nash-Sutcliffe coefficients of 0.23, 0.46 and 0.12, and the Mean Absolute Percentage Error (MAPE) of 15.8, 19.9 and 11.0, respectively. Despite better accuracy is shown, weighted ENA Nash-Sutcliffe and MAPE values aren't satisfactory, therefore, it is recommended the extension of the current methodology for other forecast dates in diferente hydrological scenarios, and the variation of temporal and spacial aggregation applied.

**Keywords:** Sistema Interligado Nacional, ECMWF-Extended-Range, cluster analysis, K-Means.



## LISTA DE FIGURAS

Figura 1 - Matriz Energética Brasileira no ano de 2022 (capacidade instalada).....	20
Figura 2 - Bacias do SIN .....	21
Figura 3 - Fluxograma representando o Dilema do Operador.....	22
Figura 4 - Estruturas de uma central hidrelétrica .....	24
Figura 5 - Estruturas de uma central hidrelétrica .....	29
Figura 6 - FPH de Furnas .....	31
Figura 7 – Principais domínios climáticos brasileiros.....	35
Figura 8 – Previsibilidade climática de diferentes regiões da América do Sul .....	37
Figura 9 – Dimensões de um MCGA .....	39
Figura 10 – Representação de uma previsão por conjunto.....	40
Figura 11 - Interação entre os componentes do modelo integrado de previsão ECMWF.....	42
Figura 12 - Método de agrupamento hierárquico por Single Linkage .....	43
Figura 13 - Método de agrupamento hierárquico por Complete Linkage .....	44
Figura 14 - Método de agrupamento hierárquico por Average Linkage .....	44
Figura 15 – Taxonomia das análises de agrupamento.....	45
Figura 16 - Fluxograma do algoritmo K-Means.....	47
Figura 17 - Exemplo do Elbow Method .....	48
Figura 18 - Exemplo do Silhouette Score.....	50
Figura 19 - Fluxograma da metodologia .....	54
Figura 20 - Agregação espacial escolhida para a análise de agrupamentos .....	55
Figura 21 - Precipitação acumulada nas semanas previstas .....	62
Figura 22 - Elbow Method .....	63
Figura 23 - Membros representativos de cada grupo x ENA Observada – Agregação Norte.....	65

Figura 24 - Membros representativos de cada grupo x ENA Observada – Agregação Sul ....	66
Figura 25 - Comparação de previsões – Agregação Norte .....	67
Figura 26 - Comparação de previsões – Agregação Sul.....	68
Figura 27 - Submercados do SIN .....	68
Figura 28 - Membros Representativos – Submercado Nordeste .....	69
Figura 29 - Membros Representativos – Submercado Norte .....	70
Figura 30 - Membros Representativos – Submercado Sudeste .....	71
Figura 31 - Membros Representativos – Submercado Sul .....	72
Figura 32 - Comparação de previsões – Submercado Nordeste.....	73
Figura 33 - Comparação de previsões – Submercado Norte .....	74
Figura 34 - Comparação de previsões – Submercado Sudeste.....	75
Figura 35 - Comparação de previsões – Submercado Sul .....	76
Figura 36 - %MLT por agrupamento .....	78
Figura 37 – Erro Médio Absoluto Percentual.....	79
Figura 38 - Coeficiente de Nash-Sutcliffe.....	80

## LISTA DE TABELAS

Tabela 1 - Estrutura dos dados de ENA utilizados .....	54
Tabela 2 - Agregação espacial utilizada .....	56
Tabela 3 - Dados de ENA com agregação espacial adotada .....	57
Tabela 4 - Dados de ENA com agregação espacial e temporal aplicada. ....	58
Tabela 5 - Número de membros por grupo.....	64

## **LISTA DE ABREVIATURAS E SIGLAS**

CGH - Centrais Hidrelétricas de Geração  
CPTEC - Centro de Previsão do Tempo e Estudos Climáticos  
ECMWF - European Centre for Medium-Range Weather Forecasts  
ENA - Energia Natural Afluyente  
ENS - Ensemble  
FPH - Função de Produção Hidrelétrica  
GEFS - Global Ensemble Forecast System  
NOAA - National Oceanic and Atmospheric Administration  
ONS - Operador Nacional do Sistema Elétrico  
PCH - Pequenas Centrais Hidrelétricas  
PDDE - Programação Dinâmica Dual Estocástica  
PMO - Planejamento Mensal de Operação Energética  
PMOP - Planejamento de Médio Prazo da Operação Hidrotérmica  
POE - Planejamento da Operação Energética  
SIN - Sistema Interligado Nacional  
SMAP - Soil Moisture Accounting Procedure  
SSE - Sum of Squared Errors  
UHE - Usina Hidrelétrica  
VNA - Vazão Natural Afluyente

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO .....</b>	<b>15</b>
1.1	OBJETIVOS .....	17
1.1.1	Objetivo Geral.....	17
1.1.2	Objetivos Específicos .....	17
<b>2</b>	<b>REVISÃO BIBLIOGRÁFICA .....</b>	<b>18</b>
2.1	BREVE HISTÓRICO DO SETOR ELÉTRICO BRASILEIRO.....	18
2.2	SISTEMA INTERLIGADO NACIONAL .....	19
2.3	SMAP .....	23
<b>2.3.1</b>	<b>VARIÁVEIS DE ENTRADA DO SMAP .....</b>	<b>26</b>
2.4	CÁLCULO DA ENERGIA NATURAL AFLUENTE .....	28
<b>2.4.1</b>	<b>Geração de energia elétrica.....</b>	<b>28</b>
<b>2.4.2</b>	<b>Vazão Natural Afluente (VNA) .....</b>	<b>32</b>
<b>2.4.3</b>	<b>Energia Natural Afluente (ENA).....</b>	<b>33</b>
2.5	CLIMA E TEMPO NO BRASIL .....	34
<b>2.5.1</b>	<b>Clima.....</b>	<b>34</b>
<b>2.5.2</b>	<b>Tempo .....</b>	<b>38</b>
2.6	PREVISÕES POR CONJUNTO .....	40
2.7	Métodos de Agrupamento.....	42
<b>2.7.1</b>	<b>K-Means.....</b>	<b>45</b>
2.8	Normalização dos dados .....	50
2.9	ÍNDICES ESTATÍSTICOS .....	51
<b>2.9.1</b>	<b>Coefficiente de Nash-Sutcliffe .....</b>	<b>51</b>
<b>2.9.2</b>	<b>Erro médio absoluto percentual .....</b>	<b>52</b>
<b>3</b>	<b>MATERIAIS E MÉTODOS .....</b>	<b>52</b>
3.1	MATERIAIS.....	52

3.2	MÉTODOS .....	52
3.2.1	Dados de entrada .....	54
3.2.2	Agregação espacial.....	54
3.2.3	Agregação temporal.....	57
3.2.4	Normalização dos dados.....	58
3.2.5	Aplicação do <i>K-Means</i> em espaço tridimensional.....	58
3.2.6	Elbow Method .....	59
3.2.7	ENA do membro representativo.....	59
3.2.8	ENA média ponderada .....	59
3.2.9	Comparação com ENA observada .....	60
3.2.10	COMPARAÇÃO COM ENA OFICIAL ONS .....	60
3.2.11	COMPARAÇÃO COM MÉDIA DO ENSEMBLE.....	60
3.2.12	MÉDIAS DE LONGO TERMO .....	61
3.2.13	MÉTRICAS CALCULADAS .....	61
4	RESULTADOS .....	61
4.1	MÉDIA DA PRECIPITAÇÃO PREVISTA.....	61
4.2	ESCOLHA DO NÚMERO DE GRUPOS.....	63
4.3	DISPERSÃO DOS MEMBROS REPRESENTATIVOS POR AGREGAÇÃO ..	64
4.4	COMPARAÇÃO COM OUTRAS PREVISÕES POR AGREGAÇÃO.....	66
4.5	DISPERSÃO DOS MEMBROS REPRESENTATIVOS POR SUBMERCADO	68
4.6	COMPARAÇÃO COM OUTRAS PREVISÕES POR SUBMERCADO .....	73
4.7	DISPERSÃO DOS MEMBROS EM RELAÇÃO À MLT .....	76
4.8	MÉTRICAS .....	78
5	CONCLUSÕES.....	81
6	REFERÊNCIAS.....	83
7	ANEXO A – CÓDIGO BASE DO ALGORITMO K-MEANS .....	88

## 1 INTRODUÇÃO

A água é um recurso natural essencial para o desenvolvimento humano, sendo utilizada para fins como abastecimento, geração de energia, agricultura, navegação, manufatura e resfriamento de indústrias. Entretanto, a necessidade de consumo contínuo desse recurso por parte da humanidade não condiz com o ciclo natural de disponibilidade deste na natureza, surgindo, assim, um dos fortes motivos para o entendimento da Hidrologia, que estuda a distribuição de água na superfície da terra e seu movimento sobre e sob a superfície, e pela atmosfera (DAVIE, 2008).

Reservatórios são infraestruturas que desempenham um papel importante no controle de cheias, irrigação, navegação, abastecimento de água e geração hidrelétrica. A geração hidrelétrica é o principal motivo para a construção de grandes reservatórios no Brasil (BRAVO et al., 2009) e, também, a principal fonte da matriz energética brasileira, chegando a representar 91% da energia produzida entre 2000 e 2012 (PASSAIA et al., 2020). Para que os reservatórios possam desempenhar a função para a qual foram projetados, é essencial que exista uma adequada gestão do mesmo, com conhecimento sobre demandas e fluxos d'água futuros (BRAVO et al., 2009).

O Sistema Interligado Nacional (SIN) abrange quase toda a energia gerada e distribuída no país. É um sistema hidro-termo-eólico-solar de grande porte, com forte predominância de usinas hidrelétricas e com múltiplos proprietários e que é administrado pelo Operador Nacional do Sistema Elétrico (ONS). Dentro deste contexto, as previsões de vazões naturais afluentes (VNA) a usinas hidrelétricas são informações fundamentais para o planejamento e programação do SIN, tendo em vista que essas subsidiam a tomada de decisão no Setor Elétrico e proporcionam ganhos sinérgicos de operação do SIN (PAIVA, 2020). O modelo de previsão de vazões adotado pelo ONS para previsões diárias e semanais é o SMAP (Soil Moisture Accounting Procedure), um modelo hidrológico conceitual, que utiliza como dados de entrada a precipitação observada, a precipitação prevista, a vazão observada e a evapotranspiração potencial (ÁVILA et al., 2022).

Atualmente, o ONS utiliza três modelos de previsão de precipitação distintos para gerar o dado de entrada de precipitação prevista para as duas primeiras semanas, denominando este produto de precipitação conjunta. Os modelos utilizados são: (i) o modelo ETA, do Centro de Previsão do Tempo e Estudos Climáticos (CPTEC); (ii) o modelo GEFS, da *National Oceanic and Atmospheric Administration* (NOAA) e (iii) o modelo ECMWF, do *European Centre for Medium-Range Weather Forecasts* (ECMWF). Ressalta-se que todos os modelos de

previsão passam por um processo remoção de viés antes da combinação dos dados em único resultado (ONS, 2022a). Os modelos (ii) e (iii) utilizam a técnica de previsão por conjunto (*Ensemble*), caracterizada pela implementação de perturbações iniciais nas condições iniciais, nas condições de contorno e nos parâmetros do modelo (CUO; PAGANO; WANG, 2011). Desta forma. Diferentes cenários de precipitação são gerados ao longo do horizonte previsto.

Além disso, o ONS desenvolve estudos sobre a extensão do horizonte de utilização do produto de precipitação conjunta no modelo chuva-vazão, saltando de duas semanas para todo o primeiro mês previsto (ONS, 2022a). Tal substituição requer um modelo de precipitação com horizonte de previsão maior ou igual a um mês.

Especificamente, o modelo ECMWF possui 51 cenários de precipitação, e conta com uma versão estendida, com horizonte de previsão de 46 dias, denominado *ECMWF-Extended-Range*. Até o final de 2022, há a expectativa de que o modelo ECMWF passe a contar com 100 cenários por rodada, além de aumentar a frequência das rodadas de 2 vezes por semana, para uma frequência diária (ECMWF, 2022a). Tal quantidade de cenários pode elevar o custo computacional e gerar grandes incertezas em modelos hidrológicos e de preços de energia elétrica (ONS, 2022a). Diante da necessidade de incorporação da evolução observada nas tecnologias de previsão do tempo no processo operacional de gestão de reservatórios do SIN, faz-se necessário uma redução do número de cenários de precipitação, com perda mínima de informação e minimizando custos, de modo a proporcionar melhores tomadas de decisões para o sistema.

Dentro do contexto de estágio obrigatório na empresa Norus, trabalhando na equipe do Pluvia, um produto que prevê afluências naturais, o presente trabalho propõe-se a fazer um estudo de caso sobre uma rodada do modelo *ECMWF-Extended-Range*, objetivando avaliar a assertividade de uma previsão de Energia Natural Afluenta (quantidade de energia capaz de ser gerada por determinada vazão natural que chega à uma UHE) gerada a partir de uma análise de agrupamento dos 51 membros dessa previsão, de acordo com a metodologia vigente sendo aplicada pelo ONS em estudos de análise de agrupamento (ONS, 2022a).



## 1.1 OBJETIVOS

### 1.1.1 Objetivo Geral

Avaliar a assertividade de previsões de ENA selecionadas com base em análise de agrupamento dos 51 membros da previsão meteorológica *ECMWF-Extended-Range*.

### 1.1.2 Objetivos Específicos

- Calcular uma ENA ponderada considerando os resultados da análise de agrupamento.
- Analisar e comparar os resultados do agrupamento entre as regiões.
- Comparar os resultados do método de agrupamento com dados observados.

## 2 REVISÃO BIBLIOGRÁFICA

### 2.1 BREVE HISTÓRICO DO SETOR ELÉTRICO BRASILEIRO

A história do setor elétrico no Brasil iniciou-se na última década do Império no Brasil, com a instalação da iluminação elétrica interna na Estação Central da Estrada de Ferro Dom Pedro II (atual central do Brasil), na cidade do Rio de Janeiro (PAULO, 2011). Em 1883 foi construída a primeira hidrelétrica brasileira, no município de Diamantina (MG), aproveitando as águas do ribeirão do Inferno, afluente do rio Jequitinhonha. Esta hidrelétrica foi seguida por outras duas, nos anos de 1885 e 1887, porém, assim como a primeira, eram voltadas para a autoprodução no setor industrial (GOMES et al., 2002).

Em 1888 foi criada a Companhia Mineira de Eletricidade, concessionária de serviço público para gerar e distribuir energia elétrica, pública e particular, para a cidade de Juiz de Fora (MG) e demais indústrias na região. Na virada do século, a exploração dos serviços urbanos de utilidade pública, como transportes, iluminação pública e produção e distribuição de eletricidade foi efetuada pelo capital estrangeiro canadense, nas cidades de São Paulo e Rio de Janeiro (GOMES et al., 2002).

Com a expansão urbana, pressionou-se a demanda por serviços urbanos. O desenvolvimento na geração de energia elétrica se deu principalmente pelo aproveitamento do potencial hidráulico do eixo Rio-São Paulo, resultando num aumento de mais de 600% da capacidade instalada no país entre 1907 e 1911. A crise de 1929 tensionou os conflitos por uso de energia entre o setor agroexportador e grupos ligados a atividades urbano-industriais. Tal correlação de forças foi determinante para a reformulação de políticas econômicas, que impactam o setor elétrico, e conseqüentemente, o uso dos recursos hídricos (GOMES et al., 2002).

Em 1934, se materializou, no âmbito regulatório, o Código de Águas (Decreto 24.643, de 10 de Julho de 1934). O código submetia ao governo a concessão e autorização de exploração da energia hidráulica, serviços de transmissão, transformação e distribuição de energia. Em 1945, surgiu a Companhia Hidroelétrica do São Francisco (Chesf), com o intuito de aproveitar os recursos hídricos do rio São Francisco. Criou-se, em 1952, o Banco Nacional do Desenvolvimento Econômico (BNDES), que financiou 46,3% do crescimento de capacidade instalada durante o período de 1956 e 1961, sob um projeto de desenvolvimento do setor elétrico sob o comando do Estado (GOMES et al., 2002). Tal política foi descontinuada efetivamente a partir de 1990, com a instituição do Plano Nacional de Desestatização (PND), iniciando-se às

desestatizações no setor elétrico a partir de 1995. A condição prévia para implementação dessa nova linha política no setor foi a desverticalização da cadeia de produção, separando as atividades de geração, transmissão, distribuição e comercialização de energia elétrica, que se tornaram áreas de negócio independentes (DA SILVA, 2001; GOMES et al., 2002).

Tais acontecimentos culminaram na criação da Agência Nacional de Energia Elétrica (Aneel) e Operador Nacional de Recursos Hídricos (ONS), em 1997 e 1998, respectivamente. Consolidou-se, portanto, na segunda metade dos anos 90, o complexo arcabouço regulatório para sustentar o novo padrão de concorrência no setor (GOMES et al., 2002).

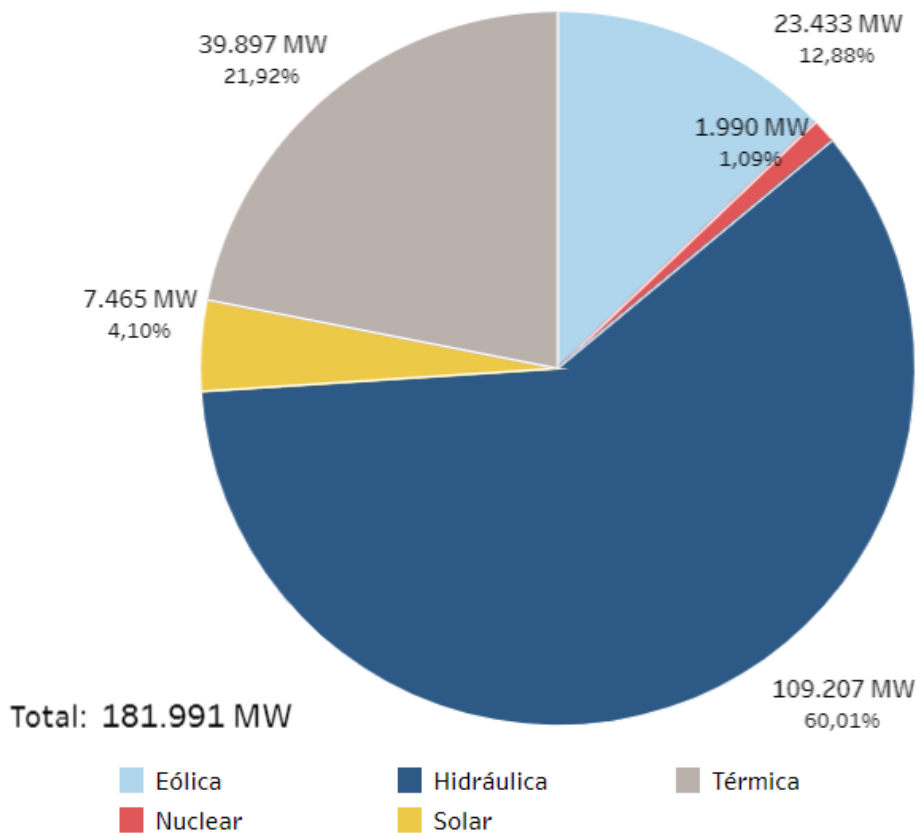
## 2.2 SISTEMA INTERLIGADO NACIONAL

O setor de energia elétrica atual brasileiro pode ser subdividido em quatro segmentos: geração, transmissão, distribuição e comercialização. O sistema de geração de energia é interligado nacionalmente pelo sistema de transmissão, com exceção de poucas regiões, sendo, este, denominado Sistema Interligado Nacional (SIN) (FUGIMOTO, 2010). De acordo com dados do Operador Nacional do Sistema Elétrico (ONS, 2020a), somente 3.

4% da capacidade de geração de energia elétrica do país se encontra fora do SIN. O segmento de distribuição é responsável por prover o serviço de acesso à rede de distribuição aos consumidores localizados em sua área geográfica de concessão (FUGIMOTO, 2010).

O Sistema Elétrico Brasileiro pode ser classificado como um sistema hidrotérmico, de predominância hidrelétrica (MATOS, 2012), com esta, chegando a representar 91% da geração de energia do país entre 2000 e 2012 (PASSAIA et al., 2020). Figura 1 mostra a matriz energética brasileira no ano de 2022.

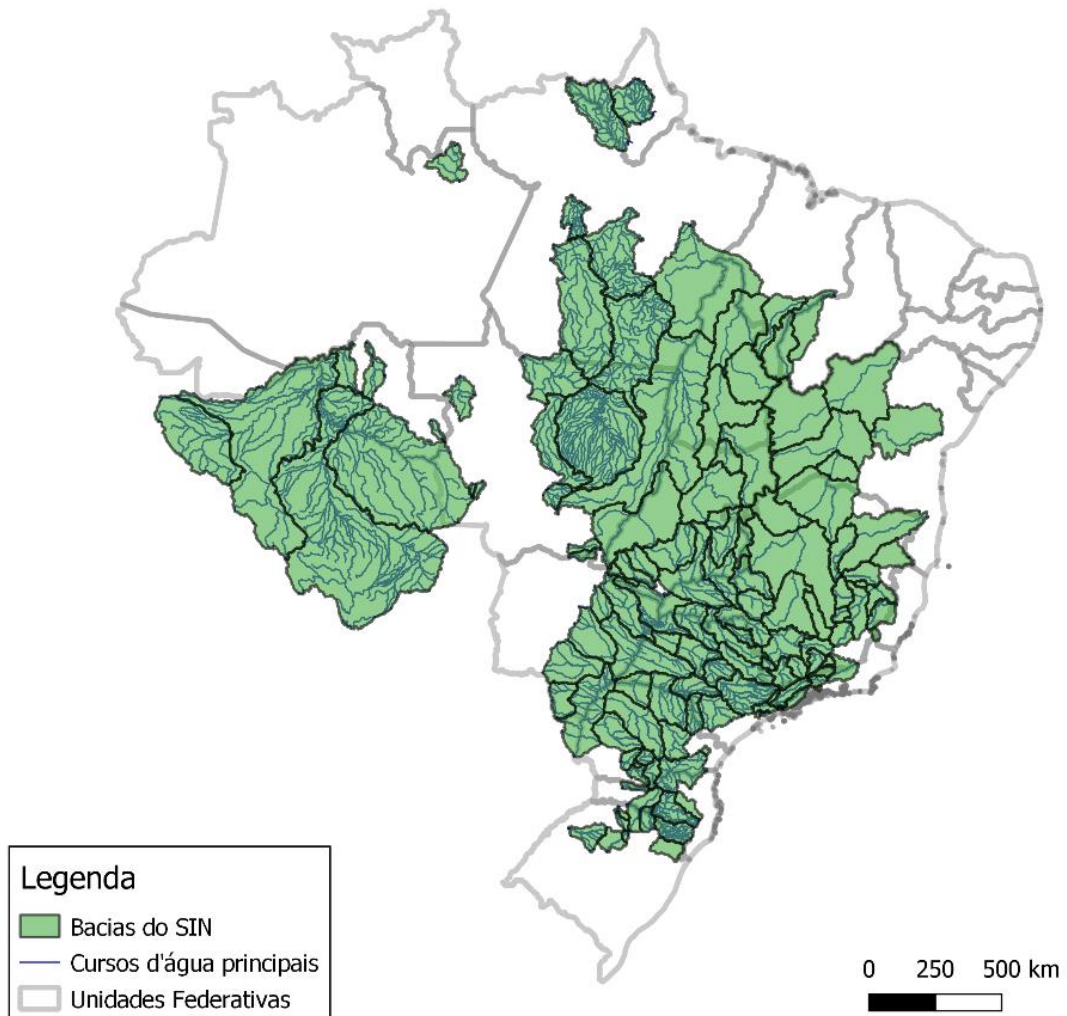
Figura 1 - Matriz Energética Brasileira no ano de 2022 (capacidade instalada)



Fonte: Adaptado de ONS (2022)

Em Sistemas de Energia Elétrica com predominância de recursos hídricos, o problema do Planejamento da Operação Energética (POE) é uma tarefa complexa de se realizar (MATOS, 2012), visto que a operação do sistema é dependente de um recurso natural de disponibilidade sazonal. Entretanto, um aspecto singular encontrado no Brasil é a dimensão do sistema elétrico, onde 99% deste é interconectado, permitindo com que blocos de energia gerados em uma região possam ser transmitidos para outras regiões, ou seja, permitindo que períodos de hidrologia favorável em uma região, possam compensar estiagens em regiões distintas (ARRUDA; SIECIECHOWICZ; BECKER, 2021). Outra particularidade do sistema elétrico brasileiro são a quantidade pequena de interconexões com países vizinhos e, portanto, a necessidade de garantir sua autossuficiência energética.

Figura 2 - Bacias do SIN

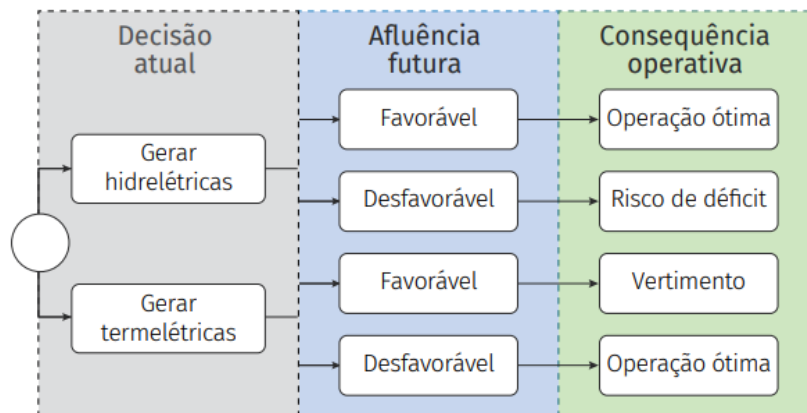


Fonte: Elaboração própria.

O ONS é o agente responsável pela operação e modelagem de todas as bacias mostradas na Figura 2, tomando decisões em escala semanal sobre quais usinas deverão gerar energia para atender a demanda ou não, com base em informações sobre a incerteza de vazões afluentes futuras (ARRUDA; SIECIECHOWICZ; BECKER, 2021), num horizonte de até 120 meses à frente.

O processo de decisão sobre a operação é conhecido como o “dilema do operador”, e é representado na Figura 3.

