Universidade de Brasília
Faculdade de Tecnologia

Revisão da Metodologia de Definição dos Limites para os indicadores de continuidade DEC e FEC das Distribuidoras

Giovane Nunes Cornelio Rêgo

TRABALHO DE CONCLUSÃO DE CURSO

ENGENHARIA ELÉTRICA

Brasília

2023

Universidade de Brasília
Faculdade de Tecnologia

Revisão da Metodologia de Definição dos Limites para os indicadores de continuidade DEC e FEC das Distribuidoras

Giovane Nunes Cornelio Rêgo

Trabalho de Conclusão de Curso submetido como requisito parcial para obtenção do grau de Engenheiro Eletricista

Orientador: Prof. Dr. Kleber Melo e Silva

Brasília
2023

Nunes Cornelio Rêgo, Giovane.
N000r Revisão da Metodologia de Definição dos Limites para os indicadores de continuidade DEC e FEC das Distribuidoras / Giovane Nunes Cornelio Rêgo; orientador Kleber Melo e Silva. -- Brasília, 2023.
73 p.

Trabalho de Conclusão de Curso (Engenharia Elétrica) -- Universidade de Brasília, 2023.

1. Métodos de Inteligência Artificial. 2. Métodos de Aprendizado de Máquina. 3. Indicadores de Qualidade Energia Elétrica. I. Melo e Silva, Kleber, orient. II. Título

Universidade de Brasília
Faculdade de Tecnologia

Revisão da Metodologia de Definição dos Limites para os indicadores de continuidade DEC e FEC das Distribuidoras

Giovane Nunes Cornelio Rêgo

Trabalho de Conclusão de Curso submetido como requisito parcial para obtenção do grau de Engenheiro Eletricista

Trabalho aprovado. Brasília, 20 de Dezembro de 2023:

Prof. Dr. Kleber Melo e Silva,
UnB/FT/ENE
Orientador

Esp. Davi Vidal Rola Almeida,
ANEEL/STD
Examinador externo

MSc. João Gabriel Martin Del Solar
Examinador externo

Brasília
2023

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.

Agradecimentos

Primeiramente, agradeço aos meus pais Amilton Mariano Rêgo e Ione Nunes Cornelio Rêgo pelo apoio incondicional em todas as decisões da minha vida. Por todo suporte dado nos momentos bons e ruins da vida

Ao meu irmão Leonardo Nunes Cornelio Rêgo por ser um exemplo de maturidade e sabedoria durante toda minha vida

À minha família em geral por ser um pilar no qual posso me sustentar

À minha namorada Cecília Farage Ramos por ter me acompanhado durante todo esse processo e me tranquilizado no dia a dia

À Universidade de Brasília e ao Departamento de Engenharia Elétrica

Ao meu orientador Kleber Melo e Silva pelo apoio profissional e por acreditar no meu potencial

À ANEEL por me dar a oportunidade de estudar esse tema

Aos meus supervisores Renato Eduardo Farias de Sousa e Davi Vidal Rôla Almeida por terem me guiado durante todo esse estudo e me amadurecerem profissionalmente

Aos meus amigos Gabriel Germano, João Rafael, Tomás Gaudino, João Pedro Cardoso, Guilherme Maciel, Pedro Borges, João Lucas e João Kleber que passaram por todas as dificuldades da graduação

Aos meus amigos do peito, que me trazem descontração nos momentos difíceis.

Resumo

Este estudo teve como objetivo revisar a metodologia de definição de limites para os indicadores de continuidade DEC e FEC, visando atualizar os dados das distribuidoras e a seleção de atributos. A proposta central é proporcionar maior explicação à metodologia de definição dos limites dos indicadores DEC e FEC, buscando equidade na comparação entre conjuntos discrepantes. Os objetivos incluem a extração de novos atributos, obtidos de fontes confiáveis como IBGE, INMET, ANA, e BDGD, visando enriquecer a análise. Além disso, será realizada a comparação de métodos para a seleção de atributos, incluindo Stepwise Regression, Decision Tree Regression (DTR) e Multi-Layer Perceptron Regression (MLP). O trabalho inicia com a extração de atributos importantes de várias bases de dados governamentais. Em seguida, os dados passam por uma transformação estatística, como a normalização, para otimizar o desempenho dos modelos. A etapa seguinte envolve a execução de três modelos de regressão (Stepwise Regression, Decision Tree Regression e MLP Regressor) por meio de um código Python. As métricas comparativas (MSE, MAE e R^2) são utilizadas para avaliar a precisão dos modelos em relação aos valores reais da variável dependente. Comparando três métodos (Stepwise Regression, Decision Tree Regression e Multi-Layer Perceptron Regression), o MLP Regressor com ativação relu se destacou nas métricas propostas, embora não tenha alcançado valores ideais. A MLP Regression foi geralmente superior nas métricas (MAE, MSE, R^2), seguida por árvores de decisão e, por último, Stepwise Regression. Limitações incluem a escassez de dados sobre a gestão das distribuidoras, sugerindo melhorias futuras com a inclusão de novos dados. A conclusão destaca este trabalho como ponto de partida para debates acadêmicos, contribuindo ao avanço do conhecimento em métodos de aprendizado de máquina, especialmente na Metodologia de Definição de Limites de indicadores de continuidade.