Figura 3 - Fluxograma representando o Dilema do Operador



Fonte: ARRUDA; SIECIECHOWICZ; BECKER (2021).

Caso se opte em utilizar hidrelétricas e a afluência for favorável, a decisão é tida como ótima, pois os níveis dos reservatórios serão restabelecidos ao final do horizonte. Por outro lado, se as afluências não forem favoráveis, o sistema corre risco de déficit energético. Em outra esfera, caso a decisão seja gerar termelétricas e a afluência futura for desfavorável, a decisão é tida como ótima, pois os níveis dos reservatórios estarão baixos para suprir a demanda. Entretanto, se as vazões forem favoráveis, as usinas sofrerão com vertimento e, portanto, desperdício energético (ARRUDA; SIECIECHOWICZ; BECKER, 2021).

As informações descritas acima representam apenas parte da complexidade do SIN, descritas a partir de um recorte da geração hidrelétrica e problemas derivados da gestão hídrica.

A Câmara de Comercialização de Energia Elétrica (CCEE) possui a atribuição de calcular o Preço de Liquidação das Diferenças (PLD), através de modelos como o NEWAVE, DECOMP e DESSEM, que serve como base para o cálculo do Custo Marginal de Operação (CMO).

O cálculo do CMO utiliza de uma cadeia de modelos matemáticos e estatísticos (SMAP, PREVIVAZ, GEVAZP, DECOMP e DESSEM) de forma a otimizar o custo de energia e de assegurar o seu fornecimento. Dentre os insumos destes modelos, estão os dados de Vazão Natural Afluyente (VNA) e Energia Natural Afluyente (ENA), de cada UHE do SIN. A ENA pode ser entendida como a quantidade de energia capaz de ser gerada por determinada vazão natural que chega à uma UHE, e sua utilização é propícia para obtenção da quantidade de energia que um conjunto de sub-bacias agregadas podem vir a gerar.

A cadeia de modelos utilizada durante o Planejamento Mensal de Operação (PMO) do SIN é estruturada da seguinte maneira:

- **SMAP:** modelo chuva-vazão determinístico utilizado nas 2 primeiras semanas previstas;

- **PREVIVAZ**: modelo estatístico e estocástico, que utiliza resultados do modelo SMAP para previsão da segunda à sexta semana;
- **GEVAZP**: modelo estocástico que utiliza resultados do primeiro mês previsto, vindos do SMAP + PREVIVAZ, para fornecer cenários de vazão no segundo mês previsto;
- **DECOMP**: modelo de otimização do planejamento da operação a curto prazo, que possui parte estocástica e parte determinística. Tal modelo utiliza previsões de vazões semanais do mês corrente, vindos do SMAP + PREVIVAZ, juntamente com cenários de afluência para o segundo mês previsto, vindos do GEVAZP, para calcular o CMO;

Atualmente, o ONS desenvolve estudos para substituição da previsão estocástica no primeiro mês previsto por uma previsão feita exclusivamente pelo modelo chuva-vazão (ONS, 2022a). Tal substituição requer um modelo de precipitação com horizonte de previsão maior ou igual à um mês. O escopo do presente trabalho está no horizonte de previsão que é atualmente coberto pelos modelos SMAP e PREVIVAZ, estudando a assertividade que o modelo estendido de precipitação do ECMWF possui em relação à ENA.

### 2.3 SMAP

De acordo com o NOS (2017) o Soil Moisture Accounting Procedure (SMAP) vem sendo aplicado como modelo oficial de previsão de vazões do ONS desde 2008, as quais são utilizadas para o cálculo da ENA. Tal modelo foi desenvolvido por LOPES (1982), baseado na aplicação do modelo Stanford Watershed IV (HOLMES; LINSEY, 1966) e do modelo Mero pelo Departamento de Águas e Energia Elétrica do estado de São Paulo (DAEE).

O modelo SMAP utilizado pelo ONS possui algumas adaptações em relação ao modelo original, descritas a seguir (ONS, 2017):

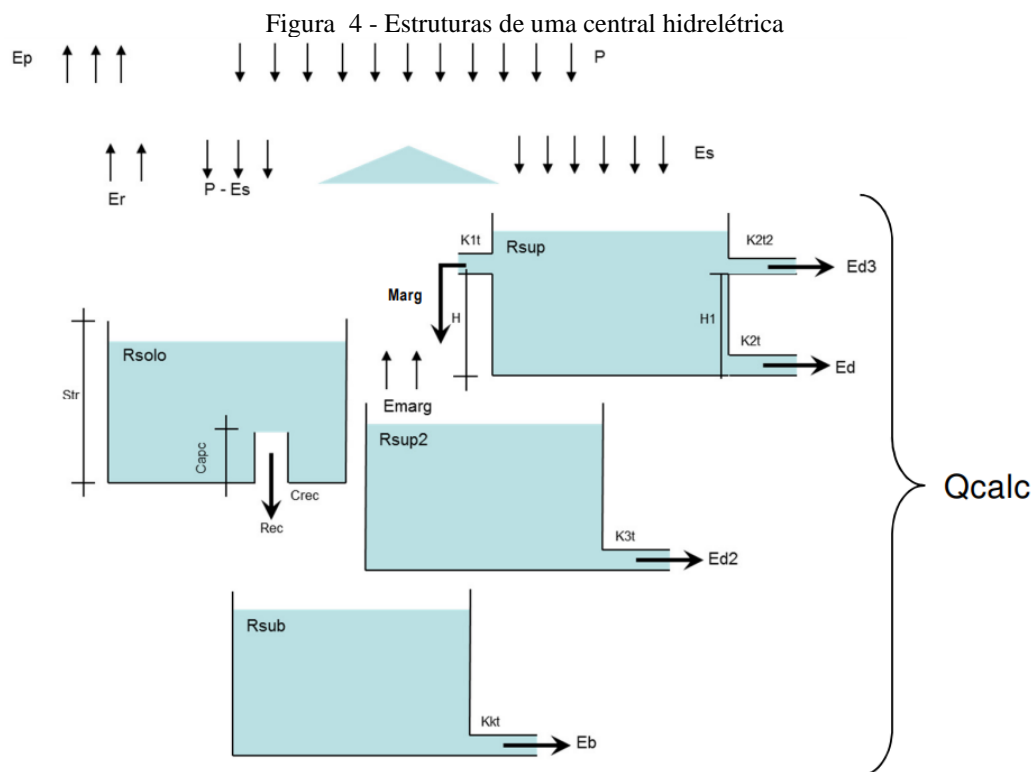
- Uso de coeficientes de representação temporal da precipitação;
- Uso de coeficiente de ajuste da evapotranspiração potencial;
- Criação de um quarto reservatório (reservatório de planície);
- Possibilidade de utilização de até dois coeficientes de recessão do reservatório de escoamento superficial.

Deste ponto do texto em diante, ao nos referirmos ao modelo SMAP, nos referiremos ao modelo SMAP com adaptações do ONS.

O SMAP é um modelo hidrológico do tipo chuva-vazão, pontual, conceitual, determinístico e dinâmico. O SMAP também pode ser definido da seguinte maneira:

O modelo SMAP baseia-se na divisão da vazão em escoamento superficial e em escoamento subterrâneo, uma vez que eles possuem características distintas. Essa divisão é feita utilizando a equação de *runoff (curve number) do Soil Conservation Service (SCS)* do Estados Unidos, sendo cada parcela armazenada em reservatórios fictícios que são responsáveis por descrever a forma e a duração dos escoamentos (ONS, 2017).

A simplificação hidrológica conceitual do modelo está na utilização de reservatórios, que podem ser observados na Figura 4.



Fonte: ONS (2017).

Como apresentado na Figura 4, o modelo SMAP possui 4 reservatórios, sendo eles:

- **Rsup:** Reservatório de superfície (representa o escoamento superficial);
- **Rsup2:** Reservatório de superfície/planície (representa o armazenamento de água que extravasa a calha natural do rio);
- **Rsolo:** Reservatório do solo (representa o armazenamento de água que infiltra no solo);



- **Rsub**: Reservatório subterrâneo (representa o armazenamento de água em aquíferos).

Outras siglas de processos representados nas Figura 4 são classificadas e descritas abaixo:

Dados de entrada:

- P: Precipitação média;
- Ep: Evapotranspiração potencial;

Parâmetros fixos do modelo:

- Capc: Capacidade de campo do solo;
- Crec: Coeficiente de recarga subterrânea;
- Ep: Evapotranspiração inicial;
- H: altura representativa para transbordamento para planícies;
- H1: altura representativa para início do segundo escoamento superficial;
- Str: altura representativa da capacidade de saturação do solo;
- K1t: constante de recessão do escoamento para planícies;
- K2t: constante de recessão do primeiro escoamento superficial;
- K2t2: constante de recessão do segundo escoamento superficial;
- K3t: constante de recessão do escoamento da superfície/planícies;
- Kkt: constante de recessão do escoamento básico.

Processos calculados a cada passo de tempo:

- Er: Evapotranspiração real do solo;
- Es: Escoamento para o reservatório de superfície;
- P-Es: Parcela da chuva que infiltra no solo;
- Marg: Extravasamento pelas margens;
- Emarg: Evaporação da planície de inundação;
- Rec: Recarga subterrânea;
- Ed: primeiro escoamento superficial;
- Ed2 : segundo escoamento superficial;
- Ed3: escoamento superficial;
- Eb: escoamento de base.

Variáveis de saída:

- Qcalc: vazão total calculada pelo modelo.

Cada dado, de ambos os tipos de parâmetros descritos acima, é obtido a partir de calibrações, de inicialização do modelo, ou diárias, antes de cada rodada do modelo.

Os parâmetros fixos são obtidos através de calibrações de configuração do modelo, e parte dos parâmetros calculados a cada passo de tempo são calibrados a cada rodada diária do modelo. O processo de calibração, também chamado de assimilação de dados, consiste em comparar os dados previstos pelo modelo (inicializado com um conjunto de dados qualquer) com os dados observados, e ajustando i.e (otimizando), os parâmetros até que os melhores valores sejam obtidos (DAVIE, 2008).

A classificação entre variáveis de entrada, parâmetros fixos do modelo, processos calculados a cada passo de tempo e variáveis de saída é importante pois de acordo com Davie (2000), em modelos conceituais, não é possível considerar que parâmetros otimizados do modelo possam ter qualquer significado físico, visto que em alguns casos é possível obter hidrogramas similares com um quadro completamente diferente de parâmetros otimizados. Portanto, o modelo SMAP não deve ser utilizado para prever escoamentos particionados, como o escoamento de base ou escoamento superficial, mas sim a vazão total da bacia, classificada como dado de saída.

### **2.3.1 VARIÁVEIS DE ENTRADA DO SMAP**

Os dados de entrada do modelo SMAP são os totais diários de precipitação, previstos e observados, os totais diários de evapotranspiração potencial e as vazões médias diárias observadas do período de assimilação de dados. O processo de assimilação de dados do SMAP consiste no recebimento de dados de tempos anteriores ao da previsão para estimar condições iniciais da bacia, e se necessário, corrigir variáveis com o intuito de aproximar a condição inicial de vazão, da vazão observada (ONS, 2017). Para realizar essa correção, o ONS utiliza um algoritmo de otimização heurística bioinspirada na ecolocalização de morcegos (Bat Algorithm) (ONS, 2017; YANG; HOSSEIN GANDOMI, 2012). Tal algoritmo corrige o escoamento de base (Ebin), escoamento superficial (Supin) e precipitação observada ( $P_{(t)}$ ). Aqui, abordaremos os dados de precipitação, que estão no contexto do presente trabalho.

### 2.3.1.1 Precipitação observada

A partir de 2021, o ONS passou a utilizar uma metodologia de combinação entre dados pluviométricos e dados de precipitação estimados por satélite, descrita em ONS (2020b). Os dados de satélite utilizados são fornecidos pelo algoritmo denominado Integrated Multisatellite Retrievals for Global Precipitation Mission (IMERG). Tal produto possui três versões, com diferentes latências de dados: EARLY, LATE e RES. O ONS utiliza a versão EARLY para a metodologia de combinação de dados de precipitação, que possui latência de aproximadamente 4 horas.

A metodologia consiste, brevemente, em retirar quaisquer pontos de grade do produto IMERG que estejam dentro de um raio de 20 km de uma estação pluviométrica, sendo esta lacuna gerada nos pontos de grade é denominada ‘área de influência da estação pluviométrica’. As coordenadas retiradas são então preenchidas através da interpolação, pelo inverso do quadrado da distância, entre: dado da estação pluviométrica e dados estimados por satélite (fora do raio de influência da estação pluviométrica).

Após o processo de interpolação, obtém-se o produto da combinação, sobre o qual serão calculadas as precipitações médias de cada bacia do SIN.

### 2.3.1.2 Precipitação prevista

A precipitação prevista de entrada do SMAP é a média da previsão de precipitação por conjunto na sub-bacia modelada. Esta previsão é montada a partir de 3 modelos de previsão de precipitação: *ECMWF-ENS*, ETA (CPTEC) e GEFS (NOAA).

A previsão *ECMWF-ENS* consiste na média de precipitação dos 51 membros da previsão por conjuntos do ECMWF. Uma das desvantagens do uso da média das previsões de precipitação é o fato de que elas não representam um estado único da atmosfera de forma genuína e dinamicamente consistente, além da menor capacidade de representar eventos extremos (ECMWF, 2022b), visto que cenários de eventos extremos são atenuados pelo cálculo da média.

Cada uma dessas previsões passa, individualmente, por um processo de remoção de viés, e posteriormente, são combinadas através de pesos para cada modelo que compõe a previsão, de forma a minimizar o erro médio de um agrupamento de dias e o erro absoluto de cada dia individualmente (ONS, 2020c). Esta metodologia é aplicada para as primeiras duas semanas previstas.

## 2.4 CÁLCULO DA ENERGIA NATURAL AFLUENTE

Nesse capítulo são apresentados alguns conceitos necessários para entender o processo de cálculo da Energia Natural Afluente (ENA), dado de entrada principal da análise de agrupamento que é utilizada neste trabalho.

### 2.4.1 Geração de energia elétrica

A geração de energia elétrica é realizada através da transformação de qualquer tipo de energia em energia elétrica. Em energias provenientes da água, este processo é dividido em duas partes (VASCONSELOS, 2017): transformação de energia potencial hidráulica em energia cinética de rotação, através de turbinas, e transformação da energia cinética em energia elétrica, através de um gerador elétrico.

A classificação de uma usina hidrelétrica é dependente da potência instalada nela (ANEEL, 2022):

- Centrais Hidrelétricas de Geração (CGH): até 1 MW de potência instalada;
- Pequenas Centrais Hidrelétricas (PCH): entre 1,1 MW e 30 MW de potência instalada;
- Usina Hidrelétrica de Energia (UHE): mais de 30 MW de potência instalada.

Vale ressaltar que todas as usinas hidrelétricas que compõem o SIN são classificadas como UHE.

De acordo com VASCONSELOS (2017), uma central hidrelétrica pode ser definida como uma estrutura capaz de produzir energia por meio do acionamento de um conjunto entre turbinas hidráulicas e água. Podemos classificar os componentes de uma hidrelétrica em três categorias principais (OLIVEIRA, 2014):

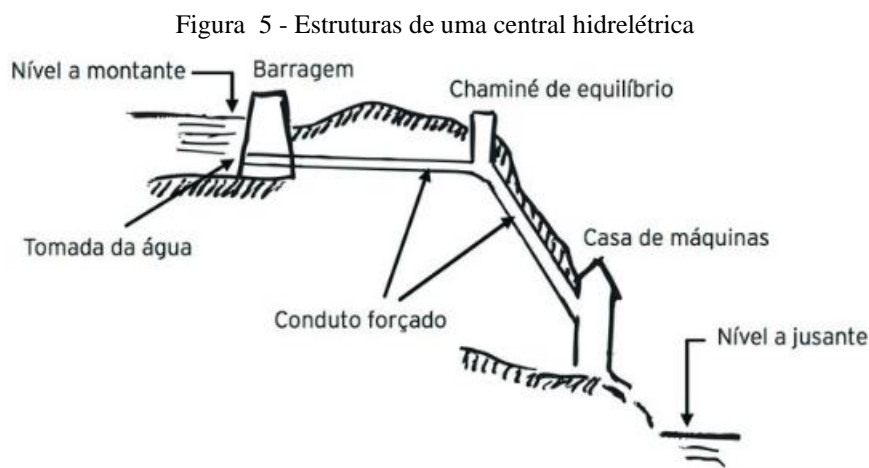
- Estruturas hidráulicas;
- Turbinas;
- Equipamentos elétricos.

#### 2.4.1.1 Estruturas hidráulicas

Dentre as estruturas hidráulicas, estão a barragem, vertedouro, reservatório e comportas d'água. A Figura 5 demonstra parte dos componentes de uma central hidrelétrica.

A barragem é uma estrutura de concreto ou pedra, responsável por armazenar água dentro do reservatório. O tipo, arranjo e quantidade de barragens, dependerão da topografia do local de instalação e das condições de fundação (OLIVEIRA, 2014). O vertedouro é uma estrutura que pode descarregar o excesso de água do reservatório, servindo como ferramenta de

gestão hídrica em conjunto com o reservatório. O reservatório, por sua vez, é o lago artificial da hidrelétrica, formado pelo acúmulo de água gerado pela construção de barragens em locais estratégicos da bacia, considerando as curvas de nível do terreno. Para provocar a queda d'água necessária para a movimentação das turbinas, o reservatório é construído em uma cota superior às mesmas (OLIVEIRA, 2014). As comportas são estruturas que permitem o fluxo de água entre o reservatório e turbinas.



Fonte: VASCONSELOS (2017)

#### 2.4.1.2 Modelagem da Função de produção das hidrelétricas

Para entender a capacidade de geração energética de uma UHE, dada determinada vazão, é necessário entender o conceito da Função de Produção Hidrelétrica (FPH).

As turbinas são responsáveis por transformar a energia proveniente da queda d'água em energia mecânica. A Equação 1 demonstra a potência de uma turbina, que é dependente da densidade d'água ( $\rho$ ), vazão turbinada ( $Q$ ), altura da queda d'água ( $H$ ), aceleração da gravidade ( $g$ ) e rendimento do sistema ( $\eta$ ).

$$P = \rho * Q * H * g * \eta \quad (1)$$

De acordo com FREDO (2016), a Função de Produção Hidrelétrica (FPH) de uma unidade geradora é uma expressão matemática que representa o comportamento da potência de saída da unidade em função de variáveis de estado e controle que compõem o processo de geração de hidroeletricidade. Tal expressão depende de maneira não linear do produto entre queda líquida, vazão turbinada e rendimento global da unidade geradora (FREDO, 2016). A

potência de saída de uma unidade hidrelétrica é dada pela Equação 2. Ressalta-se aqui que a unidade geradora não é o mesmo que a UHE, sendo, por exemplo, apenas um dos conjuntos turbina-gerador que compõem uma UHE.

$$p = G * \eta(u, hl) * hl * u \quad (2)$$

Onde:

G: constante dada por 0,00981 (MW/(m<sup>3</sup>/s).m);

p: potência ativa da unidade geradora (MW);

u: vazão turbinada na unidade geradora (m<sup>3</sup>/s);

hl: queda líquida da unidade (m);

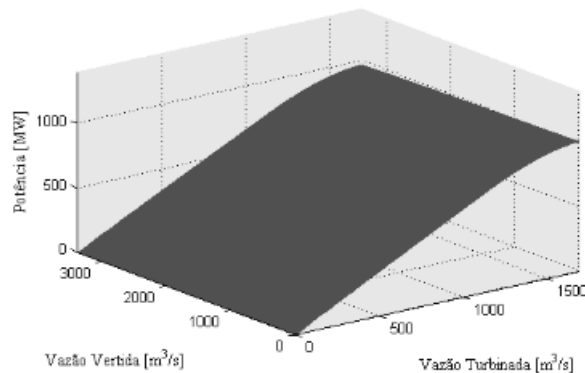
$\eta(u,hl)$ : rendimento hidráulico da unidade geradora.

Na gestão do SIN, um dos principais problemas que envolvem a gestão do sistema é o Planejamento de Médio Prazo da Operação Hidrotérmica (PMOP), o qual é influenciado pela capacidade de geração das usinas hidrelétricas. Para a resolução do PMOP, pode ser utilizada a Programação Dinâmica Dual Estocástica (PDDE) (MATOS, 2012). Entretanto, para que tal algoritmo possa atingir a convergência, é necessário que todas as equações envolvidas no PMOP formem um conjunto convexo (FREDO, 2016). Na Figura 6 é apresentada a FPH de Furnas.

Ao analisarmos a Figura 6, podemos fazer uma interpretação gráfica acerca da definição de uma função convexa: para quaisquer conjunto de pontos  $(a,b,c)$  pertencentes ao do domínio da função, que geram as imagens  $f(a)$ ,  $f(b)$  e  $f(c)$ , o plano que une estes três pontos deve estar sempre ‘acima’ de qualquer outra imagem da função, ou seja, ao criarmos um plano a partir de três pontos da função, este plano não pode interceder com a imagem da função em qualquer momento. Portanto, fica evidente que a FPH apresentada na Figura 6 não é convexa. Na verdade, de acordo com FREDO (2016), a definição da FPH é de uma função não convexa, portanto, para resolver o problema da PMOP através da PDDE, se faz necessário aproximar a FPH por meio de um modelo que mantenha a convexidade do problema.

Uma das maneiras mais simples de se realizar tal aproximação é através da consideração de produtividade constante, onde a FPH passa a depender linearmente, apenas, da vazão turbinada.

Figura 6 - FPH de Furnas

**Figura 2.13 – FPH de Furnas com  $v = 5733 \text{ hm}^3$ .**

Fonte: (FREDO, 2016)

A aproximação mais simples para a representação linear da FPH de uma UHE é aquela denominada de produtividade constante, em que se calculam coeficiente que representa a taxa de variação de potência (MW) por vazão turbinada total ( $\text{m}^3/\text{s}$ ). Neste tipo de modelagem, a queda líquida e o rendimento são considerados constantes. O produto entre queda, rendimento e a constante  $G$  apresentada no capítulo anterior, definem a produtibilidade da UHE (FREDO, 2016).

A Equação 2 pode ser generalizada para uma UHE, obtendo-se a FPH de uma UHE, apresentada na Equação 3.

$$p = G * \eta(u, hl) * hl * q \quad (3)$$

Onde:

$G$ : constante dada por  $0,00981 \text{ (MW}/(\text{m}^3/\text{s}).\text{m})$ ;

$p$ : potência ativa da UHE(MW);

$q$ : vazão turbinada pela UHE ( $\text{m}^3/\text{s}$ );

$hl$ : queda líquida equivalente da UHE (m);

$\eta(u,hl)$ : rendimento da UHE.

Da Equação 3 é possível retirar um coeficiente constante para cada UHE, denominado produtibilidade, que engloba  $G$ ,  $hl$  e  $\eta(u,hl)$ . Tal coeficiente é apresentado na Equação 4.

$$\rho = G * H * (HM - HJ - HP) \quad (4)$$

Onde:

$\rho$ : produtibilidade da UHE [MW/(m<sup>3</sup>/s)];

G: constante dada por 0,00981 (MW/(m<sup>3</sup>/s).m);

H: valor fixo de rendimento equivalente da UHE;

HM: valor fixo de cota montante equivalente (m);

HJ: valor fixo de cota jusante equivalente (m);

HP: valor fixo de perda hidráulica equivalente na UHE (m);

FREDO (2016) concluí que existem diversas maneiras para se definir a produtibilidade, visto que diferentes pontos operativos da UHE podem ser escolhidos. No que tange a operação e gestão do SIN, de acordo com ONS (2011), em relatório que define os critérios para estudo, cálculos de Energia Natural Afluyente (ENA) são realizados a partir de produtibilidades equivalentes ao armazenamento de 65% do volume útil (VU) dos reservatórios dos aproveitamentos hidrelétricos. Ou seja, os parâmetros H, HM, HJ e HP são calculados a partir do ponto de operação do reservatório que equivale à 65% de seu VU.

A produtibilidade, além de ser parte essencial do cálculo de ENA, também é utilizada nesse trabalho durante o processo de cálculo de Energias Naturais Afluentes observadas, a partir de vazões observadas consistidas.

#### **2.4.2 Vazão Natural Afluyente (VNA)**

De acordo com ONS (2018), a implantação de reservatórios modifica a distribuição espacial e temporal das vazões, regularizando a mesma de forma a atender uma capacidade de geração de energia, e outros usos consuntivos. Portanto, o regime de vazões em um determinado ponto à jusante de um reservatório não é necessariamente aquele que se apresentaria caso a bacia não possuísse alterações antrópicas. Além das regularizações, a existência de um reservatório também impõe outras modificações no balanço hídrico da bacia, como a aumento da evaporação em superfícies líquidas, que deixa de atuar somente sobre a calha do rio, mas sob toda a área do reservatório.

No setor elétrico, há a adoção de dois termos para se referir à vazão em uma bacia: vazão natural e vazão afluyente. A vazão natural é definida como a vazão que ocorreria em uma seção de rio caso não houvesse alterações antrópicas em sua bacia contribuinte (ONS, 2018), enquanto a vazão afluyente é definida como a vazão que chega à um aproveitamento hidrelétrico sob influência das obras antrópicas, a montante do aproveitamento.



A vazão natural afluente (VNA) consiste na união dos dois conceitos acima, sendo a vazão que chega à um aproveitamento hidrelétrico caso não houvesse obras antrópicas na bacia. A importância de tal reconstituição está no fato de que funções de planejamento e operação do SIN, como por exemplo, as políticas de operação obtidas na otimização do modelo DECOMP, utilizam séries de vazões naturais como dados de entrada (ONS, 2018).

Dentre as principais ações antrópicas que afetam o regime fluvial, estão: desmatamento, implantação e operação de reservatórios, transposição de vazões e captação para usos consuntivos. Para eliminar tais efeitos, as vazões naturais são calculadas a partir de séries de vazões observadas em estações fluviométricas, vazões de afluência e defluência nos aproveitamentos hidrelétricos e incorporando vazões relativas à usos consuntivos e à evaporação líquida (ONS, 2018).

#### 2.4.3 Energia Natural Afluente (ENA)

A ENA pode ser definida como uma estimativa da energia que uma bacia do SIN teria dada uma certa afluência natural à ela.