Palavras-chave: Métodos de Inteligência Artificial. Métodos de Aprendizado de Máquina. Indicadores de Qualidade Energia Elétrica.

Abstract

This study aimed to review the methodology for defining limits for the continuity indicators DEC and FEC, aiming to update the data of the distributors and the selection of attributes. The central proposal is to provide a more detailed explanation of the methodology for defining limits for the DEC and FEC indicators, seeking equity in the comparison between discrepant sets. Objectives include the extraction of new attributes from reliable sources such as IBGE, INMET, ANA, and BDGD, aiming to enrich the analysis. Additionally, a comparison of methods for attribute selection will be conducted, including Stepwise Regression, Decision Tree Regression (DTR), and Multi-Layer Perceptron Regression (MLP). The work begins with the extraction of important attributes from various government databases. Next, the data undergo statistical transformation, such as normalization, to optimize the performance of the models. The subsequent step involves the execution of three regression models (Stepwise Regression, Decision Tree Regression, and MLP Regressor) through Python code. Comparative metrics (MSE, MAE, and R^2) are used to evaluate the accuracy of the models compared to the actual values of the dependent variable. Comparing three methods (Stepwise Regression, Decision Tree Regression, and Multi-Layer Perceptron Regression), the MLP Regressor with relu activation stood out in the proposed metrics, although it did not reach ideal values. MLP Regression generally outperformed in metrics (MAE, MSE, R^2), followed by decision trees, and lastly, Stepwise Regression. Limitations include the scarcity of data on the management of distributors, suggesting future improvements with the inclusion of new data. The conclusion highlights this work as a starting point for academic discussions, contributing to the advancement of knowledge in machine learning methods, especially in the methodology for defining limits of continuity indicators.

Keywords: Artificial Intelligence Methods. Machine Learning Methods. Electrical Energy Quality Indicators.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ANA	Agência Nacional de Águas e Saneamento Básico ... 38
ANEEL	Agência Nacional de Energia Elétrica ... 14
ANN	*Artificial Neural Network* ... 15
BDGD	Bases de Dados Geográficas das Distribuidoras ... 16
CTMT	Circuito de Média Tensão ... 44
DEC	Duração Equivalente de Interrupção por Unidade Consumidora ... 14
EQRE	Equipamento Regulador ... 44
FEC	Frequência Equivalente de Interrupção por Unidade Consumidora ... 14
FUNAI	Fundação Nacional dos Povos Indígenas ... 38
GMDH	*Group method of data handling* ... 16
HVAC	*Heating, Ventilating and Air Conditioning* ... 16
IA	Inteligência Artificial ... 15
IBGE	Instituto Brasileiro de Geografia e Estatística ... 16
ICMBIO	Instituto Chico Mendes de Conservação da Biodiversidade ... 16
INCRA	Instituto Nacional de Colonização e Reforma Agrária ... 38
INMET	Instituto Nacional de Metereologia ... 46
INPE	Instituto Nacional de Pesquisas Espaciais ... 16
LSSVM	*Least-squares Support Vector Machine* ... 16
MAE	*Mean Absolute Error* ... 28
MMA	Ministério do Meio Ambiente ... 38
MSE	*Mean Squared Error* ... 28
PCA	Principal Component Analysis ... 34
RAMLIG	Ramal de ligação ... 44
RASE	*Root of Absolute Squared Error* ... 15
ReLU	*Rectified linear unit* ... 62
RF	*Random Forests* ... 16
RMSE	*Root of Mean Squared Error* ... 16
RNA	Rede Neural Artificial ... 15
R^2	Coeficiente de Determinação ... 15
SAS	*Statistical Analysis System* ... 16
SFS	*Sequential Feature Selector* ... 66
SIG-R	Sistema de Informações Geográficas Regulatório ... 16
SSDBT	Segmento do Sistema de Distribuição de Baixa Tensão ... 44
SSDMT	Segmento do Sistema de Distribuição de Média Tensão ... 44
SVM	*Support Vector Machine* ... 16

Sumário

1 Introdução

1.1 Contextualização

A metodologia de definição dos limites para os indicadores de continuidade das distribuidoras de energia elétrica é revisada esporadicamente pela Agência Nacional de Energia Elétrica (ANEEL). Os indicadores que são estudados e que são definidos os limites nessa metodologia são os de Duração Equivalente de Interrupção por Unidade Consumidora (DEC) e de Frequência Equivalente de Interrupção por Unidade Consumidora (FEC). Esses indicadores mostram, respectivamente, o intervalo de tempo e o número de interrupções que cada consumidor, em média, ficou sem energia elétrica, considerando interrupções iguais ou superiores a três minutos.