Em geral, a Energia Natural Afluente (ENA) é obtida através da multiplicação da produtividade, equivalente à 65% do VU do reservatório, pela VNA do aproveitamento hidrelétrico (SILVA, 2017). O cálculo da ENA pode ser feito com diferentes agregações temporais (diária, mensal ou anual) ou espaciais (bacia ou subsistema). As Equações 5 e 6 representam, respectivamente, a ENA por bacia e ENA por subsistema.

$$ENA_{bacia}(t) = \sum_{i=1}^n [Qnat(i, t) * \rho(i)] \quad (5)$$

$$ENA_{subsistema}(t) = \sum_{i=1}^m [Qnat(j, t) * \rho(j)] \quad (6)$$

Onde:

Qnat: vazão natural do aproveitamento no intervalo de tempo considerado (m<sup>3</sup>/s);

ENAbacia: energia natural afluente no intervalo de tempo considerado (MWmed);

ENAsubsistema: energia natural afluente no intervalo de tempo considerado (MWmed);

n: número de aproveitamentos existentes na bacia considerada;

m: número de aproveitamentos existentes no subsistema considerado.

## 2.5 CLIMA E TEMPO NO BRASIL

A superfície da terra se diferencia significativamente de um lugar para outro por conta da combinação resultante de múltiplos agentes naturais, como: estrutura geológica, relevo, clima, solo, rios, vegetação, fauna etc. (TORRES; MACHADO, 2008). No estudo da climatologia, se faz necessário uma distinção entre tempo e clima, visto que o tempo é o estado momentâneo da atmosfera em determinado estado de tempo, enquanto o clima pode ser definido como a sucessão ou conjunto de variações desses estados médios e que caracteriza a atmosfera de certo lugar (TORRES; MACHADO, 2008). De acordo com AYOADE (1996), a determinação do clima de uma região não pode ser feita com menos de 30/35 anos de dados.

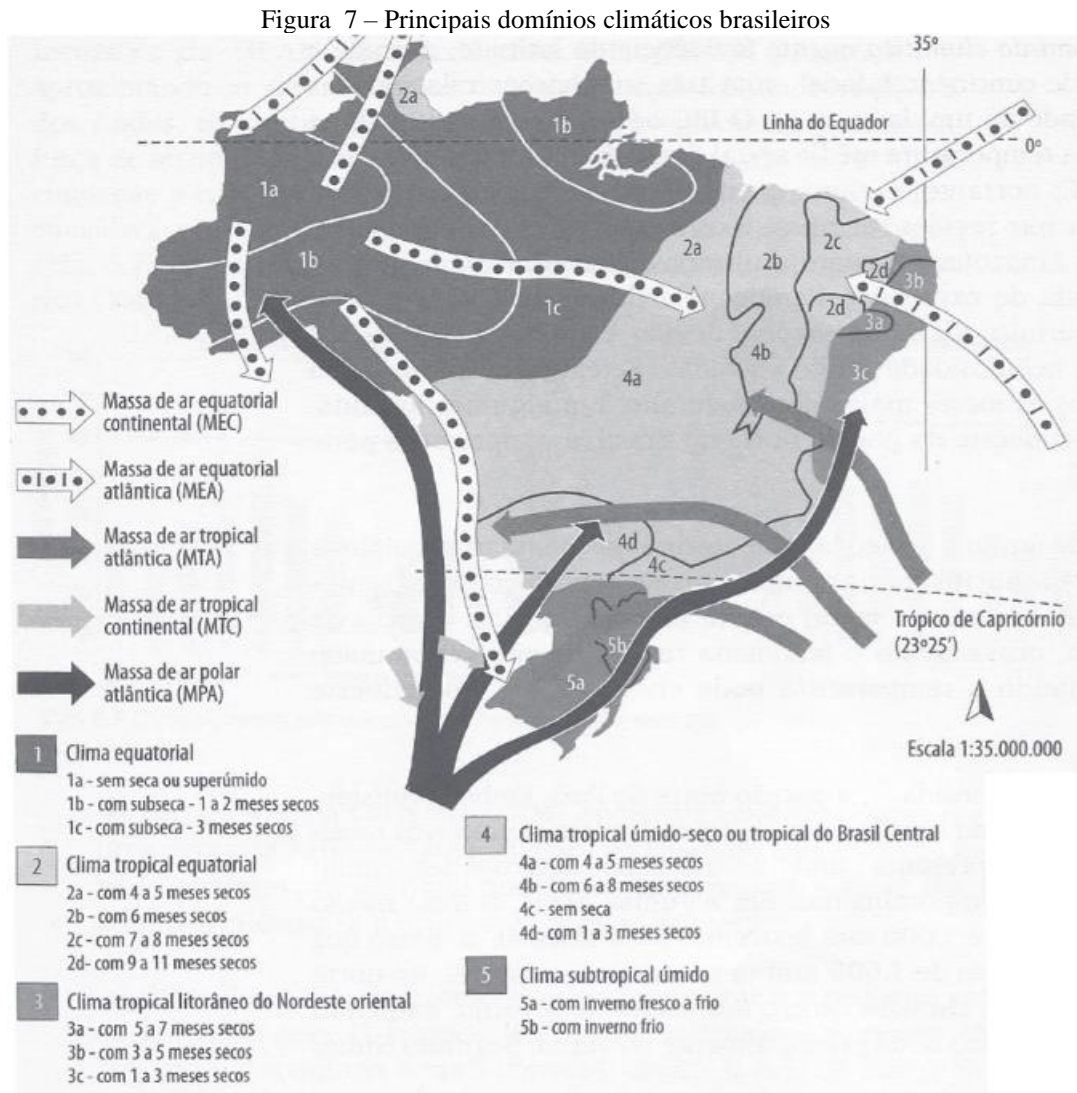
Elementos do clima (temperatura, umidade do ar, pressão atmosférica, ventos, radiação solar etc.) são seus componentes principais do clima, que se relacionam entre si para formar o tempo atmosférico e o clima em si. Já fatores do clima (latitude, longitude, maritimidade, solos, correntes marítimas etc.) trazem alterações significativas nos elementos climáticos (TORRES; MACHADO, 2008).

A classificação climática visa ordenar metodologicamente dados pesquisados, proporcionando uma melhor compreensão das manifestações climáticas existentes na superfície terrestre (ZANATTA et al., 2016). Zanatta et al. (2016), afirma que de modo geral os elementos climáticos mais utilizados para a elaboração dessas classificações são a temperatura e precipitação pluvial, porém, além dos valores absolutos, algumas classificações consideram também a variação dos elementos climáticos ao longo do ano e mudanças ocorridas ao longo de vários anos.

### 2.5.1 Clima

A grande variação de tipos climáticos que existe no Brasil é influenciada por diversos fatores, como: formato triangular da América do Sul, favorecendo maior influência das águas oceânicas no Hemisfério Sul; grande extensão territorial, que proporciona faixas banhadas por águas quentes na região próxima do Equador e regiões banhadas por águas frias em direção ao Sul; influência da maritimidade e continentalidade; influencia de diferentes massas de ar, ora de caráter equatorial, ora tropical, ora atlântica (ZANATTA et al., 2016).

Mendonça e Danni-Oliveira (2007) propõem a existência de cinco grandes comportamentos climáticos no país, baseando-se na distribuição da temperatura e pluviosidade, associadas às características geográficas e à dinâmica das massas de ar. A Figura 7 apresenta os 5 grandes domínios climáticos propostos.



Fonte: MENDONÇA; FRANCISCO; OLIVEIRA-DANNI. INÊS MORESCO (2007)

O Clima Equatorial abrange os estados do Amazonas, Pará, Acre, Rondônia, Amapá, parte do Mato Grosso e Tocantins, sendo caracterizado como um clima quente, com temperatura média anual de 24° a 26°C, além de úmido, devido à grande quantidade de rios e à densa floresta úmida presente. Frentes frias vindas do continente podem fazer a temperatura cair para 8°C na parte sudoeste, durante os meses junho a agosto (ZANATTA et al., 2016).

A distribuição da precipitação não é homogênea, tanto no quesito espacial quanto temporal. Diante dessas características, esse clima ainda pode ser subdividido em: Clima Equatorial superúmido; Clima Equatorial com subseca e Clima Equatorial seco (ZANATTA et al., 2016).

O Clima Tropical-Equatorial está ligado a mudança de vegetação entre Floresta Amazônica e Caatinga, sendo um clima de transição. Tal clima apresenta grande variabilidade pluviométrica, com regiões que registram médias anuais de 2300mm, próximas a São Luís (MA), e locais de índice baixos, como Campos Sales (CE), onde a média anual não alcança 620mm. De acordo com as características pluviométricas, este pode ser subdividido em: Clima Tropical-Equatorial com quatro a cinco meses secos; Clima Tropical-Equatorial com seis meses secos; Clima Tropical-Equatorial com seis a oito meses secos e Clima Tropical-Equatorial com nove a onze meses secos (ZANATTA et al., 2016).

O Clima Tropical Litorâneo do Nordeste Oriental trata-se de um clima quente, com influência da maritimidade e das massas de ar vindas do atlântico. A pluviosidade chega à média anual de 2500mm na região leste, porém é reduzida à oeste, próxima ao sertão, onde não atinge 700mm. As características térmicas e pluviométricas permitem subdividi-lo em: clima tropical litorâneo do nordeste oriental com um a três meses secos; clima tropical litorâneo do nordeste oriental com três a cinco meses secos; e clima tropical litorâneo do nordeste oriental com cinco a sete meses secos (ZANATTA et al., 2016).

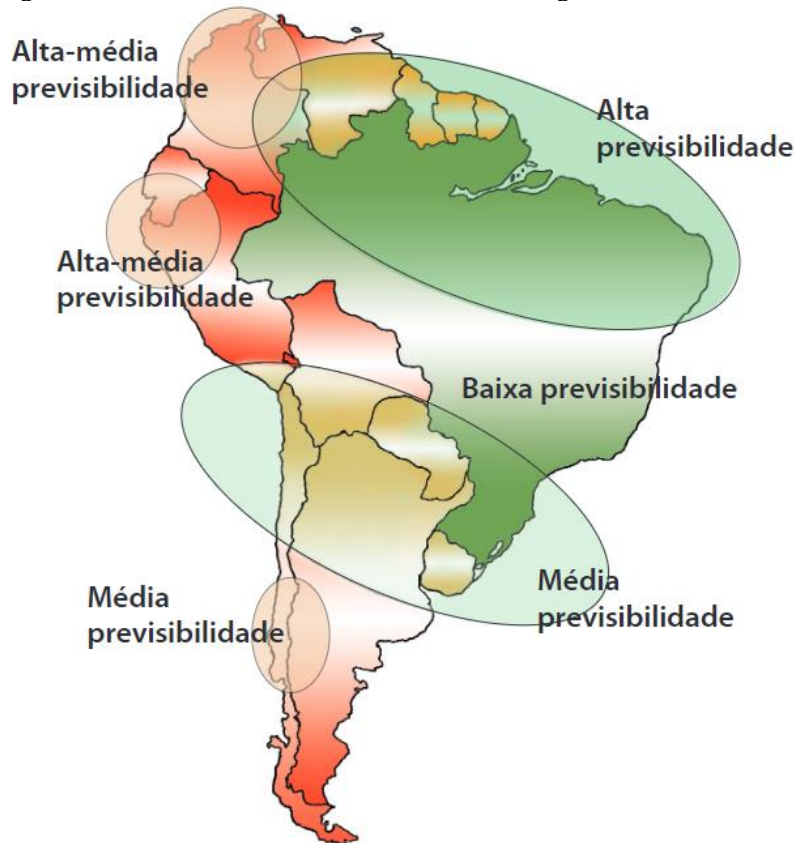
O Clima Tropical Úmido Seco está relacionado ao Centro-Oeste brasileiro, sofrendo influência das massas de ar equatoriais e tropicais, com alta variabilidade pluviométrica e térmica ao longo do ano. As médias pluviométricas anuais ficam entre 2000 e 3000mm, entretanto, as chuvas concentram-se no verão. Uma das principais características marcantes desse clima é a alta pluviosidade associada à elevadas temperaturas. Estas características pluviométricas são resultado da interação entre massas de ar e relevo (ZANATTA et al., 2016).

O Clima Subtropical Úmido é predominante na região Sul do país, apresentando chuva bem distribuída ao longo do ano e médias térmicas mais baixas em relação ao resto do país. Além disso é influenciado pelas massas de ar tropical e polar, vindas do Atlântico Sul. As médias pluviométricas anuais variam entre 1200 e 2000mm (ZANATTA et al., 2016).

ROWELL (1998) afirma que na região tropical há maior previsibilidade climática sobre os oceanos, principalmente o Pacífico e Atlântico. Nos trópicos a forçante de estado do oceano domina a variabilidade interanual, tornando a previsibilidade mais alta que nos extra trópicos, onde a variabilidade associada a não linearidades do escoamento atmosférico compete

com os efeitos de variações lentas do estado do oceano (CHARNEY; SHUKLA, 1981). Especificamente para o Brasil, no Nordeste, temos uma das previsões de chuva com maior previsibilidade em regiões continentais de todo o globo, devido à sua grande dependência com o estado dos oceanos Atlântico e Pacífico tropicais (SAMPAIO; DIAS, 2014). Na Figura 8 podemos observar a previsibilidade climática para regiões da América do Sul.

Figura 8 – Previsibilidade climática de diferentes regiões da América do Sul



Fonte: (SAMPAIO; DIAS, 2014)

As regiões Sudeste e Centro-Oeste apresentam as menores confiabilidades devido a menor dependência dos oceanos e aos sistemas meteorológicos que atuam nela. No período chuvoso a ocorrência de chuva dessas regiões é controlada pelas condições de umidade, instabilidade do ar e por frentes frias. No período de inverno estas frentes frias podem gerar quedas significativas de temperatura, porém, em geral, não geram chuvas abundantes. Portanto, a previsão para essas regiões torna-se mais difícil pela necessidade de precisão no deslocamento, número e intensidade dessas frentes (SAMPAIO; DIAS, 2014).

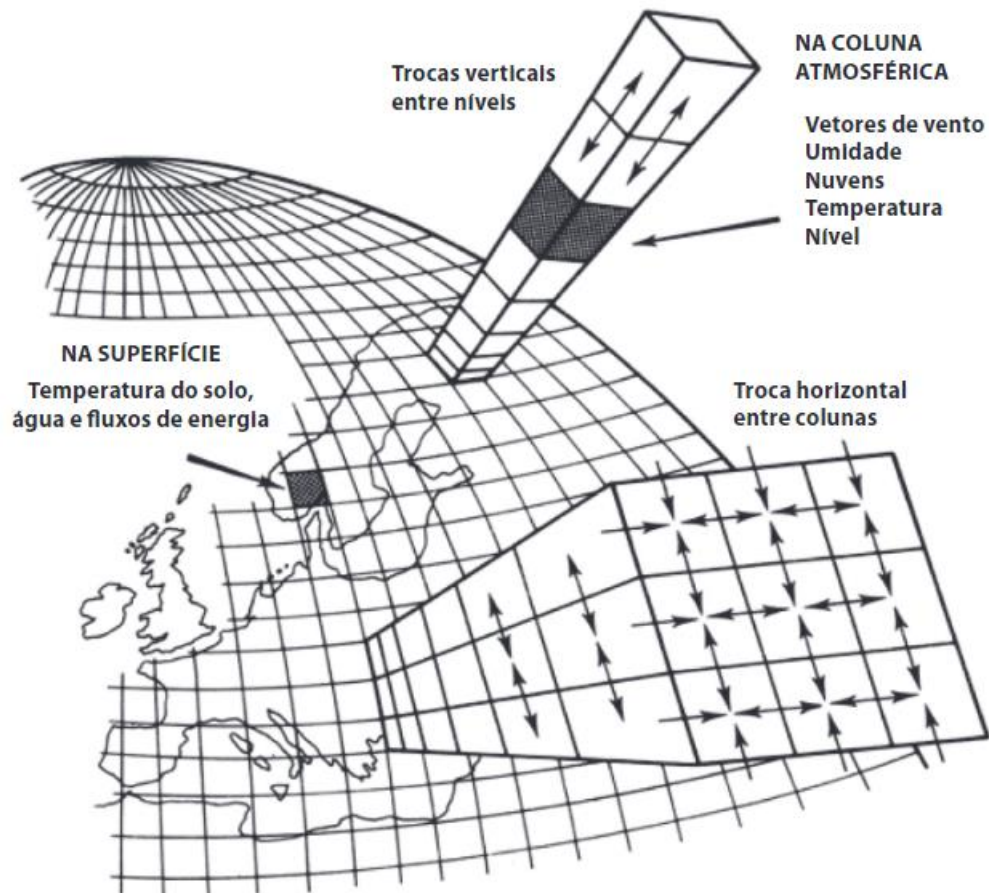
### 2.5.2 Tempo

Dentre os diversos avanços científicos que ocorreram no último século está a habilidade em simular complexos sistemas físicos utilizando modelos numéricos que constituem uma representação matemática aproximada da realidade. Nas últimas décadas, esse avanço também ocorreu no entendimento dos processos que controlam o estado da atmosfera, dos componentes do sistema climático, dos métodos de solução numérica, das equações que representam cada parte do sistema e do poder de processamento dos computadores (SAMPAIO; DIAS, 2014).

Os Modelos de Circulação Geral da Atmosfera (MGCAs) possibilitam prever as condições do tempo para vários dias (SAMPAIO; DIAS, 2014). Esses modelos são códigos computacionais que representam aproximações numéricas de equações matemáticas representativas das leis físicas que regem o movimento da atmosfera, assim como sua interação com outros componentes climáticos. Tais modelos possibilitam prever as condições do tempo para vários dias (SAMPAIO; DIAS, 2014).

As equações deste modelo podem ser discretizadas em quatro dimensões: tempo, latitude, longitude e altitude (Figura 9).

Figura 9 – Dimensões de um MCGA



Fonte: SAMPAIO; DIAS (2014).

Exemplos de variáveis previstas através de MGCAs são: temperatura, umidade, precipitação, vento, pressão atmosférica etc. Os modelos são alimentados com informações meteorológicas vindas de toda a Terra, através de redes integradas de comunicação, como por exemplo a *Global Telecommunication System* (GTS). Com a evolução da tecnologia e precisão de satélites, a partir da década de 70, também passaram a ser incorporadas informações de satélite coletadas. No século 21, ao juntar os dados de satélite com a técnica de assimilação de dados, houve um avanço ainda maior na confiabilidade nesses modelos (SAMPAIO; DIAS, 2014).

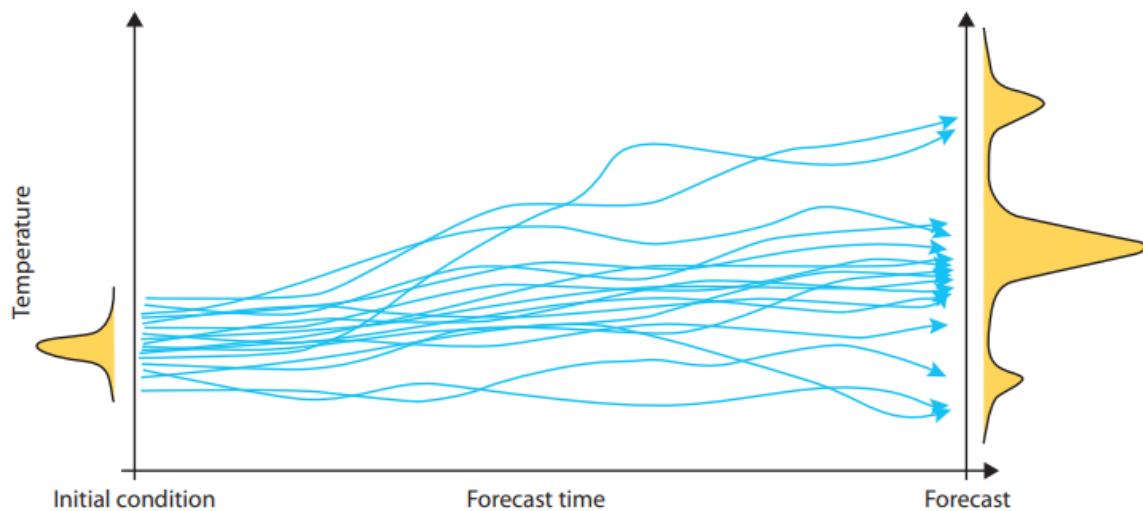
Segundo LORENZ (1965), matemático e meteorologista que estabeleceu a base teórica de previsão climática e do tempo, a previsibilidade atmosférica possui um limite teórico de duas semanas. Portanto, na prática, os mais avançados serviços meteorológicos não conseguem desenvolver previsões de tempo com taxas de acertos úteis para prazos superiores

a 7 dias, e somente em situações excepcionais a previsibilidade se estende à 15 dias (SAMPAIO; DIAS, 2014).

## 2.6 PREVISÕES POR CONJUNTO

Modelos numéricos de previsão do tempo, e atmosfera em si, podem ser vistos como sistemas dinâmicos não-lineares, nos quais a evolução depende fortemente das condições iniciais atribuídas ao modelo (LEUTBECHER; PALMER, 2008). Visto que o processo de determinação do estado da atmosfera é uma tarefa difícil e que carrega incertezas de todos os dados meteorológicos utilizados, a previsão por conjunto, também chamada de *Ensemble*, tem como objetivo representar as incertezas do processo através de perturbações nas condições iniciais do modelo. O modelo é rodado múltiplas vezes, partindo de diferentes condições atmosféricas iniciais, de modo a observarmos a dispersão da previsão ao longo do tempo (Figura 22), que está atrelada as diferentes condições atmosféricas iniciais adotadas. O ritmo de crescimento da dispersão ao longo da previsão pode ser usado como um indicativo de previsibilidade da mesma (ECMWF, 2022c).

Figura 10 – Representação de uma previsão por conjunto



Fonte: ECMWF (2022c).

O Centro Europeu de Previsões do Tempo de Médio Prazo (ECMWF) foi pioneiro na criação de um sistema para prever a confiabilidade de uma previsão. Este sistema, que opera desde 1992, é o *Ensemble Prediction System* (EPS) (ECMWF, 2022d). Em 2008 o EPS foi fundido com um sistema de previsão mensal e acoplado com um modelo oceânico dinâmico.

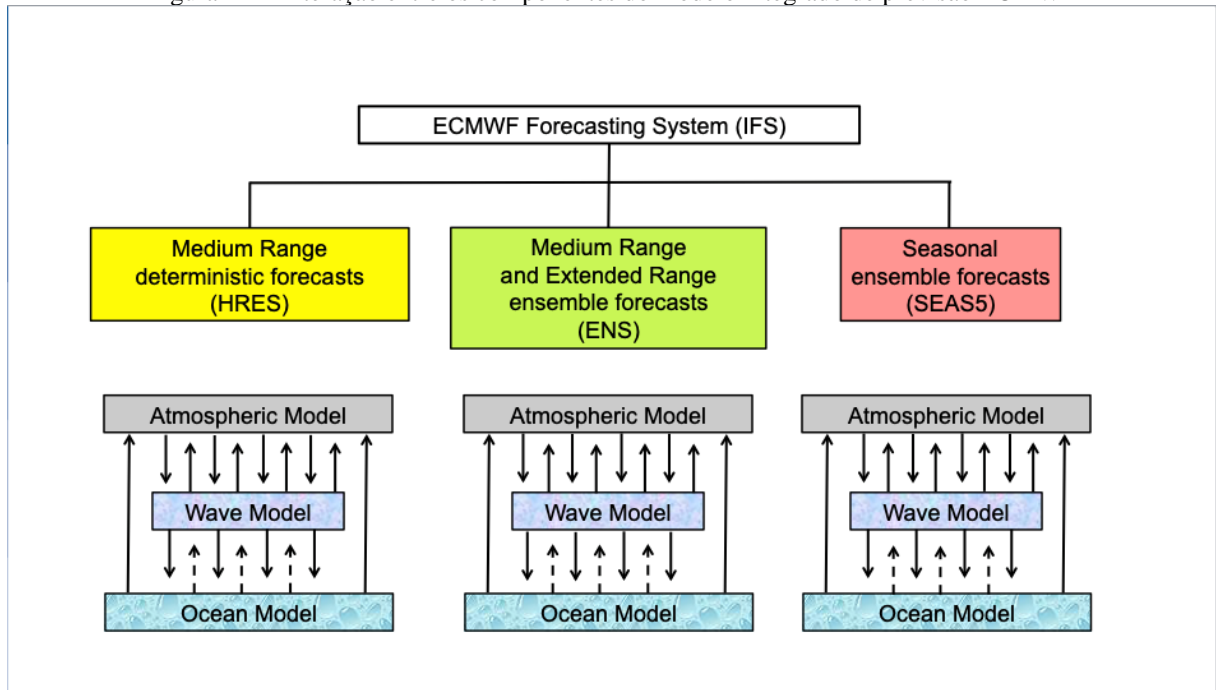


Em 2010 o modelo probabilístico EPS passou a ter 51 integrações com resolução espacial de aproximadamente 32 km (ECMWF, 2022d). Atualmente o modelo integrado de previsão do ECMWF consiste em vários componentes agrupados, sendo eles:

- Um MCGA de diferentes resoluções;
- *High Resolution* (HRES): Modelo com rodada única, duas vezes ao dia, com horizonte de 10 dias (ECMWF, 2022e);
- *Ensemble* (ENS): Modelo rodado duas vezes ao dia, com horizonte de 15 dias. É uma previsão conjunta gerada de outras 51 previsões com resolução espacial de 18 km (ECMWF, 2022f);
- *Extended-Range* (EXT): Nas segundas e quintas-feiras o modelo ENS tem seu horizonte de previsão estendido de 15 para 46 dias. Os dias estendidos tem resolução espacial de 36 km (ECMWF, 2022g);
- *Seasonal*: Modelo de previsão conjunta rodado uma vez por mês a partir de 51 previsões, com horizonte de 12 meses (ECMWF, 2022h);
- Um modelo de ondas oceânicas;
- Um modelo oceânico;
- Um modelo se superfície terrestre;
- Um sistema de análise de dados;
- Técnicas de perturbação para geração das previsões por conjunto.

Tais componentes interagem entre si em mais de uma maneira. A previsão de precipitação utilizada neste trabalho é o *Extended-Range* (EXT). A Figura 23 mostra a interação entre parte dos componentes. A motivação principal da previsão *Extended-Range* é investigar a previsibilidade no horizonte de 10 a 46 dias. Nesse horizonte, a interação entre os modelos atmosféricos e oceânicos é central para a previsão, visto que é um período curto o suficiente para a atmosfera ainda ter memória de suas condições iniciais, e longo o suficiente para que a variabilidade oceânica possa ter impacto na circulação atmosférica (ECMWF, 2022c).

Figura 11 - Interação entre os componentes do modelo integrado de previsão ECMWF



Fonte: ECMWF (2022c)

A saída do modelo *Extended-Range* provê uma visão mais generalista da previsão até o dia 46, focando principalmente nas diferenças semana a semana ao invés de tentar dar detalhes em dias individuais sobre o tempo (ECMWF, 2022i). Um dos principais objetivos deste modelo é a previsão com boa antecedência de padrões de grande escala irregulares e persistentes que podem levar a eventos extremos. Tais previsões não tem expectativa de capturar dia a dia, com precisão, variações no tempo, mas devem capturar padrões de circulação de larga escala que durem mais de uma semana, assim como indicar aproximadamente o tempo no qual os padrões de circulação sofrerão alterações (ECMWF, 2022c).

## 2.7 MÉTODOS DE AGRUPAMENTO

No contexto da utilização de previsões por conjunto em outros processos ou modelos, onde se deseja uma única previsão determinística ou a probabilidade de acontecimento de eventos semelhantes, os métodos de agrupamento se tornam uma maneira conveniente para a extração de informações dessas previsões.

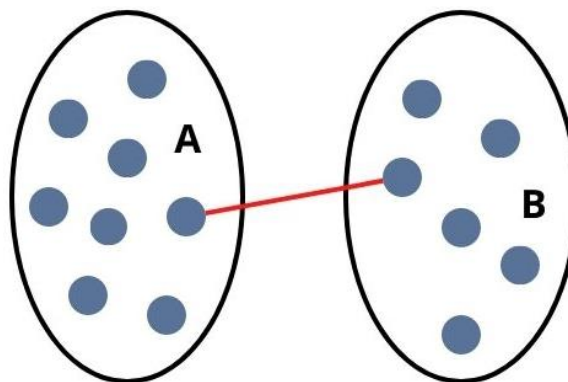
A classificação é um processo mental humano básico, onde agrupamentos relevantes podem prover uma economia de memória e poder preditivo. Com o avanço da tecnologia e do poder computacional, diversas áreas de pesquisa foram envolvidas no desenvolvimento de algoritmos automáticos ou objetivos de classificação (MILLIGAN; COOPER, 1987).

Não há definição exata do termo *Cluster*, e, portanto, diferentes abordagens são utilizadas ao tratar do assunto (SAXENA et al., 2017). As duas principais abordagens para análise de cluster são as técnicas: hierárquicas e de partição, sendo a definição de análise de cluster: classificação de dados em grupos significativos quando o número de grupos e outras informações sobre sua composição são desconhecidos (FRALEY; RAFTERY, 1998).

Os métodos hierárquicos produzem uma sequência de partições, cada uma com um número diferente de grupos, podendo ser classificados ainda em: aglomerativos (onde grupos são juntados a cada partição) ou divisivos (onde grupos são separados a cada partição). A cada estágio da análise de agrupamento hierárquica, a divisão ou aglomeração é feita de modo a otimizar determinado critério (FRALEY; RAFTERY, 1998). Este critério é usualmente denominado de distância utilizada pela análise de cluster, onde se define uma métrica que será utilizada para calcular distâncias entre diferentes grupos, além de definir a forma com que os grupos serão aglomerados ou divididos a cada estágio da análise. As distâncias mais comuns utilizadas dentro dos métodos hierárquicos são a single link (nearest neighbour), complete link (farthest neighbour) e soma dos erros quadráticos (*SSE*).

O método de Single Linkage determina a distância entre dois grupos através da menor distância entre qualquer elemento de um grupo para qualquer elemento de outro grupo (SAXENA et al., 2017). A Figura 12 ilustra a aplicação do método entre dois grupos num espaço plano.

Figura 12 - Método de agrupamento hierárquico por Single Linkage

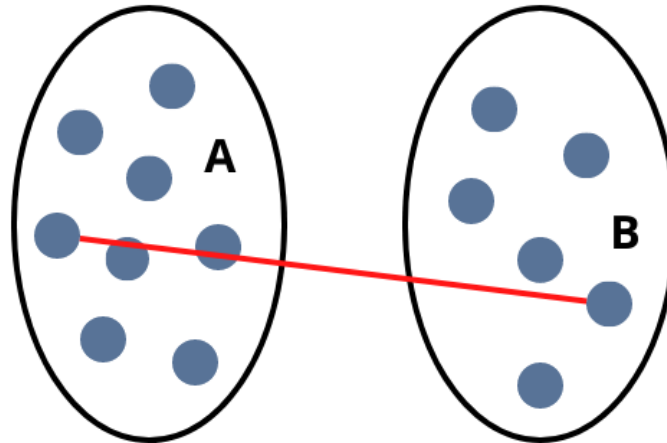


Fonte: Adaptado de SAXENA et al. (2017)

O método de Complete Linkage determina a distância entre dois grupos através da maior distância entre qualquer elemento de um grupo para qualquer elemento de outro grupo.

A Figura 13 ilustra a aplicação do método entre dois grupos num espaço plano (SAXENA et al., 2017).

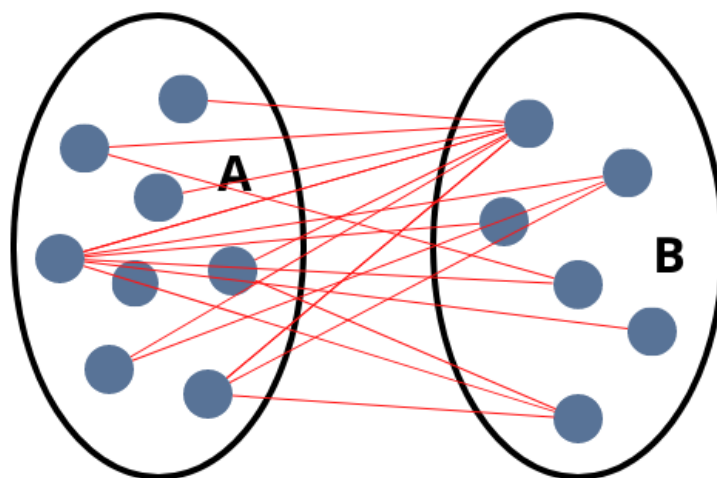
Figura 13 - Método de agrupamento hierárquico por Complete Linkage



Fonte: Adaptado de SAXENA et al. (2017)

O método Average Linkage, também conhecido como método da variância mínima, determina a distância entre dois grupos através da distância média existente entre qualquer membro de um grupo para qualquer membro de outro grupo (SAXENA et al., 2017). A Figura 14 ilustra a aplicação do método num espaço plano.

Figura 14 - Método de agrupamento hierárquico por Average Linkage

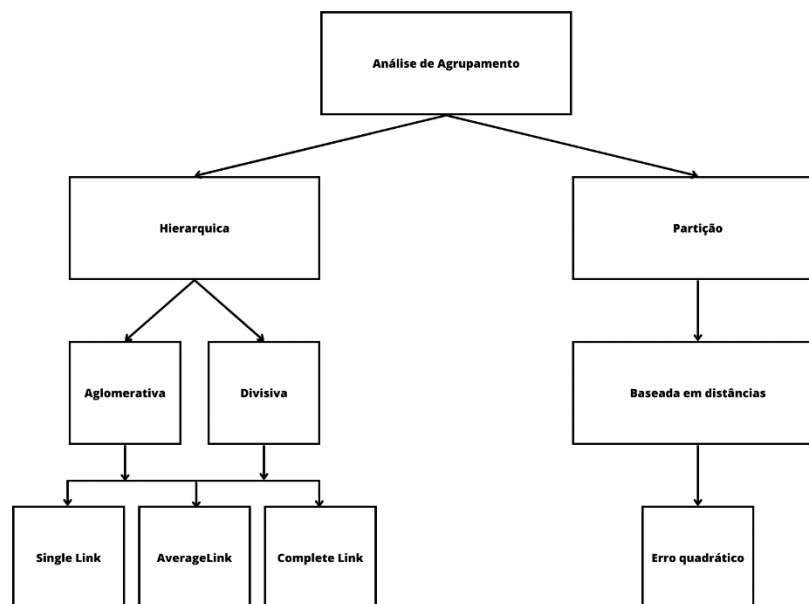


Fonte: Adaptado de SAXENA et al. (2017)

Saxena et. Al (2017) destaca que as críticas mais comuns feitas aos algoritmos hierárquicos são a falta de robustez, e, portanto, a sensibilidade que possuem para ruídos e valores discrepantes na série de dados. Uma vez que um objeto é associado à um grupo, ele não será considerado a outro grupo novamente, ou seja, algoritmos que usam a abordagem hierárquica não possuem a capacidade de corrigir agrupamentos errados feitos em uma iteração prévia.

Métodos de partição movem as observações, iterativamente, de um grupo para o outro, começando a partir de uma divisão já estabelecida, com o número de grupos precisa ser determinado com antecedência. O método de partição mais comum é o *K-Means* (FRALEY; RAFTERY, 1998). Vale ressaltar que nenhuma das abordagens tem definida em si a forma de determinação do número ideal de grupos. A Figura 15 apresenta a taxonomia das diferentes análises de agrupamento abordadas neste trabalho.

Figura 15 – Taxonomia das análises de agrupamento



Fonte: Adaptado de (FRALEY; RAFTERY, 1998).

### 2.7.1 K-Means

O K-Means é dos mais populares métodos de agrupamento não supervisionados (SINAGA; YANG, 2020) que parte da abordagem de partição, utilizando a distância euclidiana. De acordo com Sinaga e Yang (2020), o algoritmo pode ser definido como: sendo  $X =$

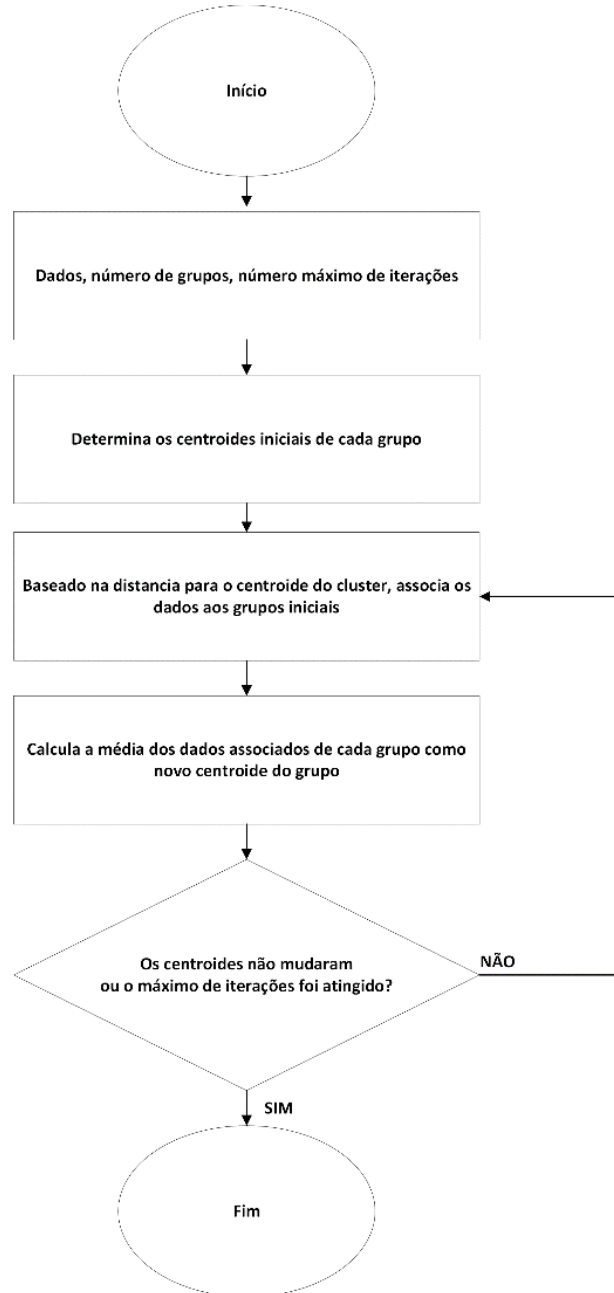
$\{x_1, \dots, x_n\}$  os dados em um espaço euclidiano de  $d$  dimensões,  $\mathbb{R}^d$ . Sendo  $A = \{a_1, \dots, a_c\}$  os  $c$  centroides de cada grupo. Sendo  $z = [z_{ik}]_{n \times c}$  onde  $z_{ik}$  é uma variável binária que indica se determinado ponto pertence ao  $k$  –ésimo grupo  $k = 1, \dots, c$ . Sendo a função objetivo do *K-Means* (distância)  $J(z, A) = \sum_{i=1}^n \sum_{k=1}^c z_{ik} (x_i - a_k)^2$ . O algoritmo *K-Means* é iterado de forma a minimizar a função objetivo, atualizando os centroides (Equação 7) de cada grupo e associações (Equação 8) a cada iteração.

$$a_k = \frac{\sum_{i=1}^n z_{ik} X_{ij}}{\sum_{i=1}^n z_{ik}} \quad (7)$$

$$z_{ik} = \begin{cases} 1, & \text{se } \|x_i - a_k\|^2 = \min_{1 \leq k \leq c} \|x_i - a_k\|^2 \\ 0, & \text{para qualquer outro caso.} \end{cases} \quad (8)$$

A Figura 16 resume o fluxo do algoritmo.

Figura 16 - Fluxograma do algoritmo K-Means



Fonte: Adaptado de SENARATHNA; HEMAPALA (2020).

A inicialização do método clássico do *K-Means* (determinação dos centroides iniciais) é aleatória, e Sinaga e Hemapala (2020) destacam que isso é um problema, visto que a inicialização pode afetar o resultado. Uma maneira de contornar isso é rodando o algoritmo N vezes e selecionando o melhor resultado dentre as N rodadas, baseado em alguma função objetivo.

### 2.7.1.1 Escolha do número ideal de grupos no *K-Means*

Kodinariya e Makwana (2013) destacam alguns métodos para a escolha do número ideal de grupos no algoritmo *K-Means*, dentre eles: Rule of Thumb, Elbow Method, Silhouette Score e validação cruzada.

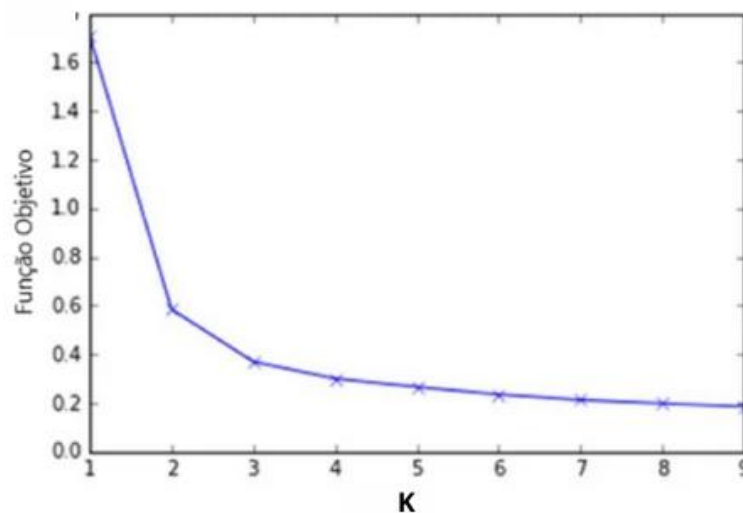
A *Rule of Thumb*, ou Regra de Ouro, é definida na Equação 9.

$$k \approx \sqrt{\frac{n}{2}} \quad (9)$$

Onde  $k$  é o número de grupos e  $n$  é o número de objetos (observações).

O método mais antigo para determinação do número de grupos é chamado de *Elbow Method*, ou Método do Cotovelo (NG, 2012). É um método visual, onde a ideia é mapear uma função objetivo em função do número  $K$  de grupos, iniciando em  $K = 2$ . Em algum momento, o valor de  $K$  irá cair significativamente, chegando num platô nos valores de  $K$  seguintes após esta queda. O número de  $K$  a ser escolhido é onde atingimos este platô (KODINARIYA; MAKWANA, 2013). A Figura 17 ilustra o método.

Figura 17 - Exemplo do Elbow Method



Fonte: Adaptado de (YUAN; YANG, 2019).

No caso ilustrado na Figura 17, a escolha do  $K = 2$  é direta, entretanto, existem casos em que o ponto ótimo não é facilmente identificado.

O conceito do *Silhouette Score* se baseia na espessura da silhueta, que por sua vez envolve a proximidade dos membros de um mesmo grupo e a separação destes de outros grupos.

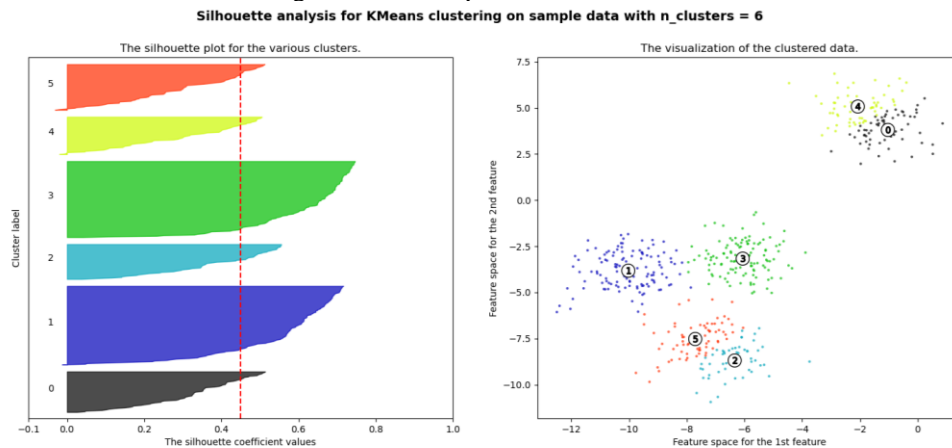


Especificamente, o *Silhouette Score*,  $s(i)$ , de um elemento  $i$  é dado pela Equação 10 (KODINARIYA; MAKWANA, 2013).

$$s(i) = \frac{b(i) - a(i)}{\max [a(i), b(i)]} \quad (10)$$

Onde  $a(i)$  é a distância média entre determinado elemento  $i$  e todos os outros elementos pertencentes ao mesmo grupo, e  $b(i)$  é a distância mínima entre a média das distâncias do elemento  $i$  e elementos pertencentes a outros grupos (KODINARIYA; MAKWANA, 2013), ou seja, a menor distância média que  $i$  possui dentre os grupos. O *Silhouette Score* varia de -1 a 1. Um valor de -1 significa que o elemento está associado ao grupo errado. Um valor próximo a 0 significa que o elemento poderia estar associado a outro grupo. Um valor próximo de 1 significa que o elemento está bem agrupado (KODINARIYA; MAKWANA, 2013). A Figura 18 apresenta uma representação visual do índice, onde o eixo das ordenadas está dividido em grupos, e cada segmento deste eixo em cada grupo representa um elemento  $i$ . A extensão (valor) deste elemento  $i$  no eixo das abscissas é o *Silhouette Score*,  $s(i)$ . A espessura de silhueta é o valor médio dos  $s(i)$  de cada grupo, podendo também se inferir a média do agrupamento como um todo. Quanto maior esta média, melhor agrupados estão os elementos na análise de agrupamentos.

Figura 18 - Exemplo do Silhouette Score



Fonte: SCIKIT-LEARN (2022).

## 2.8 NORMALIZAÇÃO DOS DADOS

Durante o processo de análise de cluster um dos problemas metodológicos envolve a decisão de normalizar ou não os dados de entrada antes da aplicação de qualquer algoritmo de agrupamento. A normalização de variáveis se faz necessária quando uma métrica de dissimilaridade é aplicada, como a distância euclidiana, que é sensível a diferenças de escala e magnitude dos dados de entrada (MIUIGAN et al., 1988; MOHAMAD; USMAN, 2013). Mohamad e Usman (2013) afirmam que distancias calculadas entre elementos que possuem grande variação irá implicar em maiores métricas em relação a elementos de menor variação. Além disso, tais elementos precisam ser adimensionais, visto que a variação numérica de tais elementos depende da escolha da unidade de medida, e, portanto, a seleção da unidade de medida pode afetar significativamente o resultado do processo de análise de agrupamentos. Portanto, funções objetivo que utilizam distancias geométricas, como a distância euclidiana, não devem ser utilizadas sem a normalização prévia dos dados (AKSOY; HARALICK, 2001). A normalização previne que sobrepesos sejam aplicados a elementos de valores numéricos altos em relação a elementos com valores numéricos baixos (MOHAMAD; USMAN, 2013).

Dentre os métodos de normalização, podem ser citados: Z-Score, Min-Max Scaling e Decimal Scaling (MOHAMAD; USMAN, 2013).

Segundo Miuligan e Cooper (1988), pesquisadores das áreas de ciências sociais geralmente definem que uma variável normalizada foi transformada de modo a ter média zero e variância unitária. Este é o caso de um dos métodos de normalização denominado *Z-score* (Equação 11).

$$Z = \frac{X - \bar{X}}{s} \quad (11)$$

Onde  $X$  é o dado original,  $\bar{X}$  é a média da série de dados e  $s$  é o desvio padrão.

A normalização Decimal Scaling, ou de Escala Decimal, move o ponto decimal dos dados de um elemento  $X$ , com base no valor máximo absoluto de  $X$  (MOHAMAD; USMAN, 2013). Tal método é definido pela Equação 12.

$$DS(X_{ij}) = \frac{X_{ij}}{10^c} \quad (12)$$

Onde  $c$  é o menor inteiro que respeite que  $\max[|DS(X_{ij})|] < 1$ .

A normalização Min-Max transforma os dados em valores entre 0 e 1, onde o menor valor é definido como 0 e o maior valor definido como 1. Tal técnica provê uma maneira fácil de se comparar valores medidos com diferentes unidades de medida. Tal normalização é definida na Equação 13.

$$MM(X_{ij}) = \frac{X_{ij} - X_{min}}{X_{max} - X_{min}} \quad (13)$$

## 2.9 ÍNDICES ESTATÍSTICOS

### 2.9.1 Coeficiente de Nash-Sutcliffe

O coeficiente de Nash-Sutcliffe é amplamente utilizado, e potencialmente confiável, para quantificar quão bem um modelo pode prever o valor observado (NAGELKERKE, 1991). A métrica é apresentada na Equação 14. Um valor de 1 representa o cenário ideal, onde não há erro em nenhum valor previsto. O coeficiente pode variar do infinito negativo à 1.

$$E = 1 - \frac{\sum_{i=1}^n (V_o^n - V_p^n)^2}{\sum_{i=1}^n (V_o^n - \overline{V_o})^2} \quad (14)$$

Onde  $\overline{V_o}$  é a média dos valores observados,  $V_o^n$  é o  $n$  – ésimo valor observado e  $V_p^n$  é o  $n$  – ésimo valor previsto.

### 2.9.2 Erro médio absoluto percentual

O erro médio absoluto percentual, em inglês *Mean Absolute Percentage Error* (MAPE), é utilizado frequentemente na avaliação de performance de modelos por conta de sua interpretação intuitiva em termos de erro relativo (DE MYTTENAERE et al., 2016) Tal métrica pode ser descrita pela Equação 15.

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{O_t - P_t}{O_t} \right| \quad (15)$$

Onde  $O_t$  é o valor observado,  $P_t$  é o valor previsto,  $n$  e  $n$  é o número de observações.

## 3 MATERIAIS E MÉTODOS

### 3.1 MATERIAIS

A metodologia desenvolvida neste estudo foi aplicada a dados de ENAs previstas para um horizonte de 46 dias, fornecidas pela Empresa Norus. O produto Pluvia/Norus gerou dados de ENA para cada bacia do SIN, a partir da conversão de dados de vazão resultantes da aplicação do SMAP usando produto *Extended-Range* do ECMWF (36 km de resolução espacial), do dia 23 de Setembro de 2021. O produto *Extended-Range* do ECMWF possui 51 cenários de precipitação prevista (51 membros).

A data de 23/09/2022 foi escolhida em função da observação da grande diferença entre os dados de ENA observados no período, e os resultados da previsão de ENA oficial do ONS, que utiliza um modelo de precipitação conjunta com os primeiros 14 dias do modelo *EC-Ensemble*. Ressalta-se que esta data está dentro de um período de transição de um período seco para um período molhado para grande parte das bacias do SIN.

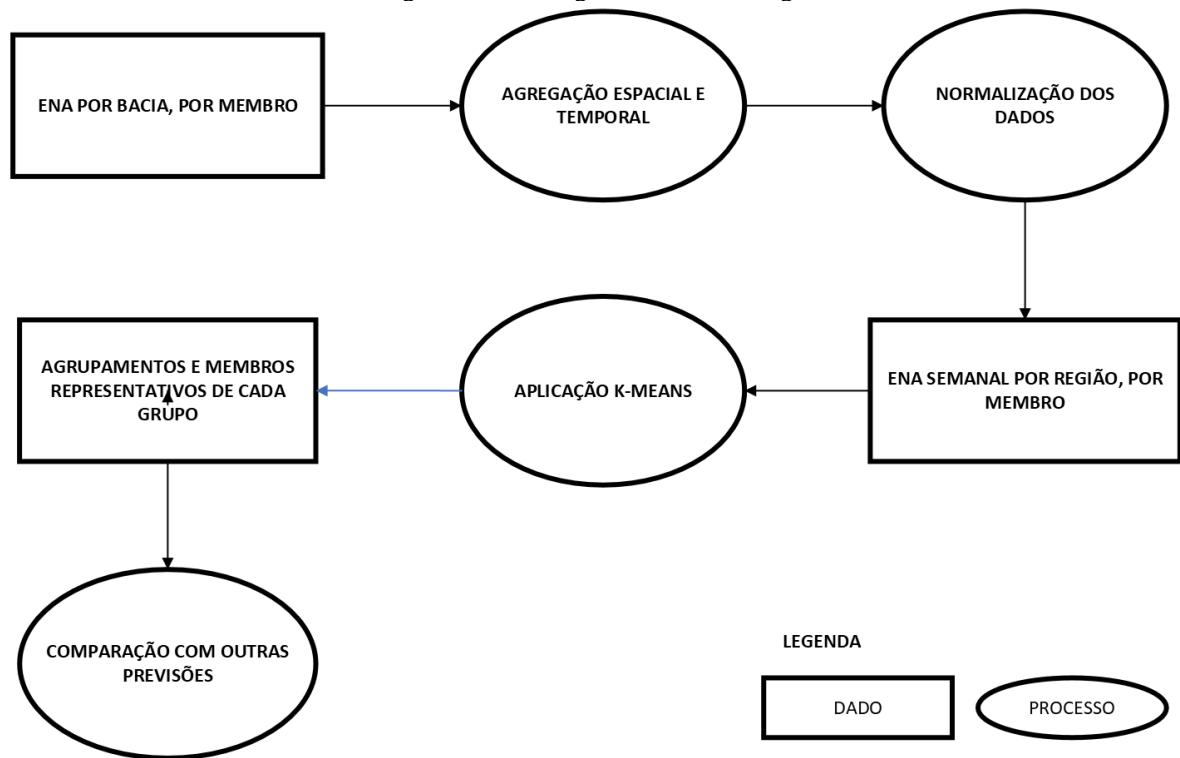
### 3.2 MÉTODOS

Este trabalho constitui-se de um estudo de caso, feito com base em dados de ENAs previstas para um horizonte de 45 dias, a partir de resultados do modelo SMAP com o uso dos 51 membros de precipitação prevista do produto *Extended-Range* do ECMWF, para 23/09/2021. Para atingir o objetivo geral de avaliar a assertividade de uma previsão de ENA gerada a partir de uma análise de agrupamento dos 51 membros dessa previsão, foram desenvolvidos os passos metodológicos apresentados na Figura 19.

Primeiro, para cada bacia do SIN, foram adquiridos os dados de ENA, para um período de 45 dias. Ressalta-se que, para cada um dos 45 dias, há 51 previsões de ENA, visto que cada uma das 51 previsões de precipitação do *Extended-Range* do ECMWF gera um resultado no SMAP. Para as análises posteriores, os dados de ENA foram separados em dois grupos (grupos das bacias localizadas a Norte do Brasil e Grupo das Bacias Localizadas mais ao Sul do Brasil), em função das diferentes características de precipitação e ENA dessas regiões. Também foi feita uma agregação temporal, que consistiu em calcular a ENA média semanal de cada um dos grupos, de forma a reduzir a dimensão dos elementos que entram na análise de agrupamentos. Após isso, foi realizado um processo de normalização dos dados através do *z-score*. Por fim o algoritmo *K-Means* é aplicado nas 51 previsões. Foi calculado o membro representativo de cada um dos grupos gerados pelo algoritmo, que é o cenário de previsão de precipitação mais próximo ao centroide final do grupo. Tendo em mãos os grupos, número de previsões associadas à cada grupo e o membro representativo de cada grupo, calculou-se a ENA média ponderada, de modo incluir as informações obtidas pela análise de agrupamento num único cenário de previsão. Ao final, essa previsão foi comparada, graficamente e estatisticamente, com a previsão oficial do ONS realizada na mesma data e com a previsão realizada a partir da média das 51 previsões de precipitação do produto *Extended-Range*.

Tal metodologia é a vigente em estudos de análise de agrupamentos desenvolvidos pelo ONS, conforme descrito por ONS (2022a).

Figura 19 - Fluxograma da metodologia



Fonte: Elaboração própria.

### 3.2.1 Dados de entrada

Os dados e processamento necessários para se chegar no primeiro dado da Figura 19, foram obtidos da empresa Norus, através do produto Pluvia. A Tabela 1 exemplifica a estrutura na qual os dados de cada cenário de previsão são recebidos. As linhas representam diferentes dias da previsão, e as colunas representam as diferentes bacias do SIN.

Tabela 1 - Estrutura dos dados de ENA utilizados

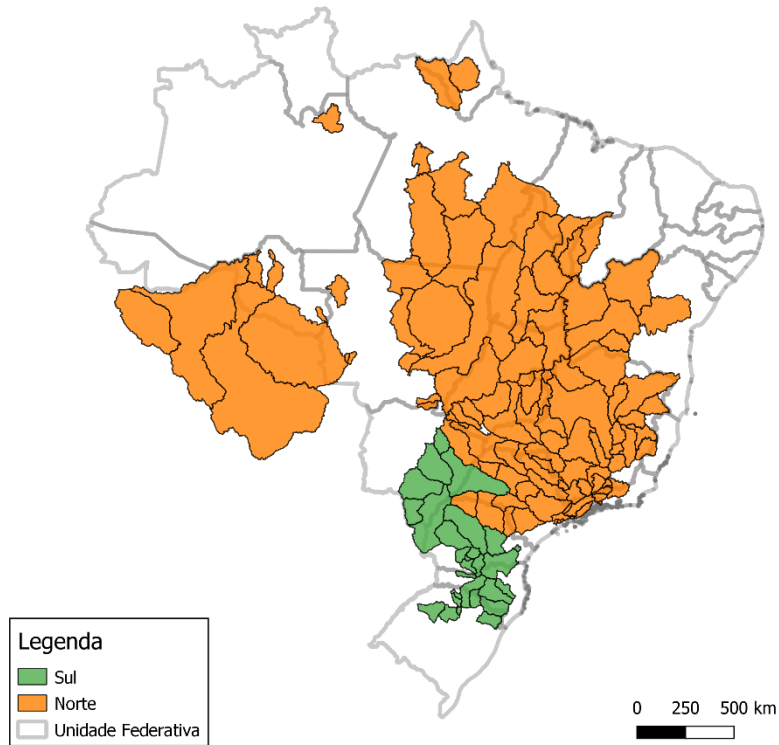
<i>Dia previsto</i>	<i>B1</i>	...	<i>B30</i>
<i>D1</i>	ENAB1-D1	...	ENAB30-D1
...	...	ENAB...-D...	...
<i>D45</i>	ENAB1-D45	...	ENAB30-D45

Fonte: Elaboração própria.

### 3.2.2 Agregação espacial

Por conta da extensão territorial brasileira, incluindo os diferentes regimes de precipitação, massas de ar atuantes, períodos atuantes das massas de ar e temperatura, as ENAs diárias por bacia foram agrupadas em duas regiões, de acordo com a Tabela 2 e Figura 20.

Figura 20 - Agregação espacial escolhida para a análise de agrupamentos



Fonte: Elaboração própria.

A agregação Norte, é composta majoritariamente pela região Norte, Sudeste, Centro-Oeste e partes do Nordeste. O clima predominante nessas regiões é o Equatorial, nas bacias que afluem em Rondônia e em parte da bacia do Xingu, e Tropical Equatorial na região Sudeste e Nordeste abrangida pelas áreas de contribuição das bacias. A distribuição temporal da precipitação nessas regiões não é homogênea, ou seja, possuem o período molhado definido.

A agregação Sul, composta majoritariamente pela região Sul e parte do Centro-Oeste e Sudeste, possui predominantemente clima Subtropical Úmido, com médias térmicas mais baixas em relação ao restante do país e precipitação bem distribuída ao longo do ano.

Tabela 2 - Agregação espacial utilizada

<b>Norte</b>	<b>Sul</b>
GRANDE	ITAJAI-ACU
ALTO TIETE	URUGUAI
PARAIBA DO SUL	JACUI
DOCE	CAPIVARI
AMAZONAS (SE)	BAIXO PARANA
SAO FRANCISCO (SE)	IGUACU
SAO FRANCISCO (NE)	
JEQUITINHONHA (NE)	
PARNAIBA	
ITABAPOANA	
PARANAIBA	
TIETE	
ALTO PARANA	
PARAGUACU	
JEQUITINHONHA (SE)	
PARAGUAI	
AMAZONAS (N)	
TOCANTINS (SE)	
TOCANTINS (N)	
ARAGUARI	
MUCURI	
XINGU	
PARANAPANEMA (SE)	
PARANAPANEMA (S)	

Fonte: Elaboração própria.

Além da diferença de climas predominantes em cada agregação, também há a diferença de massas de ar que atuam sobre cada região, e acabam por influenciar nos regimes de precipitação. A agregação Norte dá origem à massa Equatorial continental, de característica quente e úmida, e que atinge regiões do extremo sul do Sudeste durante o Verão, e se torna mais tímida no inverno. A agregação Norte e Sul sofrem influências da massa Tropical atlântica, também de característica quente e úmida e atuante sobretudo no litoral do Sul ao Nordeste, durante o verão (ZANATTA et al., 2016). A massa Polar atlântica, de característica fria e úmida, influencia exclusivamente a agregação Sul durante o Verão, mas se estende até porções do Sudeste durante o inverno, e conseqüentemente, influencia a agregação Norte. A massa Tropical continental, de característica quente e seca, influencia agregação Norte através da região Sudeste durante o Verão, porém, cede espaço para a massa Polar atlântica durante o Inverno (ZANATTA et al., 2016).



Além desses fatores, tal agregação espacial também foi utilizada em estudo preliminares do ONS sobre análise de agrupamentos, justificada pelos diferentes regimes pluviométricos e diferenças de clima predominante em cada região (ONS, 2022a).

A Tabela 2, anteriormente composta por 30 bacias do SIN, passa a ter 2 colunas, representando o agrupamento Norte e Sul adotado, conforme exemplificado na Tabela 3.

Tabela 3 - Dados de ENA com agregação espacial adotada

<i>Dia Previsto</i>	<i>Norte</i>	<i>Sul</i>
<i>D0</i>	ENANorte-D0	ENASul-D0
...	...	...
<i>D45</i>	ENANorte-D45	ENASul-D45

Fonte: Elaboração própria.

O cálculo de ENA com agregação espacial seguiu o descrito na Equação 7, com a exceção de que, como os dados já estavam em MWmed, bastou somar a ENA das diferentes bacias que compõem a agregação espacial adotada.

### 3.2.3 Agregação temporal

As ENAs diárias agregadas por região foram agregadas temporalmente pela média semanal. É evidente que ao fazer a média de qualquer dado, ocorre uma perda de informação. Com isso em mente, aliado ao fato de que estamos lidando com uma previsão que possui um horizonte de 45 dias, que não é um dividendo natural do número 7, escolheu-se incluir sempre na média da primeira semana prevista os três dias que excedem as 6 semanas completas possíveis de prever com tal horizonte.

Esta escolha foi feita por conta de a dispersão entre os membros da previsão de precipitação serem menores na primeira semana prevista, ou seja, ao fazermos a média dos dias excedentes perderemos menos informação em relação às semanas posteriores. Portanto, a média da primeira semana é, na verdade, a média dos primeiros 10 dias de previsão.

Após a agregação temporal, os dados passaram a ter o formato exemplificado na Tabela 4.

Tabela 4 - Dados de ENA com agregação espacial e temporal aplicada.

<i>Semana prevista</i>	<i>Norte</i>	<i>Sul</i>
<i>S1</i>	ENANorte-S1	ENASul-S1
<i>S2</i>	ENANorte-S2	ENASul-S2
<i>S3</i>	ENANorte-S3	ENASul-S3
<i>S4</i>	ENANorte-S4	ENASul-S4
<i>S5</i>	ENANorte-S5	ENASul-S5
<i>S6</i>	ENANorte-S6	ENASul-S6

Fonte: Elaboração própria.

### 3.2.4 Normalização dos dados

A normalização dos dados também foi implementada através da linguagem de programação Python, sendo escolhido o método de normalização de dados *z-score*. Os dados da Tabela 4 foram normalizados por região, aplicando a Equação 11 e utilizando os valores de média e desvio padrão da respectiva região que a coluna representa.

### 3.2.5 Aplicação do *K-Means* em espaço tridimensional

Uma peculiaridade da análise de agrupamento aplicada neste trabalho é sua dimensionalidade, visto que ela é feita em três dimensões (espacial, temporal e membros da previsão), mas com elementos e centroides que consistem de matrizes  $M \times N$ , onde  $M$  são as semanas previstas e  $N$  as regiões espaciais nas quais as bacias foram agregadas. O objetivo do algoritmo é, portanto, agrupar  $C$  matrizes  $M \times N$  em  $K$  grupos, onde  $C$  representa o número de cenários do modelo *Extended-Range* como um todo. Em uma análise de agrupamento tradicional os valores que compõem as matrizes ao longo do espaço  $M \times N \times C$  seriam agrupados entre si, sem considerar a matriz  $M \times N$  como um elemento só, que representa uma das perturbações da previsão *Ensemble*.

O método de agrupamento *K-Means* foi implementado através da linguagem de programação *Python 3.9*, assim como a compilação dos dados no formato desejado. O algoritmo foi construído desde o início, visto que nenhuma das bibliotecas disponíveis que continham o algoritmo suportavam a aplicação dele considerando a peculiaridade discutida acima.

A distância entre cada elemento  $A$  e  $B$  da análise, matrizes  $M \times N$ , pode ser descrita pela norma de Frobenius, apresentada na Equação 16:

$$\|A - B\| = \sum_{i=1}^M \sum_{j=1}^N |(a_{ij} - b_{ij})|^2 \quad (16)$$

Onde  $a$  e  $b$  são os elementos de suas respectivas matrizes  $A$  e  $B$ . Em outras palavras, é calculada a distância euclidiana entre cada elemento correspondente entre as matrizes, e então, somados os valores da matriz de erros resultante.

### 3.2.6 Elbow Method

A escolha do número ideal de grupos se deu através do *Elbow Method*, com a função objetiva sendo a SSE. A SSE é definida como a soma das diferenças entre o dado associado à um grupo e o centro desse grupo (NAINGGOLAN et al., 2019), que neste trabalho foi definido como o próprio centroide da iteração final do algoritmo. Como cada elemento da análise de agrupamentos deste trabalho é um dado bidimensional, o SSE acaba por ser a soma da norma de Frobenius aplicada entre o centroide do grupo e seus respectivos elementos, como descrito na Equação 17.

$$SSE = \sum_{g=1}^K \sum_{i=1}^M \sum_{j=1}^N |(a_g - b_{cg})_{ij}|^2 \quad (17)$$

Onde,  $K$  é o número de grupos,  $a_g$  é o elemento associado ao  $g$  – ésimo grupo e  $b_{cg}$  é o centroide do  $g$  – ésimo grupo.

### 3.2.7 ENA do membro representativo

Calculou-se, através do Python e tomando como base a norma de Frobenius, o membro de cada grupo que estava mais próximo ao seu respectivo centroide ao final da análise de agrupamento, denominado aqui de membro representativo.

### 3.2.8 ENA média ponderada

Cada membro da previsão por conjuntos possui a mesma probabilidade de se consolidar na realidade. A partir do momento que cada membro da previsão *Extended-Range* foi alocado à um grupo, buscou-se construir uma ENA média ponderada da análise de agrupamentos, que carregue em si a informação gerada pela análise.

A ENA média ponderada da análise de agrupamento foi calculada com base na média ponderada da quantidade de membros atribuída a cada grupo, tomando os membros representativos de cada grupo como base de cálculo. A ENA em determinado espaço de tempo, após aplicação da análise de agrupamentos, pode ser descrita pela Equação 18.

$$ENA_t = \frac{\sum_{i=1}^K w_i ENA_{it}}{\sum_{i=1}^n w_i} \quad (18)$$

Onde,  $t$  é o passo de tempo sob o qual se está calculando a ENA,  $K$  é o número de grupos,  $ENA_{it}$  é a ENA do membro representativo do  $i$  – ésimo no espaço de tempo escolhido,  $w_i$  é o número de membros associados ao  $i$  – ésimo grupo e  $ENA_t$  é a ENA média ponderada no passo de tempo  $t$ .

### 3.2.9 Comparação com ENA observada

Foram compilados dados de ENA observada, que é calculada com base em dados de vazões observadas disponibilizados diariamente pelo ONS. A ENA foi obtida através da Equação 5, aplicada em um passo de tempo semanal, ou seja, com a vazão média semanal de cada bacia.

### 3.2.10 COMPARAÇÃO COM ENA OFICIAL ONS

Foram compilados dados do Relatório de Previsão de Vazões e Geração de Cenários de Afluências, do PMO de Outubro de 2021, disponibilizados pelo ONS na plataforma Sintegre (ONS, 2021), de forma a obter a previsão realizada pela metodologia atual aplicada pelo ONS para cada subsistema.

### 3.2.11 COMPARAÇÃO COM MÉDIA DO ENSEMBLE

A média do Ensemble consiste na aplicação da precipitação média dos 51 membros da previsão *Extended-Range* no modelo chuva-vazão, que é posteriormente convertida em ENA. Tal processamento é realizada pelo produto Pluvia, e foi disponibilizada juntamente com os outros dados para este trabalho.

### **3.2.12 MÉDIAS DE LONGO TERMO**

As médias de longo termo (MLTs) são geradas a partir da média de longo termo da ENA em determinada bacia. A ENA é compilada mensalmente, a partir da vazão média mensal histórica de cada bacia, e então multiplicada pela produtibilidade das usinas pertencentes à bacia para obter a ENA dela. A MLT de um subsistema é calculada através da média das MLTs de bacias que compõem o subsistema. Estas médias foram obtidas diretamente dos dados internos do produto Pluvia.

### **3.2.13 MÉTRICAS CALCULADAS**

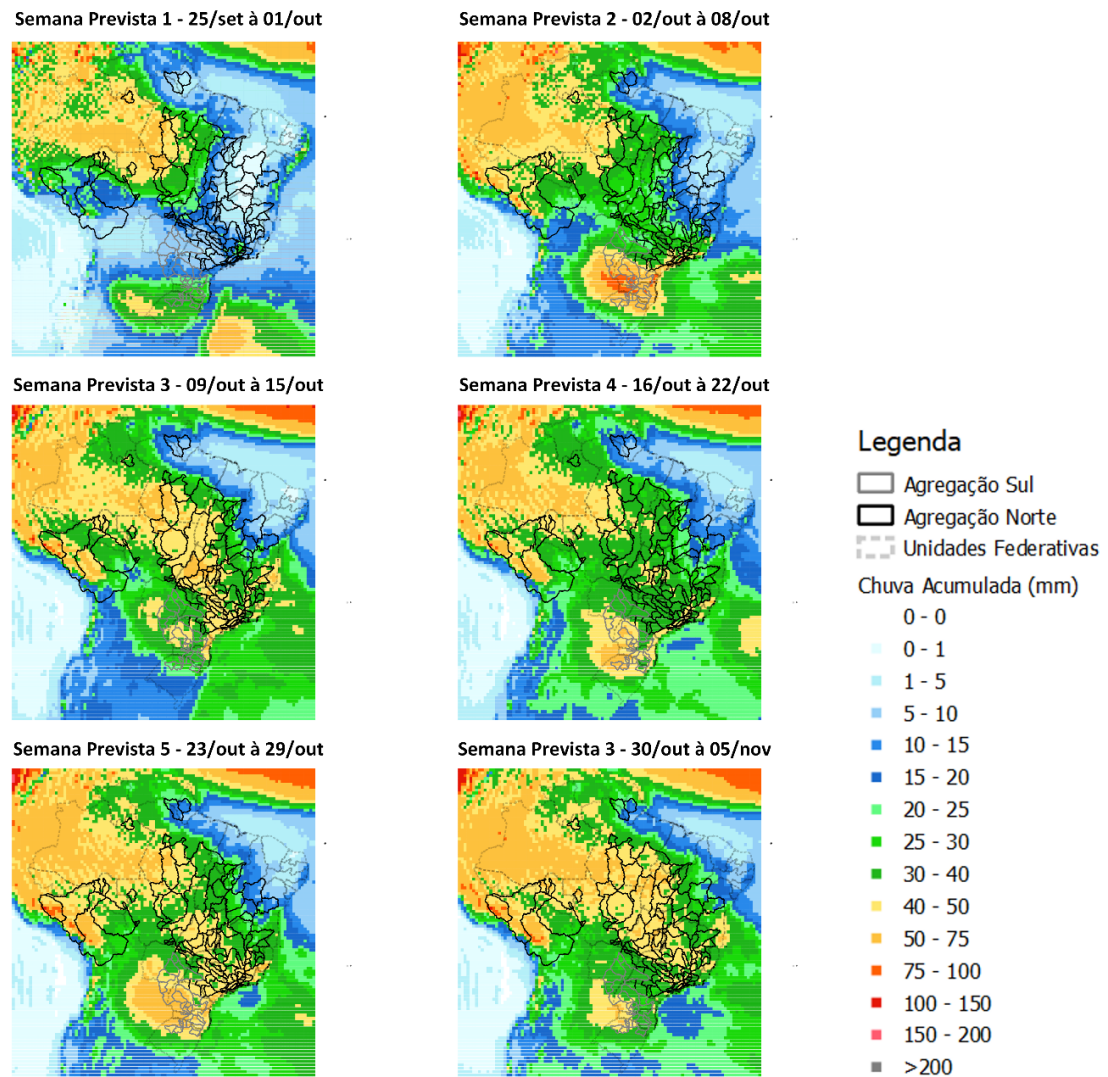
O coeficiente de Nash-Sutcliffe e Erro médio absoluto percentual foram calculados entre as médias semanais de cada previsão, sempre comparando-as com a série de ENA observada.

## **4 RESULTADOS**

### **4.1 MÉDIA DA PRECIPITAÇÃO PREVISTA**

Para a contextualização do cenário específico que está sendo estudado, calculou-se a chuva acumulada em cada semana prevista através da média de precipitação dos 51 membros, que pode ser visualizado na Figura 21.

Figura 21 - Precipitação acumulada nas semanas previstas



Fonte: Elaboração própria.

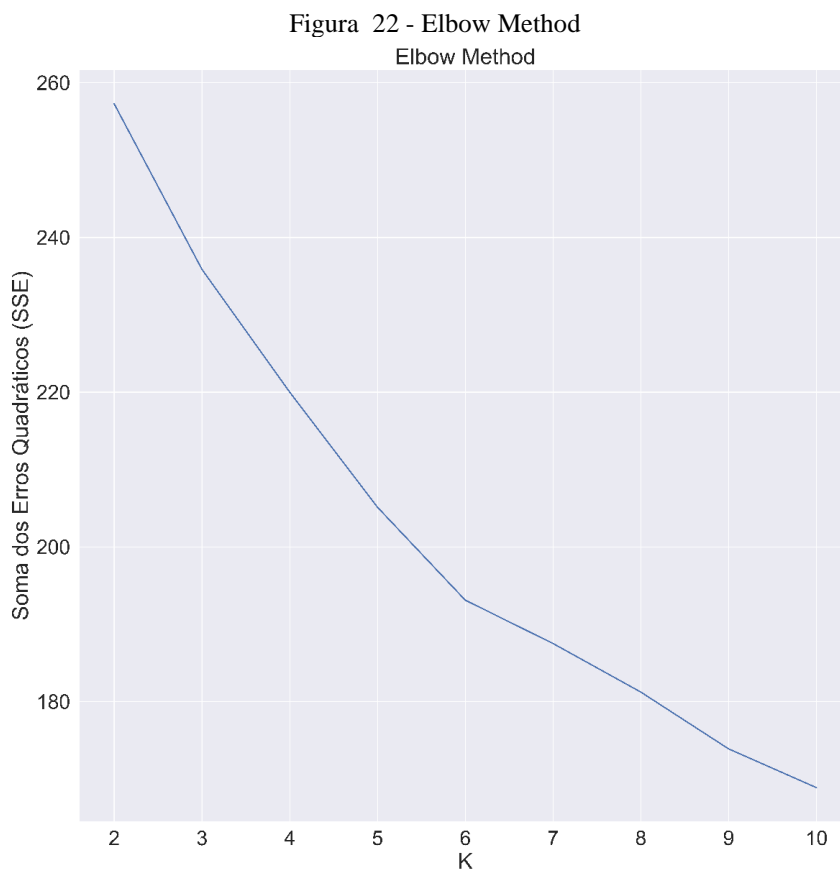
Observa-se acumulados constantes de 40 a 50 mm na região Amazônica, mas que impactam apenas áreas de contribuição da bacia do Xingu, que está relacionada à agregação e subsistema Norte. Nas regiões do Subsistema Sudeste, a chuva acumulada semanal é de até 20mm na primeira semana, mas alcança valores superiores a este durante o restante do período em estudo, chegando à 50mm semanais nas semanas 3, 5 e 6.

A região do subsistema Nordeste registra chuvas muito fracas durante a primeira semana, não passando de 5mm, porém, a chuva vinda da região Amazônica atinge as áreas de contribuição da bacia do São Francisco nas semanas 3, 4 e 5, chegando à 40mm semanais, e até mesmo 50 mm na semana 6, em áreas específicas.

O subsistema Sul observou chuvas maiores que 20 mm durante todas as 6 semanas, com um foco de chuva maior que 40 mm que atingiu principalmente à bacia do Rio Uruguai, mas que se estendeu até o Baixo Paraná nas semanas 4 e 5.

#### 4.2 ESCOLHA DO NÚMERO DE GRUPOS

A Figura 22 mostra a variação do SSE de acordo com o número de grupos. Conforme esperado, a SSE diminuí com o aumento do número de grupos, entretanto, não é possível identificar um ponto K onde houve um decréscimo significativo na SSE.



Fonte: Elaboração própria.

Diante da ausência de um decréscimo significativo no gráfico, o número de grupos foi determinado de acordo com a menor SSE gerada, onde  $K = 10$ . Um número de grupos maior que este não foi especulado visto que isto geraria um número excessivo de cenários possíveis, dificultando a análise e visualização deles.

#### 4.3 DISPERSÃO DOS MEMBROS REPRESENTATIVOS POR AGREGAÇÃO

As Figuras 23 e 24 apresentam as ENAs geradas pelos membros representativos de cada grupo, somadas conforme agregação espacial aplicada para a análise. A distribuição do número de membros por grupo pode ser visualizada na Tabela 5.

Tabela 5 - Número de membros por grupo

<b>Grupo</b>	<b>Número de Membros</b>	<b>Probabilidade de ocorrência (%)</b>
<b>0</b>	7	13,7
<b>3</b>	7	13,7
<b>5</b>	7	13,7
<b>9</b>	7	13,7
<b>2</b>	6	11,7
<b>1</b>	4	7,84
<b>4</b>	4	7,84
<b>8</b>	4	7,84
<b>7</b>	3	5,8
<b>6</b>	2	3,9

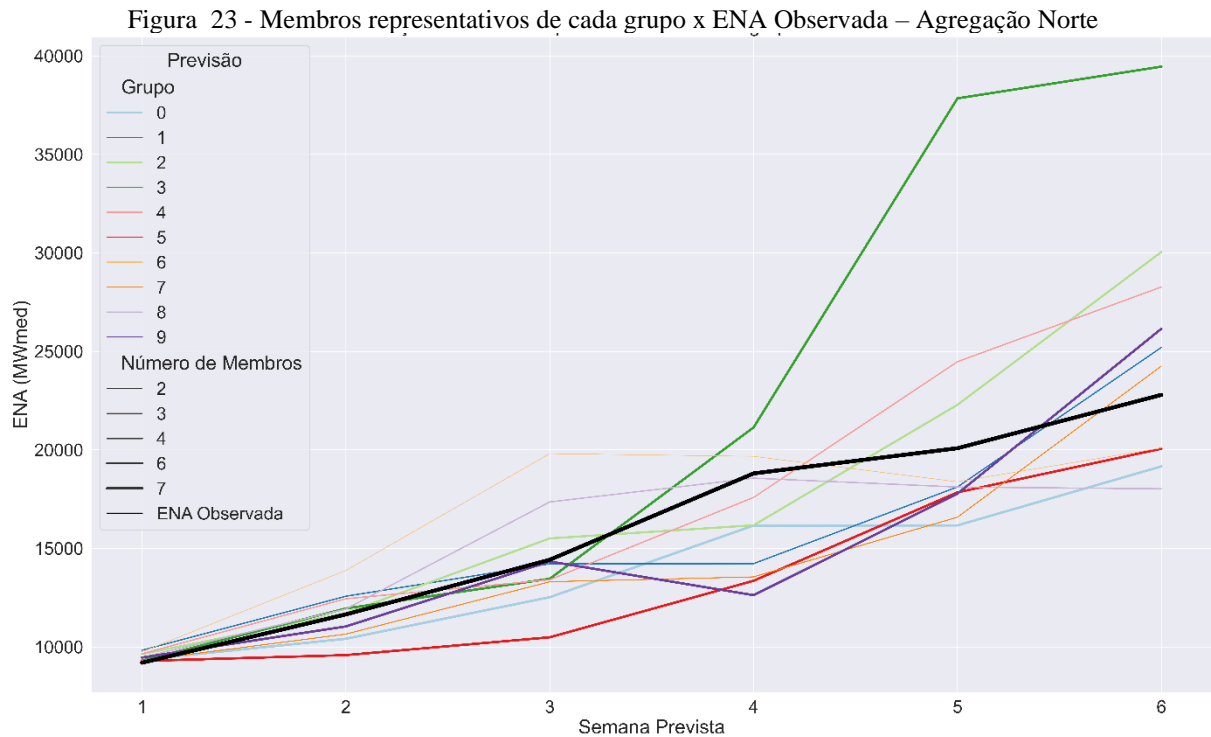
Fonte: Elaboração própria.

Os membros ficaram bem distribuídos ao longo de todos os grupos, com exceção dos grupos 6 e 7, que tiveram apenas 2 e 3 membros associados a eles, respectivamente. Tais grupos possuem a menor possibilidade de ocorrência da previsão.

Na agregação Norte (Figura 23), observamos que todos os grupos, exceto os grupos 6 e 8, apresentaram tendência de crescimento semanal da ENA prevista, porém, com intensidade e dispersões diferentes ao longo das semanas 2 à 6. A análise de agrupamento conseguiu mostrar diferenças de comportamento entre os grupos principalmente a partir da segunda semana prevista, onde os membros representativos dos grupos já apresentam uma dispersão na faixa de 10 GWmed entre si, com destaque para o grupo 3, que registrou um crescimento semanal acima da ENA observada, chegando a mais de 22 GWmed na quarta semana prevista.

Além dos valores das ordenadas, destacam-se também as espessuras das linhas que representam cada grupo, pois estas nos trazem a informação de quantos membros foram associados a determinado grupo, e conseqüentemente, o quão provável é a ocorrência daquele cenário deste grupo.

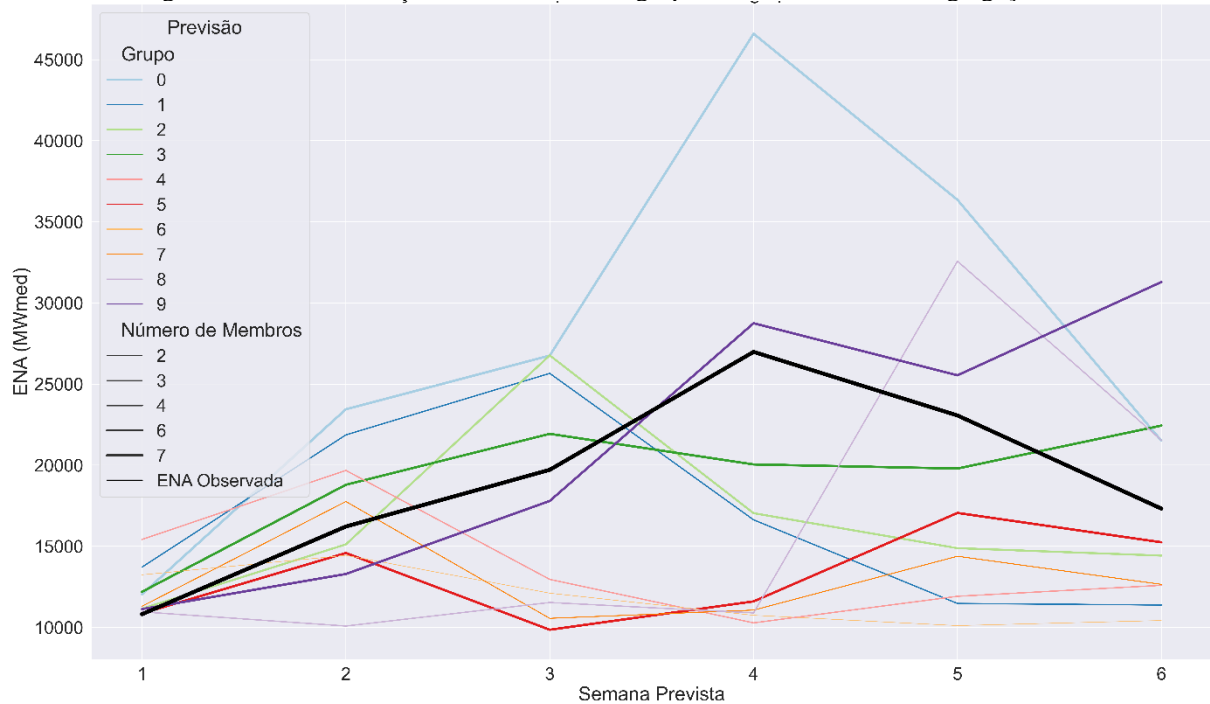




Fonte: Elaboração própria.

Na agregação Sul (Figura 24), os membros representativos de cada grupo apresentaram dispersão entre si desde a segunda semana. Observam-se também uma maior variedade de comportamentos de aumento e redução na ENA prevista entre os grupos ao longo das semanas previstas, diferentemente do que acontece na Figura 23, onde a maioria apresentou um comportamento de crescimento ao longo das semanas. Tal fato indica que as perturbações no modelo *Extended-Range* geraram cenários mais caóticos e com diferentes comportamentos na agregação Sul, compreendida entre as bacias do Jacuí e do Baixo Paraná.

Figura 24 - Membros representativos de cada grupo x ENA Observada – Agregação Sul

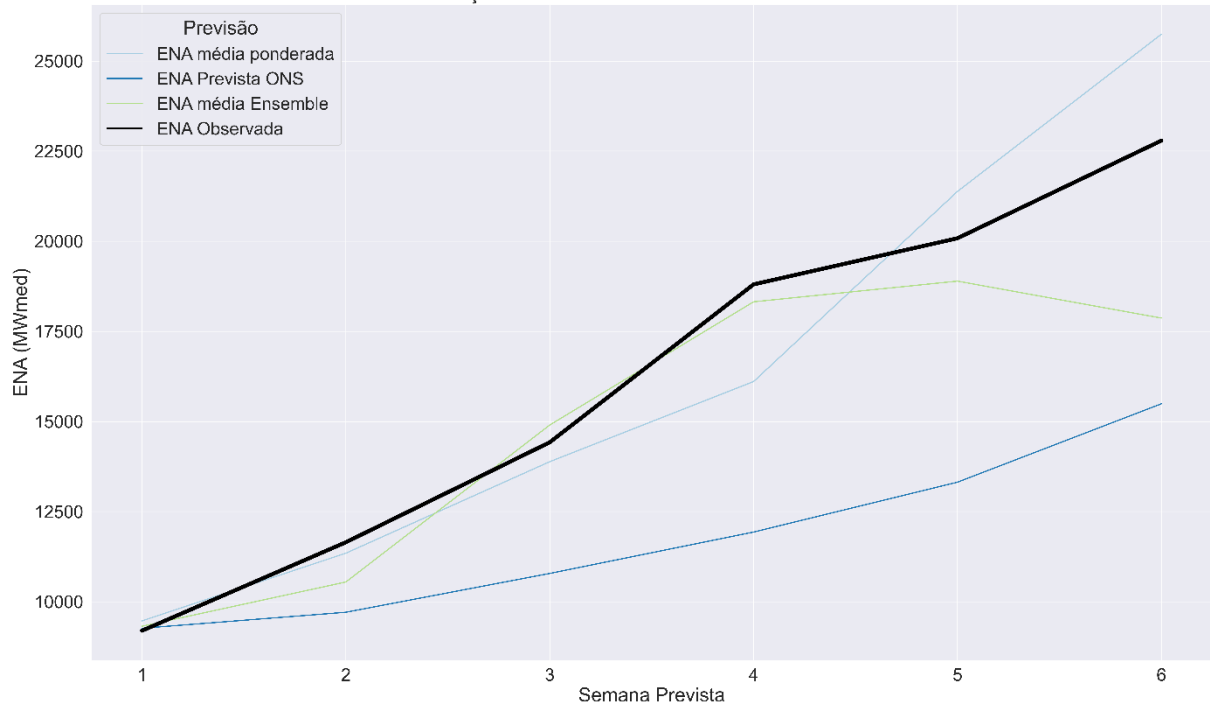


Fonte: Elaboração própria.

#### 4.4 COMPARAÇÃO COM OUTRAS PREVISÕES POR AGREGAÇÃO

Na Figura 25 temos a comparação, na agregação Norte, entre a ENA média do Ensemble, a ENA da previsão oficial NOS, a ENA observada e ENA média ponderada, calculada através da análise de cluster, como exposto anteriormente. A previsão oficial ONS subestimou a ENA observada durante o período de estudo. A ENA média Ensemble e ENA média ponderada se mantiveram próximas até a quarta semana prevista, porém, a partir da quinta semana, a previsão média do Ensemble registrou uma queda, enquanto a ENA observada e ENA média ponderada seguiram em crescimento.

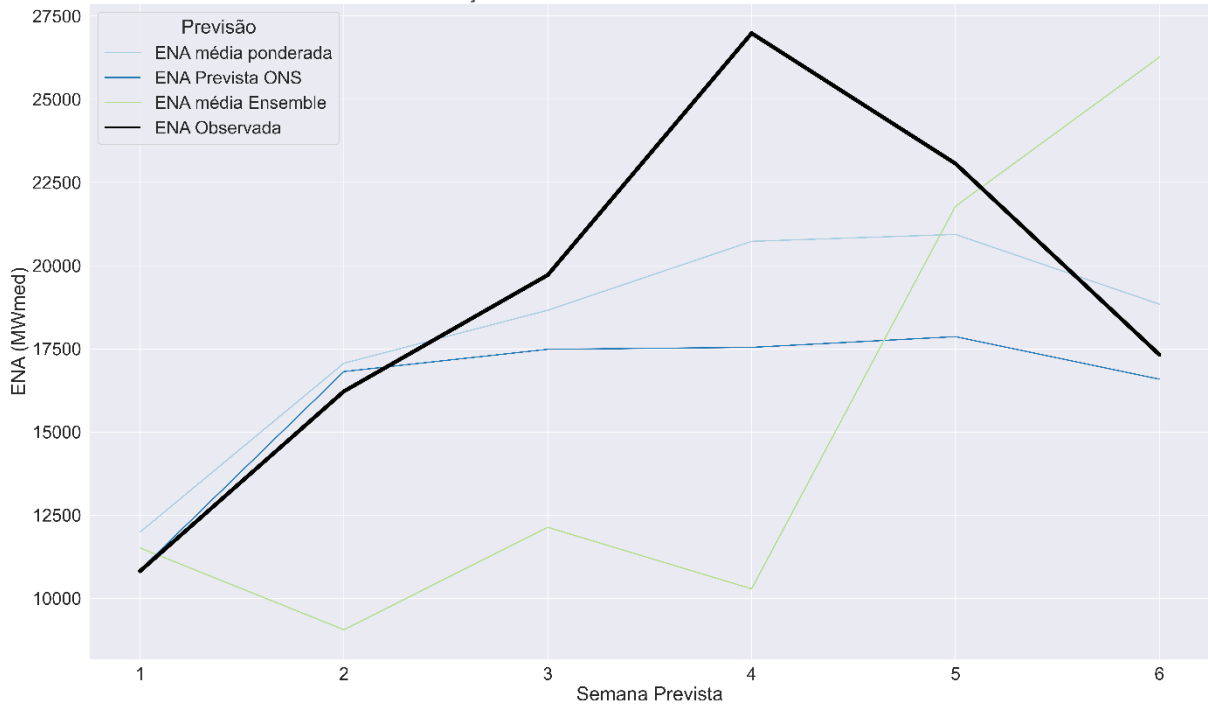
Figura 25 - Comparação de previsões – Agregação Norte  
AGREGAÇÃO ENSEMBLE x CLUSTER x ENA PREVISTA - N



Fonte: Elaboração própria.

Já na agregação Sul (Figura 26), a ENA da previsão média do Ensemble subestimou a ENA observada da primeira à quarta semana, apresentando uma tendência completamente diferente do que se realizou. A ENA da previsão oficial ONS mostrou comportamento semelhante à previsão calculada através da análise de agrupamento, com exceção de um pico na quarta semana prevista que não foi exibido pela ENA por média ponderada.

Figura 26 - Comparação de previsões – Agregação Sul  
AGREGAÇÃO ENSEMBLE x CLUSTER x ENA PREVISTA - S

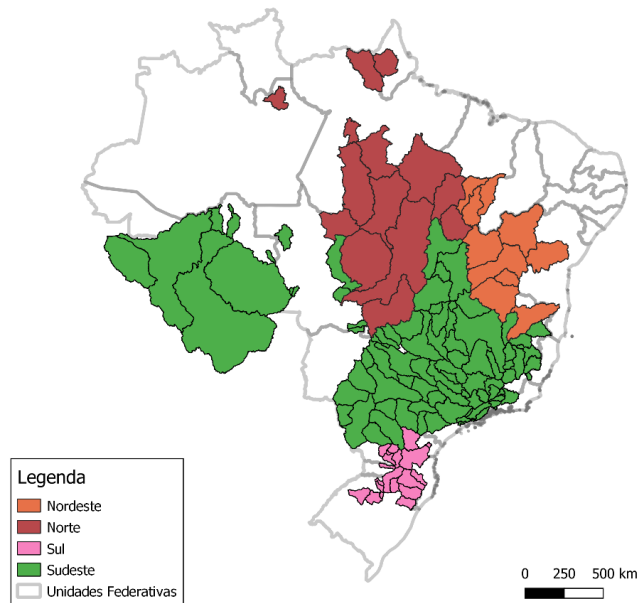


Fonte: Elaboração própria.

#### 4.5 DISPERSÃO DOS MEMBROS REPRESENTATIVOS POR SUBMERCADO

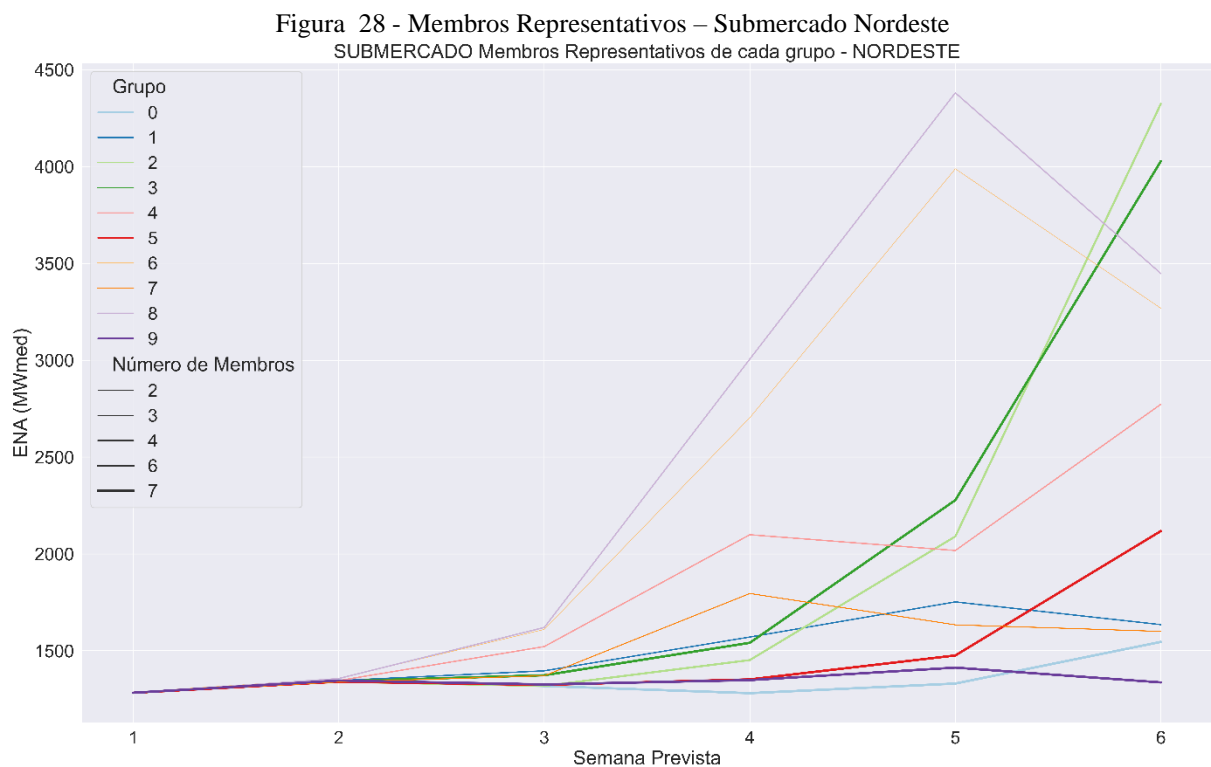
O SIN é composto por por 4 submercados principais, apresentados na Figura 27.

Figura 27 - Submercados do SIN



Fonte: Elaboração própria.

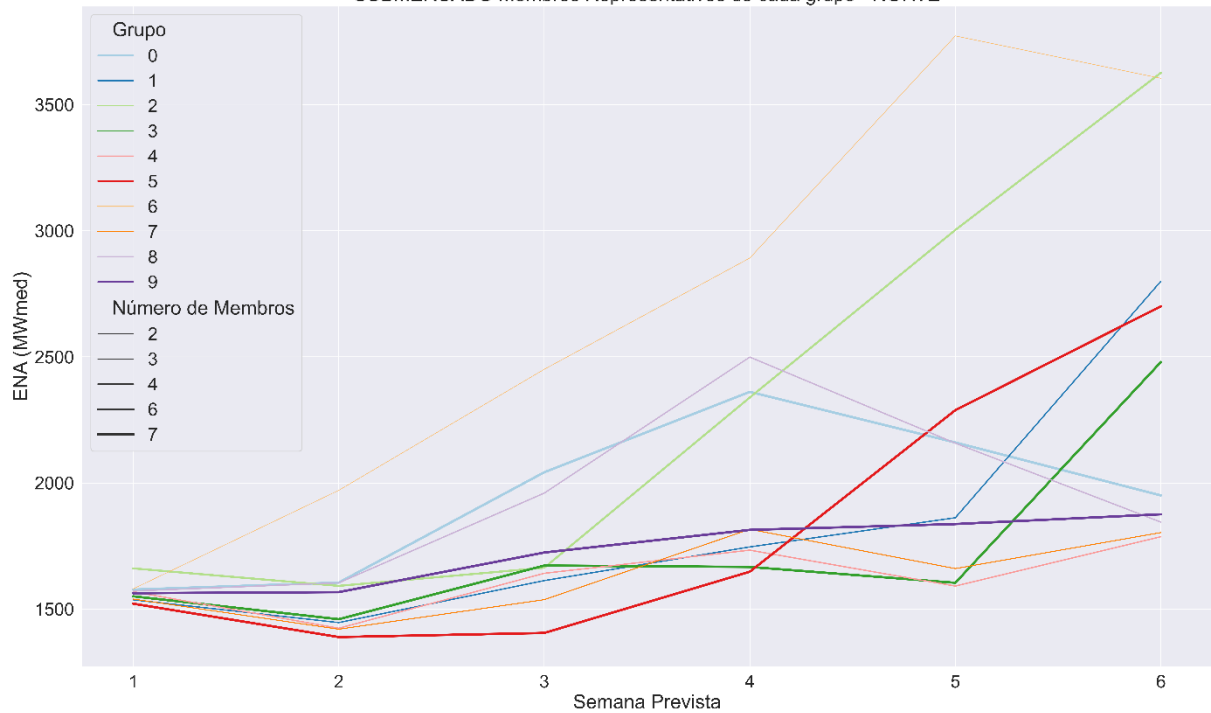
Na Figura 28 podemos observar a dispersão dos membros no submercado Nordeste. Os grupos só passaram a dispersarem entre si a partir da terceira semana, chegando na quinta semana com grupos variando entre 1 a 4,5 GWmed. Destaca-se aqui que nem todas as bacias que compõem o submercado Nordeste ocupam o mesmo espaço físico que a região Nordeste, que possui alta previsibilidade em modelos climáticos de MCGA, visto que a bacia do São Francisco possui afluentes desde o estado de Minas Gerais, e, portanto, é influenciada também pelo clima de região Sudeste.



Fonte: Elaboração própria.

A Figura 29 apresenta os membros representativos da análise de grupos no submercado Norte. A dispersão dos grupos ocorre já na segunda semana, com o grupo 1, 2 e 6 seguindo uma tendência de crescimento, enquanto os demais seguem estáveis ou decrescem.

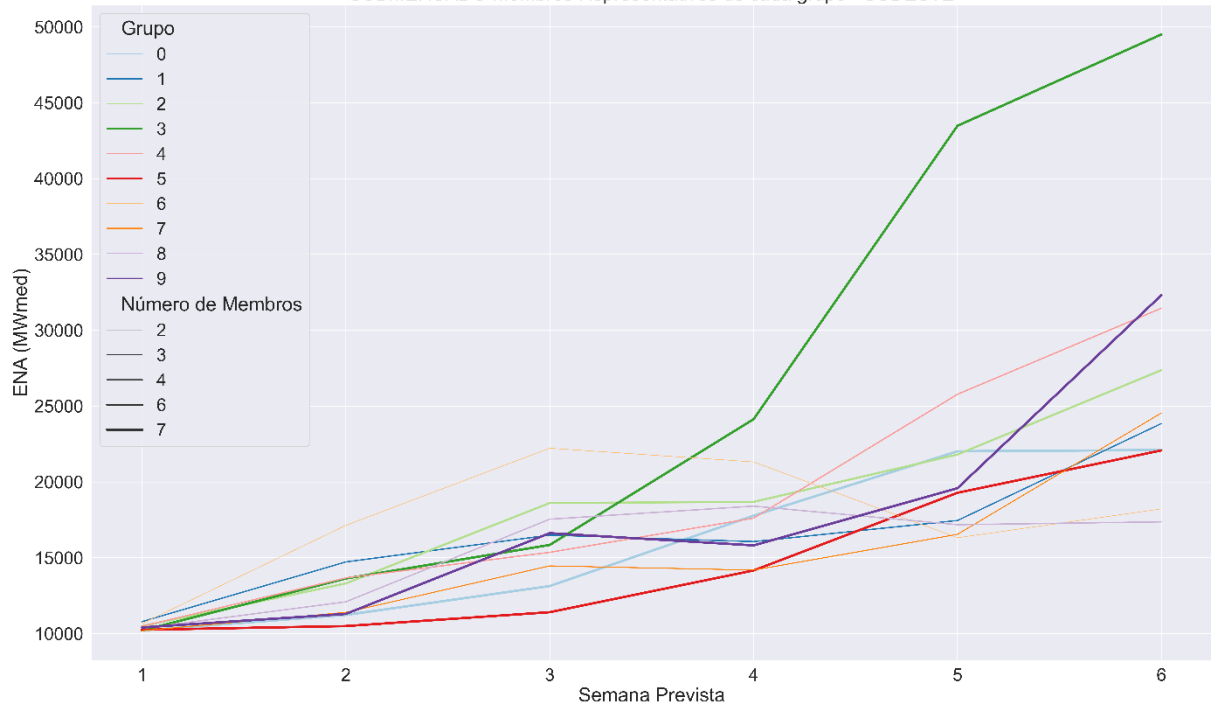
Figura 29 - Membros Representativos – Submercado Norte  
SUBMERCADO Membros Representativos de cada grupo - NORTE



Fonte: Elaboração própria.

A Figura 30 apresenta os membros representativos no submercado Sudeste. Em geral, todos demonstraram a mesma tendência de crescimento, com dispersões significativas sendo observadas por parte do grupo 6 já na segunda semana prevista. Destaca-se também a diferença na dispersão final (semana 6) entre os grupos, que vai da faixa de 1,7 GWmed à 5 GWmed.

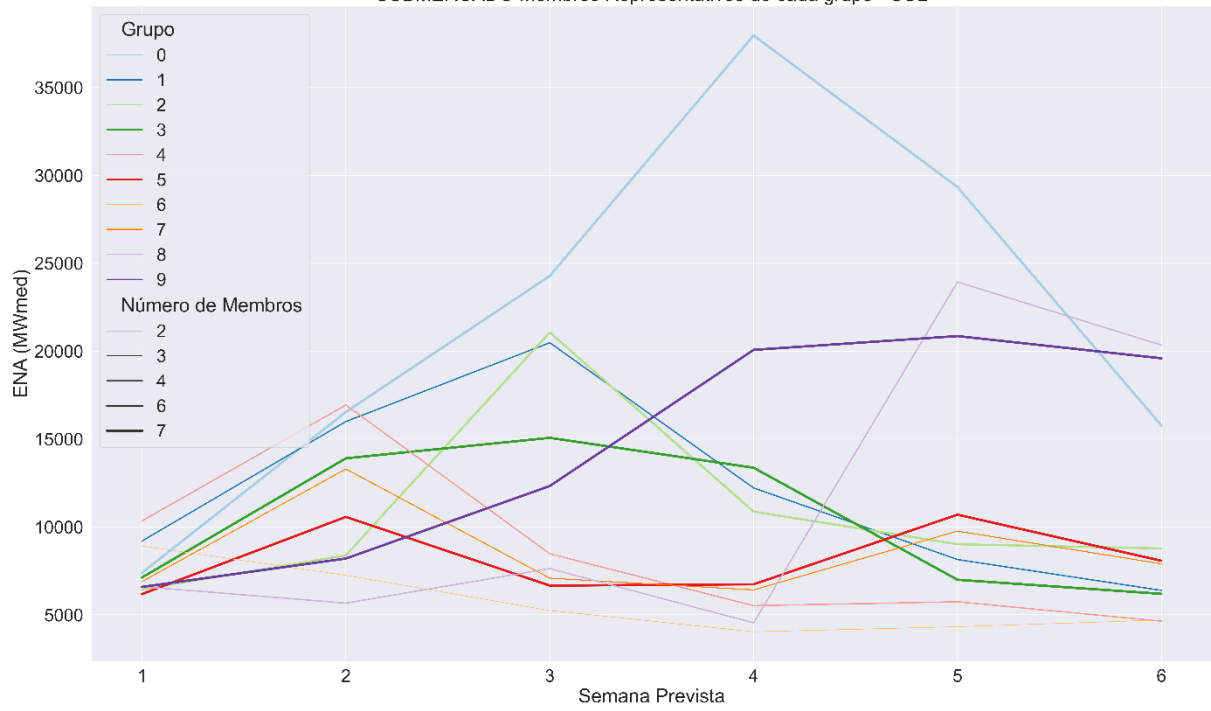
Figura 30 - Membros Representativos – Submercado Sudeste  
SUBMERCADO Membros Representativos de cada grupo - SUDESTE



Fonte: Elaboração própria.

No subsistema Sul (Figura 31), temos uma composição de bacias muito parecida com a da agregação Sul, e, portanto, os comportamentos dos grupos são semelhantes aos da Figura 31, com o grupo 0 apresentando um pico muito acima dos demais grupos na quarta semana prevista, e grande diferença de tendências entre os grupos ao longo das cinco semanas previstas. Apesar da variação apresentada ao longo das 6 semanas, destaca-se que ao fim das 6 semanas os grupos se concentram em duas faixas de ENA bem definidas.

Figura 31 - Membros Representativos – Submercado Sul  
SUBMERCADO Membros Representativos de cada grupo - SUL

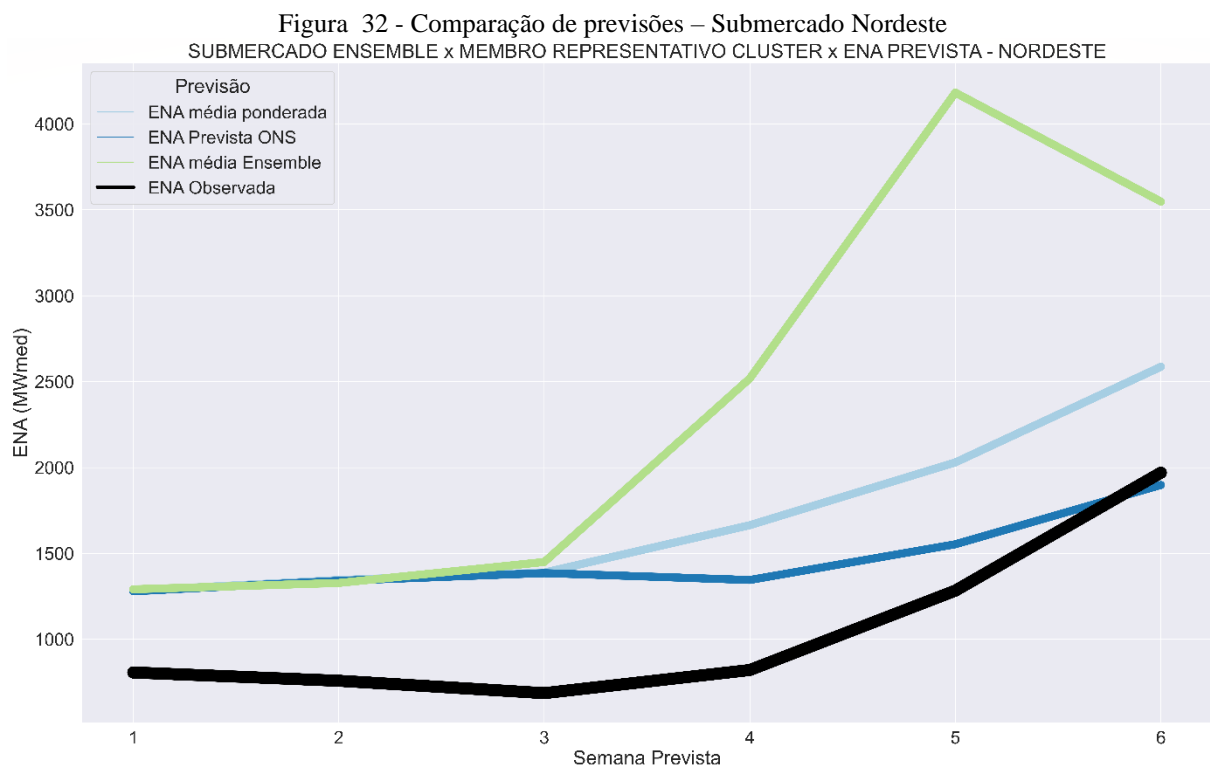


Fonte: Elaboração própria.



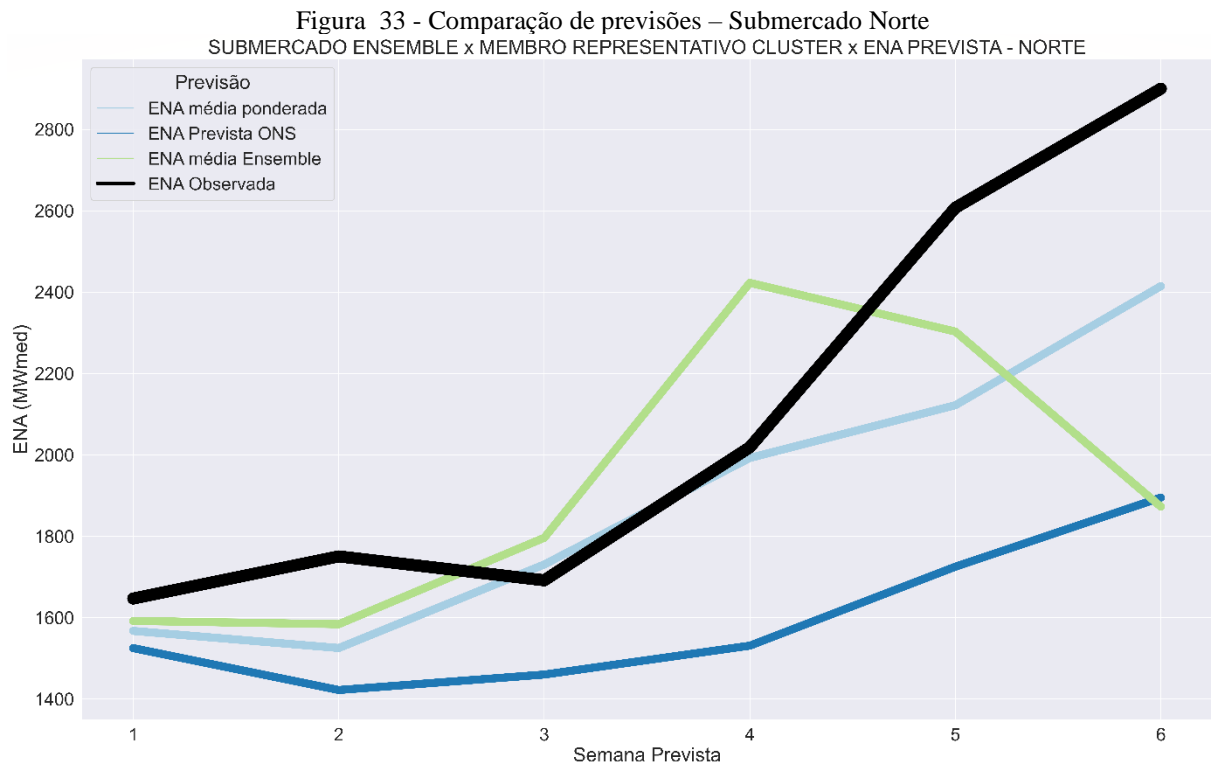
#### 4.6 COMPARAÇÃO COM OUTRAS PREVISÕES POR SUBMERCADO

Na Figura 32 apresenta-se uma a comparação entre ENA média Ensemble, ENA prevista ONS, ENA observada e ENA média ponderada, para o submercado Nordeste. Todas as previsões superestimaram a ENA observada ao longo das 6 semanas, sendo a ENA prevista ONS a que se manteve mais próxima do observado. A ENA média Ensemble registrou um crescimento a partir da terceira semana prevista à quinta semana prevista, que não se realizou. O mesmo comportamento foi observado ENA média ponderada, porém de forma atenuada.



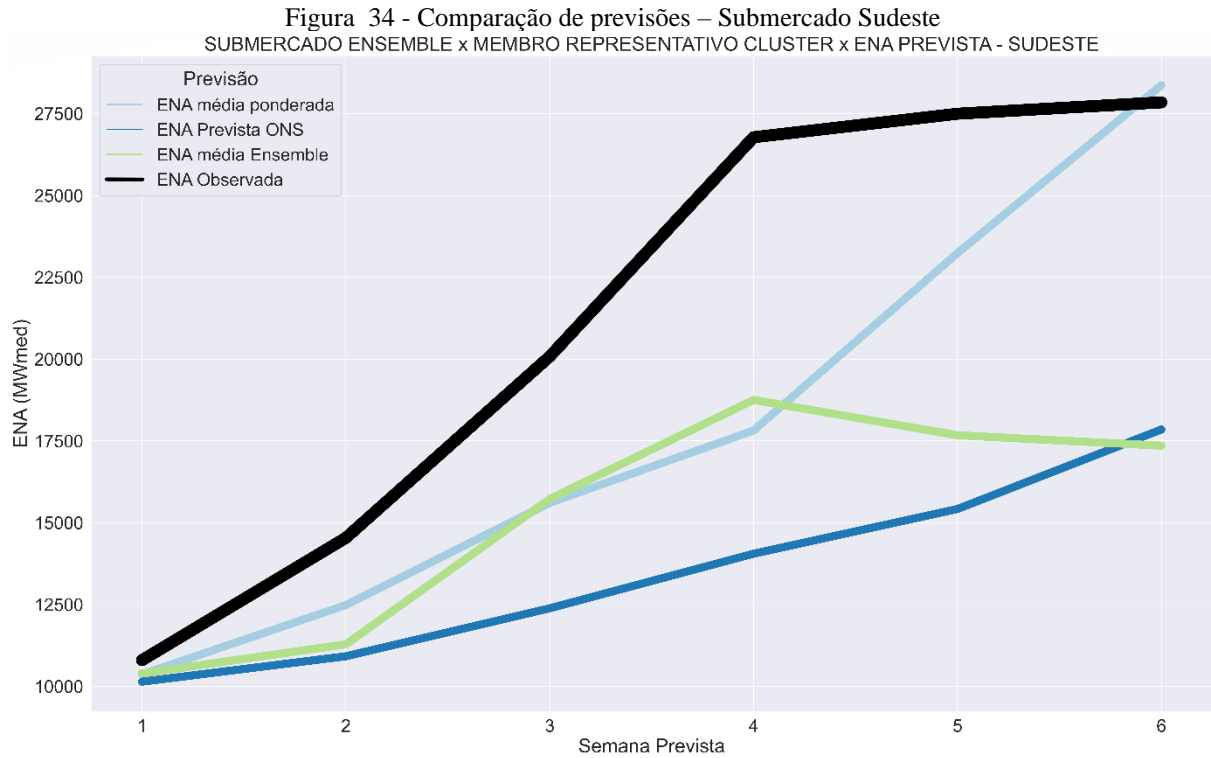
Fonte: Elaboração própria.

A Figura 33 apresenta a comparação de previsões no submercado Norte. De modo geral, todas as previsões tiveram uma tendência de crescimento, acompanhando a ENA observada, com exceção da ENA prevista ONS, que subestimou o observado durante as 5 semanas previstas. A ENA média Ensemble se distanciou do observado a partir da quarta semana, quando registrou um crescimento de mais de 0,4 GWmed, e, após isso, uma queda até a sexta semana prevista, que não se consolidou. A ENA média ponderada se manteve aderente ao observado até a quinta semana prevista, onde não apresentou um pico na sexta semana, apesar de capturar o crescimento com a mesma intensidade que o observado na sexta semana prevista.



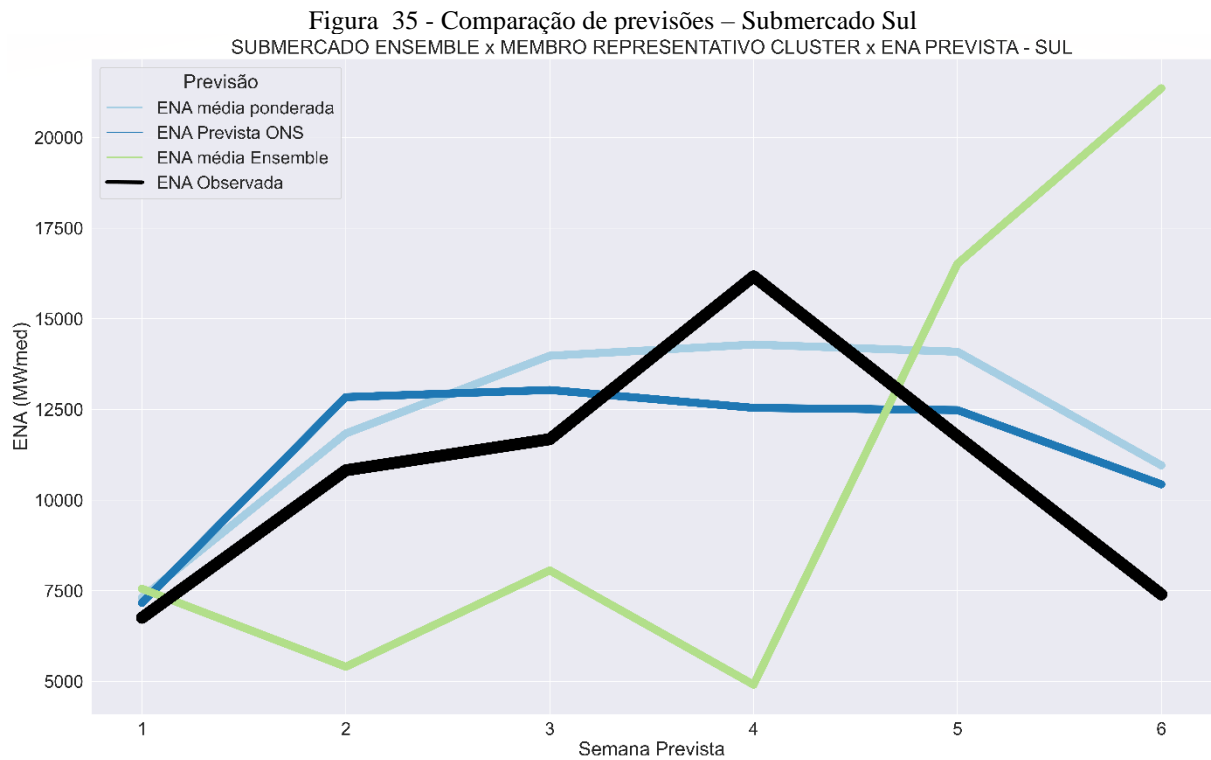
Fonte: Elaboração própria.

A Figura 34 apresenta a comparação de previsões no submercado Sudeste. Todas as previsões subestimaram a ENA observada, em pelo menos 5 das 6 semanas previstas. A previsão que mais se aproximou do observado foi a realizada ENA média ponderada, que sempre acompanhou a tendência do observado, porém com intensidades diferentes, principalmente na terceira e quarta semana prevista, onde não cresceu com a mesma intensidade. A ENA média Ensemble acompanhou a tendência do observado até a quarta semana prevista, onde entra em recessão que não se realizou. A ENA prevista ONS segue a tendência de crescimento do observado, porém com intensidade muito inferior.



Fonte: Elaboração própria.

No submercado Sul (Figura 35), a ENA média Ensemble apresenta tendências e intensidades completamente distintas do observado. A ENA prevista ONS e ENA média ponderada andam próximas, entretanto, a ENA prevista ONS não capturou com a mesma intensidade o pico na quarta semana, ficando mais de 2 GWmed abaixo do consolidado.



Fonte: Elaboração própria.

#### 4.7 DISPERSÃO DOS MEMBROS EM RELAÇÃO À MLT

A Figura 36 apresenta a dispersão que os membros possuem entre si ao longo das semanas previstas, em relação à MLT de cada submercado. Nesta seção, ao nos referirmos aos valores de chuva prevista, estamos nos referindo aos valores acumulados da média de precipitação da previsão por conjuntos *Extended-Range*, apresentada na Figura 29.

Vemos que na semana 1 os grupos não apresentam dispersão entre si nos submercados Norte, Nordeste e Sudeste, onde ocorreu chuva moderada na semana 1, entre 15 e 20 mm para a maioria das sub-bacias do SIN. No submercado Sul, é observada grande variação entre os grupos já na primeira semana prevista, onde a maioria das sub-bacias pertencentes ao submercado apresentaram chuva acumulada na faixa de 30 a 40mm.

Na semana 2 os grupos do submercado Sul apresentam ainda mais dispersão, enquanto as bacias desse subsistema passam a registrar até 75 mm acumulados durante a segunda semana. Os submercados Nordeste e Norte continuam sem diferença maior que 20% da MLT entre os grupos. O submercado Sudeste passa a apresentar variações entre os grupos, enquanto a bacia de Paranapanema passa a receber também uma chuva acumulada de 75 mm semanais, e as demais veem um aumento médio de 10 mm na chuva acumulada semanal, em relação à semana anterior.

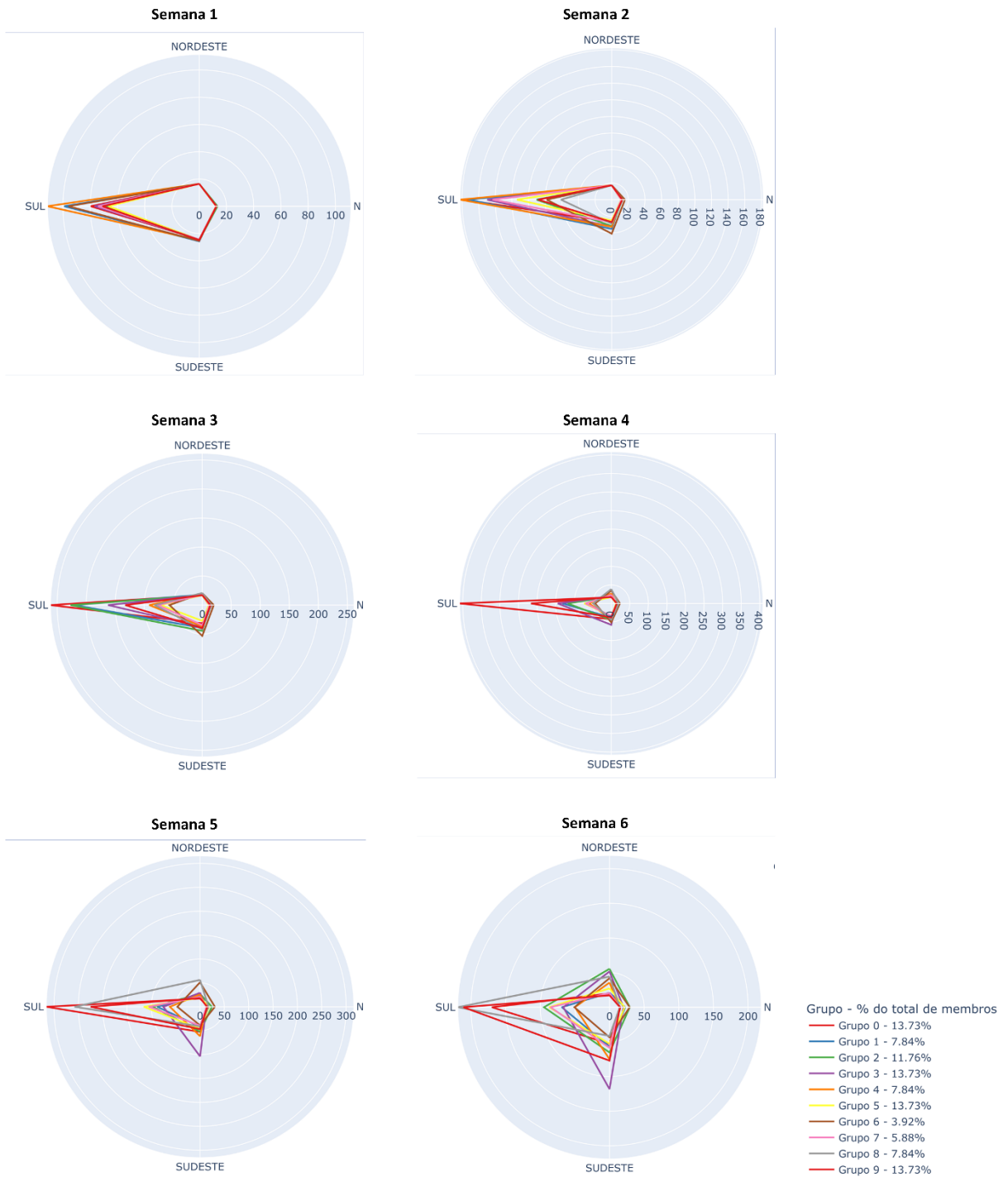
Durante a semana 3, os submercados apresentam dispersão entre os grupos muito similar ao registrado na semana 2, enquanto a grande maioria do Sudeste passa a receber uma chuva acumulada semanal de até 75 mm.

Na semana 4, os submercados Norte, Nordeste e Sudeste voltam a apresentar pouca dispersão entre os membros. No submercado Sul, o grupo 0 passa a apresentar uma ENA extremamente elevada, na faixa de 400% da MLT, enquanto os demais grupos mantêm os valores e dispersões semelhantes entre si, em relação à semana 3.

A partir da quinta semana, passamos a observar dispersão entre os membros em todos os submercados. No submercado Sul os grupos 9 e 8 se juntam ao grupo 0, e apresentam porcentagens da MLT maiores que 200%. No submercado Sudeste, o grupo 3 passa a apresentar mais de 100% da MLT, e no submercado Nordeste, os grupos 6 e 8 também chegam à esses valores. No submercado Norte a dispersão entre os grupos também existe, mas é menor, com nenhum grupo apresentando mais que 100% da MLT.

Finalmente, na semana 6, as dispersões são bruscas e são observadas em todos os submercados. A dispersão no submercado Sul se mantém semelhante à semana 5. A dispersão no submercado Sudeste mais é maior, com o grupo 3 chegando na faixa de 120% da MLT e os demais grupos ultrapassando a faixa de 50%. No submercado Norte, o grupo 2 chega à faixa de 25% da MLT, juntamente com o grupo 6 que já estava nesta faixa desde a semana 5. No submercado Nordeste, mais da metade dos grupos ultrapassam a faixa de 25% da MLT, chegando a até 50%.

Figura 36 - %MLT por agrupamento

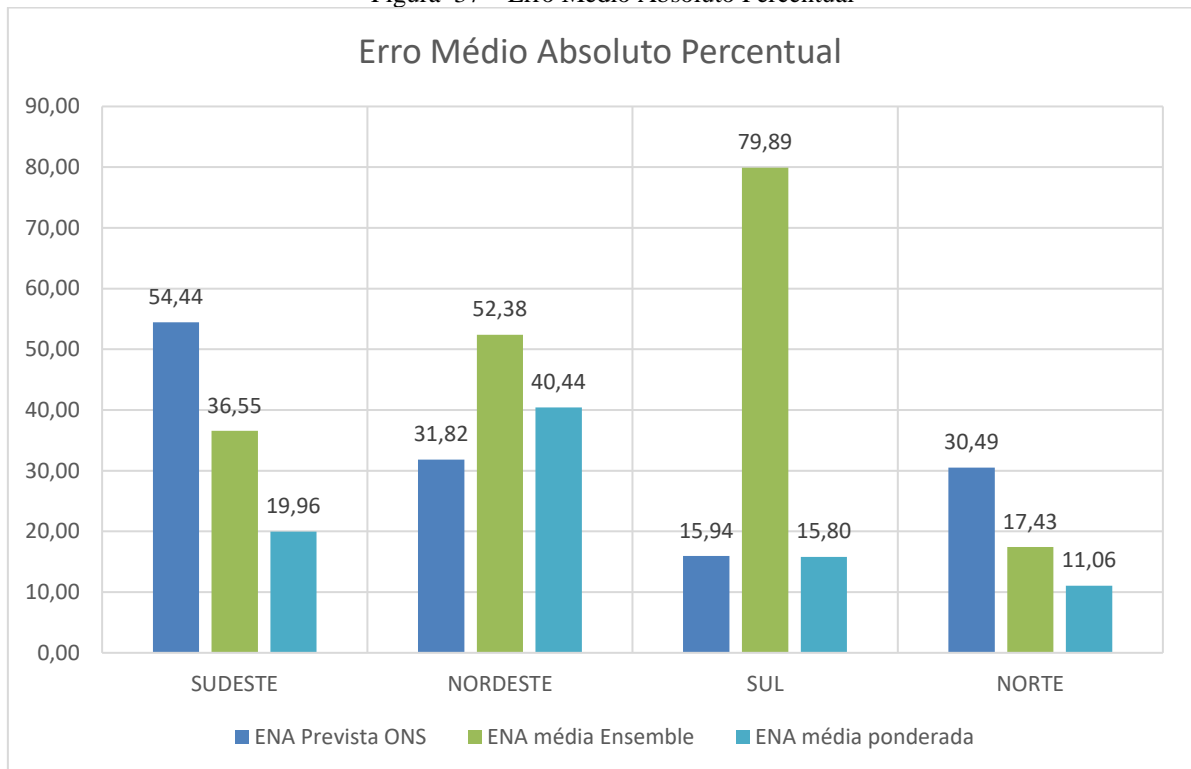


Fonte: Elaboração própria.

#### 4.8 MÉTRICAS

A Figura 37 apresenta os valores de erro médio absoluto percentual calculado entre cada previsão e a ENA observada.

Figura 37 – Erro Médio Absoluto Percentual

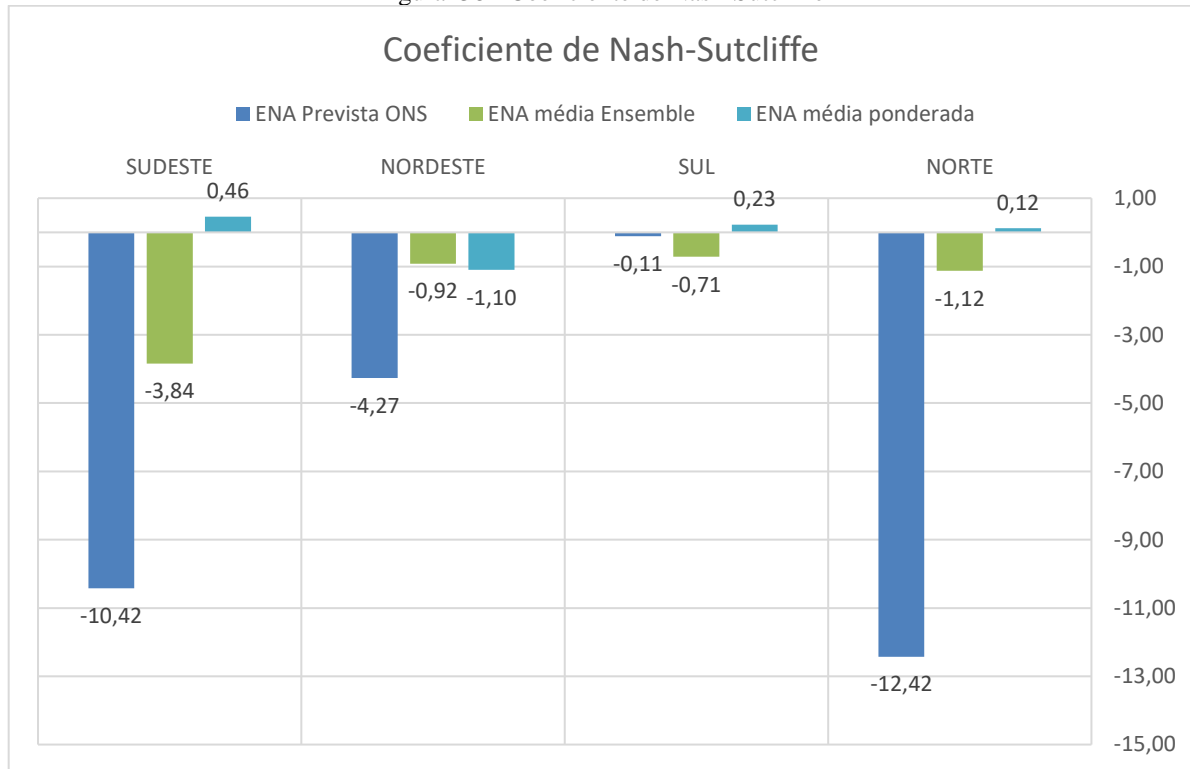


Fonte: Elaboração própria.

Observamos que a ENA média ponderada é a previsão com menor MAPE em todos os submercados, com exceção do Nordeste. Na região Nordeste, a diferença entre ENA Prevista ONS e ENA média ponderada foi de 12%, entretanto, tal submercado, neste estudo do caso, é o de menor impacto em relação à ENA total do SIN. A previsão ENA média Ensemble apresentou previsões melhores que as oficiais nos submercados Sudeste e Norte, porém, obteve um erro de quase 80% no submercado Sul.

A Figura 38 apresenta o coeficiente de Nash-Sutcliffe calculado para cada previsão, em relação à ENA observada.

Figura 38 - Coeficiente de Nash-Sutcliffe



Fonte: Elaboração própria.

Observamos que, novamente, a ENA média ponderada obtém os melhores valores dentre as previsões em todos os submercados, com exceção do Nordeste. A ENA média Ensemble obtém os melhores resultados no submercado Nordeste, e os piores no submercado Sul. A ENA prevista ONS obteve os piores valores no submercado Sudeste, Nordeste e Norte, com destaque para Sudeste e Norte, onde o coeficiente chegou a valores menores que 10 negativo. Valores negativos nos indicam que as somas dos erros residuais do modelo são maiores que o erro quadrático total (soma das diferenças entre os valores observados e sua média). A soma do erro quadrático total nos dá uma noção da variação natural da série observada em torno de sua média. Portanto, valores negativos do coeficiente nos mostram que erros entre o modelo de previsão e o observado são maiores do que a variação natural da série observada. Quanto maior a magnitude deste valor negativo, maiores são os erros entre o modelo e o observado.

Ressalta-se que apesar da maior assertividade em relação à ENA prevista ONS, a ENA média ponderada não apresentou valores satisfatórios nos índices calculados. Recomenda-se estudos futuros com a aplicação da metodologia proposta em outros períodos do ano e em outros cenários hidrológicos, além da transição seco-molhada estudada neste trabalho. Variações das



agregações temporais e espaciais também devem ser consideradas, visto que informações de variações de ENA se perdem nas médias semanais e nas bacias agregadas.

## 5 CONCLUSÕES

O algoritmo *K-Means* foi aplicado sobre cenários de ENA gerados a partir do produto *Extended-Range* do ECMWF, do dia 23 de Setembro de 2021, agregado temporalmente e espacialmente, em duas regiões distintas, variando os números de grupos de 2 a 10, conforme metodologia proposta pelo ONS em estudos de análise de agrupamento (ONS, 2022a). Verificou-se através do *Elbow Method* que o número ideal de grupos para o estudo de caso foi igual a 10.

As previsões de ENA média ponderada se mostraram mais assertivas que a previsão oficial do ONS nos submercados Sudeste, Sul e Norte, com coeficiente de Nash-Sutcliffe de 0.46, 0.23 e 0.12, respectivamente. Também foram registrados menores erros médios absolutos percentuais, em relação à outras previsões, nessas mesmas regiões.

No submercado Nordeste, a previsão ENA média Ensemble obteve o melhor índice de Nash-Sutcliffe, de -0.92. Entretanto, também obteve o maior erro médio percentual absoluto, registrando um erro 21% maior que a previsão oficial do ONS.

As previsões de ENA geradas a partir da média ponderada dos membros representativos de cada grupo, com base no número de membros associados ao grupo, conseguiu capturar o pico da ENA observada, o qual não foi mostrado pela previsão oficial utilizada pelo ONS no PMO de Outubro de 2021. Tal previsão se mostrou mais próxima da ENA posteriormente observada em três dos quatro submercados do SIN, superestimando apenas o submercado Nordeste da quarta à sexta semana. Entretanto, tal submercado é o menos impactante no SIN, registrando apenas cerca de 1 GWmed ao longo das seis semanas previstas analisadas.

A previsão gerada a partir da média dos membros da previsão conjunta capturou parte do pico de ENA e vazão nos submercados Norte, Nordeste e Sudeste. Entretanto, apresentou um comportamento totalmente distinto das outras previsões e da ENA observada no submercado Sul. No submercado Nordeste, o pico de ENA foi superestimado em relação à ENA observada a partir da quarta semana prevista, e, no submercado Sudeste, subestimado.

O subsistema Sul se mostrou mais caótico que os demais, apresentando dispersão entre seus grupos desde a primeira semana prevista, apesar de receber, em média, a mesma quantidade de chuva que outros submercados. Tal fato pode ser relacionado às características

das sub-bacias que compõem o submercado Sul, ou à maior ocorrência de eventos extremos nos membros nessa região, durante o período analisado.

Com exceção do submercado Sul, verificou-se que todos os submercados começam a apresentar uma dispersão mais brusca entre os grupos a partir da quinta semana prevista.

Recomenda-se o estudo futuro de outros métodos para escolha do número ideal de grupos, como o a verificação das distâncias dos grupos através do Silhouette Score, e a variação das agregações espaciais e temporais, visto que as métricas calculadas para a ENA média ponderada não se demonstraram satisfatórias, mesmo que superiores às da previsão oficial por parte do ONS. Para a agregação temporal, uma análise exploratória preliminar pode ser efetuada, agrupando as bacias do SIN com base em suas respectivas evoluções temporais de ENA.

## 6 REFERÊNCIAS

AKSOY, S.; HARALICK, R. M. Feature normalization and likelihood-based similarity measures for image retrieval. **Pattern Recognition Letters**, v. 22, n. 5, p. 563–582, 2001.

ANEEL. **OUTORGA**. Disponível em: <<https://www.gov.br/aneel/pt-br/assuntos/geracao/outorgas>>. Acesso em: 27 out. 2022.

ARRUDA, N.; SIECIECHOWICZ, M.; BECKER, A. C. Disponibilidade Hídrica (Water Availability). Em: **Gestão de água em complexo de geração termelétrica**. [s.l.: s.n.]. p. 40–59.

ÁVILA, L. et al. Comparative Evaluation of Five Hydrological Models in a Large-Scale and Tropical River Basin. **Water**, v. 14, n. 19, p. 3013, 25 set. 2022.

AYOADE, J. O. **Introdução à Climatologia para os Trópicos**. 4. ed. [s.l.: s.n.]. v. 4

BRAVO, J. M. et al. Incorporating Forecasts of Rainfall in Two Hydrologic Models Used for Medium-Range Streamflow Forecasting. **JOURNAL OF HYDROLOGIC ENGINEERING**, v. 14, p. 435–445, 2009.

CHARNEY, J. G.; SHUKLA, J. Predictability of monsoons. Em: LIDTHILL, J.; PEARCE, R. P. (Eds.). **Monsoon Dynamics**. Cambridge: Cambridge University Press, 1981. p. 99–110.

CUO, L.; PAGANO, T. C.; WANG, Q. J. **A review of quantitative precipitation forecasts and their use in short- to medium-range streamflow forecasting**. **Journal of Hydrometeorology**, out. 2011.

DA SILVA, E. L. **Formação de preços em mercados de energia elétrica**. 1. ed. Porto Alegre: Sagra Luzzatto, 2001.

DAVIE, T. **Fundamentals of Hydrology, Second Edition**. 3. ed. Nova Iorque: [s.n.].

DE MYTTENAERE, A. et al. Mean Absolute Percentage Error for regression models. **Neurocomputing**, v. 192, p. 38–48, 2016.

ECMWF. **Global numerical modelling at the heart of ECMWF's forecasts**. Disponível em: <<https://www.ecmwf.int/en/about/media-centre/focus/2022/global-numerical-modelling-heart-ecmwfs-forecasts>>. Acesso em: 13 out. 2022a.

ECMWF. **6.1 Medium Range Forecasting - ENS alone**. Disponível em: <<https://confluence.ecmwf.int/display/FUG/6.1+Medium+Range+Forecasting+-+ENS+alone>>. Acesso em: 30 nov. 2022b.

ECMWF. **Extended Range Predictability**. Disponível em: <<https://confluence.ecmwf.int/display/FUG/Extended+Range+Predictability>>. Acesso em: 30 nov. 2022c.

ECMWF. **The ECMWF Ensemble Prediction System - The rationale behind probabilistic weather forecasts**. [s.l: s.n.]. Disponível em: <<https://confluence.ecmwf.int/display/FUG/2+The+ECMWF+Integrated+Forecasting+System+-+IFS>>. Acesso em: 25 nov. 2022d.

ECMWF. **HRES - High-Resolution Forecast**. Disponível em: <<https://confluence.ecmwf.int/display/FUG/HRES+-+High-Resolution+Forecast>>. Acesso em: 25 nov. 2022e.

ECMWF. **ENS - Ensemble Forecasts**. Disponível em: <<https://confluence.ecmwf.int/display/FUG/ENS+-+Ensemble+Forecasts>>. Acesso em: 25 nov. 2022f.

ECMWF. **Extended-Range Forecast**. Disponível em: <<https://confluence.ecmwf.int/display/FUG/Extended-Range+Forecast>>. Acesso em: 25 nov. 2022g.

ECMWF. **Long-Range (Seasonal) Forecast**. Disponível em: <[Long-Range \(Seasonal\) Forecast](#)>. Acesso em: 25 nov. 2022h.

ECMWF. **Extended Range Output - Extended ENS**. Disponível em: <<https://confluence.ecmwf.int/display/FUG/8.2+Extended+Range+Output+-+Extended+ENS>>. Acesso em: 25 nov. 2022i.

FRALEY, C.; RAFTERY, A. E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. **The Computer Journal**, v. 41, n. 8, p. 578–588, 1998.

FREDO, G. **ANÁLISE DE DIFERENTES REPRESENTAÇÕES DA FUNÇÃO DE PRODUÇÃO HIDRELÉTRICA NO PROBLEMA DE PLANEJAMENTO DA OPERAÇÃO ENERGÉTICA DE MÉDIO PRAZO**. Florianópolis: Universidade Federal de Santa Catarina, 2016.

FUGIMOTO, S. **ESTRUTURA DE TARIFAS DE ENERGIA ELÉTRICA ANÁLISE CRÍTICA E PROPOSIÇÕES METODOLÓGICAS**. São Paulo: USP, 2010.

GOMES, A. C. et al. BNDES 50 anos - O SETOR ELÉTRICO. Em: **Banco Nacional de Desenvolvimento Econômico e Social 50 anos: histórias setoriais**. Rio de Janeiro: Dbá, 2002. p. 321–347.

HOLMES, C.; LINSEY, R. **Digital Simulation in Hydrology**. Stanford: Department of Civil Engineering, 1966.

KODINARIYA, T. M.; MAKWANA, P. R. Review on determining number of Cluster in K-Means Clustering. **International Journal of Advance Research in Computer Science and Management Studies**, v. 1, n. 6, 2013.

LEUTBECHER, M.; PALMER, T. N. Ensemble Forecasting. **Journal of Computer Physics**, v. 227, n. 7, p. 3515–3539, 2008.

LOPES, J. E. G.; BRAGA, B. P. F. JR.; CONEJO, J. G. L. **SMAP--a simplified hydrologic model. Applied modeling in catchment hydrology / ed. by V.P. Singh** Littleton, Colo. : Water Resources Publications, 1982., , 1982.

LORENZ, E. N. A study of the predictability of a 28-variable atmospheric model. **Tellus**, v. 17, n. 3, p. 321–333, 1965.

MATOS, V. L. **UM MODELO PARA O PLANEJAMENTO ANUAL DA OPERAÇÃO ENERGÉTICA CONSIDERANDO TÉCNICAS AVANÇADAS DE OTIMIZAÇÃO ESTOCÁSTICA**. Florianópolis: Universidade Federal de Santa Catarina, 2012.

MENDONÇA; FRANCISCO; OLIVEIRA-DANNI. INÊS MORESCO. **Climatologia: noções básicas e climas no Brasil**. São Paulo: Oficina de Textos, 2007.

MILLIGAN, G. W.; COOPER, M. C. Methodology Review: Clustering Methods. **Applied Psychological Measurement**, v. 11, n. 4, p. 329–354, 1987.

MIUIGAN, G. W. et al. A Study of Standardization of Variables in Cluster Analysis. **Journal of Classification**, v. 5, p. 181–204, 1988.

MOHAMAD, I. BIN; USMAN, D. Standardization and its effects on K-means clustering algorithm. **Research Journal of Applied Sciences, Engineering and Technology**, v. 6, n. 17, p. 3299–3303, 2013.

NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. **Biometrika**, v. 78, n. 3, p. 691–692, 1 set. 1991.

NAINGGOLAN, R. et al. **Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method**. Journal of Physics: Conference Series. **Anais...**Institute of Physics Publishing, 18 dez. 2019.

NG, A. Clustering with the k-means algorithm. **Machine Learning**, p. 1–2, 2012.

OLIVEIRA, M. **Energia elétrica: geração, transmissão e sistemas interligados**. 1. ed. [s.l.] LTC, 2014.

ONS. **Submódulo 23.5 Critérios para estudos hidrológicos**. [s.l: s.n.]. Disponível em: <<http://www.ons.org.br>>.

ONS. **MANUAL DE METODOLOGIA - APLICATIVO SMAP**. [s.l: s.n.]. Disponível em: <<http://www.ons.org.br>>. Acesso em: 13 out. 2022.

ONS. **METODOLOGIA DE RECONSTITUIÇÃO E TRATAMENTO DAS VAZÕES NATURAIS**. [s.l: s.n.].

ONS. **Dados Técnicos do SIN**. Disponível em: <<http://www.ons.org.br>>. Acesso em: 13 out. 2022a.

ONS. **PRECIPITAÇÃO OBTIDA PELA COMBINAÇÃO ENTRE DADOS ESTIMADOS POR SATÉLITE E DADOS OBSERVADOS DE ESTAÇÕES METEOROLÓGICAS**. [s.l: s.n.].

ONS. **APRIMORAMENTO DA PREVISÃO DE PRECIPITAÇÃO POR CONJUNTO E REMOÇÃO DE VIÉS ATÉ A SEGUNDA SEMANA OPERATIVA**. [s.l: s.n.].

ONS. **Resultados finais consistidos (Vazões Semanais - PMO)**. Disponível em: <[https://sintgre.ons.org.br/sites/9/13/79/Produtos/244/Definitivo\\_202110\\_PMO.zip](https://sintgre.ons.org.br/sites/9/13/79/Produtos/244/Definitivo_202110_PMO.zip)>. Acesso em: 20 dez. 2022.

ONS. **PREVISÃO DE PRECIPITAÇÃO PARA O PRIMEIRO MÊS DA OPERAÇÃO - Clusterização - Parte 2**. Disponível em: <<https://ctpmopld.org.br/web/ct-pmo-pld>>. Acesso em: 13 out. 2022a.

ONS. **O Sistema em números**. Disponível em: <[ons.org.br](http://ons.org.br)>. Acesso em: 13 out. 2022b.

PAIVA, L. **INTERFERÊNCIA DA PREVISÃO CLIMÁTICA NA PREVISÃO DE VAZÕES E SUAS CONSEQUÊNCIAS PARA A OPERAÇÃO E PLANEJAMENTO DO SISTEMA INTERLIGADO NACIONAL: aplicação na Bacia do Rio São Francisco**. Tese—Recife: Universidade Federal de Pernambuco, 2020.

PASSAIA, O. A. et al. Impact of large reservoirs on simulated discharges of brazilian rivers. **Revista Brasileira de Recursos Hídricos**, v. 25, p. 1–9, 2020.

PAULO, S. **EVOLUÇÃO DO SETOR ELÉTRICO BRASILEIRO NO CONTEXTO ECONÔMICO NACIONAL: UMA ANÁLISE HISTÓRICA E ECONOMETRICA DE LONGO PRAZO**. São Paulo: USP, 2011.

REIS, L. B. **Geração de Energia Elétrica**. 2. ed. São Paulo: Manole, 2000. v. 2

ROWELL, D. P. Assessing Potential Seasonal Predictability with an Ensemble of Multidecadal GCM Simulations. **Journal of Climate**, v. 11, n. 2, p. 109–120, 1998.

SAMPAIO, G.; DIAS, P. L. DA S. Evolução dos Modelos Climáticos e de Previsão de Tempo e Clima. **Revista USP**, n. 103, p. 41–54, nov. 2014.

SAXENA, A. et al. A review of clustering techniques and developments. **Neurocomputing**, v. 267, p. 664–681, 6 dez. 2017.

SCIKIT-LEARN. **Selecting the number of clusters with silhouette analysis on KMeans clustering.** Disponível em: <[https://scikit-learn.org/stable/auto\\_examples/cluster/plot\\_kmeans\\_silhouette\\_analysis.html](https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)>. Acesso em: 26 nov. 2022.

SENARATHNA, S.; HEMAPALA, K. T. M. U. Optimized Adaptive Overcurrent Protection Using Hybridized Nature-Inspired Algorithm and Clustering in Microgrids. **Energies**, v. 13, p. 3324, 30 jun. 2020.

SILVA, F. W. M. **PROJEÇÃO DE ENERGIA NATURAL AFLUENTE (ENA) A MÉDIO PRAZO PARA O SETOR ELÉTRICO DO BRASIL.** Redenção: UNILAB, 2017.

SINAGA, K. P.; YANG, M. S. Unsupervised K-means clustering algorithm. **IEEE Access**, v. 8, p. 80716–80727, 2020.

TORRES, F. T. P.; MACHADO, P. J. D. O. **INTRODUÇÃO À CLIMATOLOGIA.** [s.l: s.n.].

VASCONSELOS, F. **Geração, transmissão e distribuição de energia elétrica.** Londrina: Editora e Distribuidora Educacional S.A., 2017.

YANG, X.; HOSSEIN GANDOMI, A. Bat algorithm: a novel approach for global engineering optimization. **Engineering Computations**, v. 29, n. 5, p. 464–483, 1 jan. 2012.

YUAN, C.; YANG, H. Research on K-Value Selection Method of K-Means Clustering Algorithm. **J**, v. 2, n. 2, p. 226–235, 2019.

ZANATTA, I. F. et al. **Introdução à Climatologia.** 1. ed. Londrina: Editora e Distribuidora Educacional SA, 2016.

## 7 ANEXO A – CÓDIGO BASE DO ALGORITMO K-MEANS

```

import pandas as pd
import numpy as np
import os
from scipy.stats import zscore
from numpy import random
from numpy import median

import datetime

%% PRÉ PROCESSAMENTO DOS DADOS
%% CRIA K MEANS

k = 10

class ShapeException(BaseException):
    pass

class MatrixKMeans:
    def __init__(self, data: list, n_centroids: int = k):
        self.n_centroids = n_centroids
        self.data = data
        self.centroids = None
        self.data_current_centroid = np.zeros(len(data))

    def generate_random_centroids(self) -> None:
        self.centroids = []
        mean_centroid = sum(self.data) / len(self.data)
        for i in range(self.n_centroids):
            random_noise = np.random.normal(0, 1, size=self.data[0].shape)
            # random_noise = np.random.random(size =
self.data[0].shape)*1000

            self.centroids.append(mean_centroid + random_noise)
            # self.centroids.append(self.data[i])

    def assign_items_to_centroids(self):
        for i in range(len(self.data)):
            distance_from_centroid = np.zeros(len(self.centroids))
            for j in range(len(self.centroids)):
                distance_from_centroid[j] =
self.absolute_distance(self.data[i], self.centroids[j])
                # print(distance_from_centroid)
                # breakpoint()
            self.data_current_centroid[i] =
np.argmin(distance_from_centroid)
            # print(self.data_current_centroid)
            # breakpoint()

    def update_centroids_positions(self, tolerance: float = 0.01) -> bool:
        same_position = 0
        for i in range(len(self.centroids)):
            new_centroid_position = np.zeros(shape=self.data[0].shape)
            divisors = 0
            for j in range(len(self.data)):
                if self.data_current_centroid[j] == i:
                    new_centroid_position += self.data[j]
                    divisors += 1

```



```

        if divisors == 0:
            same_position += 1
            continue

        new_centroid_position /= divisors
        if self.absolute_distance(self.centroids[i],
new_centroid_position) < tolerance:
            same_position += 1
            self.centroids[i] = new_centroid_position
            # print(len(self.centroids))
            # print(same_position)
        return same_position == len(self.centroids)

    def calculate_centroids_positions(self, max_iterations: int = 5000,
verbose=False):
        self.generate_random_centroids()
        for i in range(max_iterations):
            self.assign_items_to_centroids()
            if self.update_centroids_positions():
                break

    @staticmethod
    def absolute_distance(a: np.array, b: np.array) -> float:
        if not a.shape == b.shape:
            raise ShapeException("Matrizes de tamanhos incompatíveis")
        abs_distance = float(np.sum(np.abs(a - b)))
        return abs_distance

    def evaluate(self) -> float:
        mean = 0
        for i in range(len(self.data)):
            mean += self.absolute_distance(self.data[i],
self.centroids[int(self.data_current_centroid[i])])
        return mean

    def representative_member(self, n_clusters):

        tmp_cluster = pd.DataFrame()
        for k in range(0, n_clusters):

            centroid_index_position =
pd.DataFrame(self.data_current_centroid, columns =
['cluster']).reset_index()
            centroid_index_position =
centroid_index_position[centroid_index_position['cluster'] == k]
            list_centroid_index_position =
list(centroid_index_position['index']).to_list()

            tmp_k = pd.DataFrame()
            for position in list_centroid_index_position:
                error = self.absolute_distance(self.data[position],
self.centroids[int(self.data_current_centroid[position])])
                d = {'erro': error, 'index membro': position, 'Grupo': k}
                tmp = pd.DataFrame([d])

            tmp_k = pd.concat([tmp_k, tmp])

        tmp_cluster = pd.concat([tmp_cluster, tmp_k])

```

```

        tmp_cluster = tmp_cluster.groupby(['Grupo'], as_index =
False).min()

        return tmp_cluster

def load_members_zscore(members_path: str = "./Datasets/membros por
região") -> list:
    files = os.listdir(members_path)
    members_list = []
    for file in files:
        members_list.append(pd.read_excel(members_path + "/" +
file).drop("Unnamed: 0", axis=1).replace(',','.', regex=
True).astype(float).apply(zscore).values)
    return members_list

def main(members_list,k):
    k_means = MatrixKMeans(members_list, n_centroids=k)
    k_means.calculate_centroids_positions(verbose=True)

    cluster_final = k_means.data_current_centroid
    erro_final = k_means.evaluate()

    representative_members = k_means.representative_member(k)

    return cluster_final, erro_final, representative_members;

### APLICAÇÃO K MEANS

import glob
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.ticker as ticker
from collections import Counter

members_list = load_members_zscore()

df_elbow = pd.DataFrame()
for k in range(2,11):
    i = 0
    df_saida = pd.DataFrame()
    for i in range (i,5000):

        print('\nIteração {} - k = {}'.format(i,k))
        cluster,erro,cluster_representativo = main(members_list,k)
        cluster = cluster.astype(str)
        cluster = [x.replace('.0','') for x in cluster]
        tmp = pd.DataFrame({'Cluster': [cluster], 'Erro': erro, 'Membro
representativo': [cluster_representativo['index membro'].to_list()]})

        df_saida = pd.concat([df_saida,tmp])

    # menor erro das 5k iterações
    df_saida = df_saida.sort_values('Erro', ascending = True)
    cluster = df_saida['Cluster'].iloc[0]
    membros_representativos = df_saida['Membro representativo'].iloc[0]
    df_saida['K'] = k
    df_elbow = pd.concat([df_elbow,df_saida.head(1)])

```