Com a ideia de estabelecer padrões de qualidade e garantir que as distribuidoras forneçam energia elétrica de maneira contínua, a ANEEL define, através de regulação e estatística, limites máximos de DEC e FEC que os conjuntos das distribuidoras devem cumprir. Caso as distribuidoras não atinjam os limites estabelecidos, estão sujeitas a penalidades e obrigações de melhorias no serviço prestado. A Nota Técnica n° 0102/2014 [6] é responsável por explicar, de forma detalhada, todos os passos da metodologia de definição desses limites.

De acordo com a Nota Técnica nº 0136/2021-SRD/ANEEL[7], as distribuidoras de energia elétrica têm apresentado contribuições à ANEEL referentes aos atributos utilizados na metodologia de comparação entre conjuntos de dados, especialmente os relacionados a índice pluviométrico e vegetação remanescente. Na Nota Técnica nº 0136/2021-SRD/ANEEL (2021, p. 37) diz-se, "Apesar de se tratar de atributos dos quais se espera menor variabilidade, alguns agentes têm alertado nas discussões em consultas públicas que os níveis de chuva em determinadas áreas de concessão nos últimos anos vêm apresentando incremento relevante em relação à base atualmente usada pela Agência. Assim, alegam ser importante a inclusão na nova metodologia de um procedimento para a atualização mais frequente desses dados."

Além disso, essa tomada de subsídio, com intuito de aprimorar a regulamentação que define a metodologia para o estabelecimento de limites de DEC e FEC dos conjuntos de unidades consumidoras das distribuidoras, também cita estudos feitos pela Celesc que dão contribuições relativas ao passo de seleção de atributos. Sendo assim, vê-se necessário o estudo de novos métodos que podem contribuir para a melhoria dessa ferramenta do setor elétrico.

1.2 Motivação

A inteligência artificial (IA) possui técnicas promissoras para abordar questões de pesquisa devido a algumas razões. Problemas do mundo real muitas vezes envolvem complexidade e não linearidade, o que pode ser desafiador para métodos tradicionais. Algoritmos de IA, como Redes Neurais Artificiais (RNAs) e Máquinas de Vetores de Suporte (SVM), podem aprender padrões complexos e lidar com relações não lineares.

Além disso, a IA pode identificar relações não triviais entre variáveis, ajudando os pesquisadores a descobrirem padrões e conexões que podem não ser evidentes devido à complexidade dos dados. Por fim, modelos de IA treinados em um conjunto de dados podem ser geralmente aplicáveis a novos dados, permitindo a generalização. Abaixo está um histórico de estudos que utilizaram técnicas de IA no âmbito da Engenharia Elétrica.

Em 1999, Kalogirou [14] utilizou Redes Neurais Artificiais no contexto da energia solar, modelando a resposta ao aquecimento de uma planta de geração de vapor através da energia solar, estimando fatores de interceptação de coletores solares. Além disso, essas redes também foram empregadas nas estimativas das cargas térmicas dos edifícios, sendo que foi utilizado uma arquitetura com múltiplas camadas ocultas. As Redes Neurais Artificiais foram aplicadas com sucesso em: classificações, previsões, sistemas de controle e otimização.

Em 2000, Kalogirou & Bojic [15] empregaram Redes Neurais Artificiais (RNAs) na previsão do consumo de energia de um edifício solar. O consumo de energia do edifício depende do isolamento de todas as paredes, da espessura da alvenaria e do isolamento e da estação do ano. Dados simulados para vários casos foram usados para treinar uma Rede Neural Artificial (RNA), a fim de relacionar as entradas com a saída através de um coeficiente R^2. O coeficiente de determinação múltipla R^2 obtido neste caso foi igual a 0,9991.

Em 2005, González & Zamarreño [11] apresentaram uma nova abordagem para a previsão de carga de curto prazo em edifícios. O método é baseado em um tipo especial de Rede Neural Artificial (RNA) que retroalimenta uma parte de suas saídas (feedback *Artificial Neural Network* (ANN)). O sistema utiliza valores atuais e previstos de temperatura, a carga atual, a hora e o dia como entradas. Os resultados obtidos demonstram a alta precisão.

Em 2007, Tso & Yau [22] compararam três técnicas de modelagem para prever o consumo de energia elétrica. As técnicas comparadas foram: Regressões Lineares, Árvores de Decisão e Redes Neurais Artificiais. A seleção do modelo é baseada na raiz quadrada do erro quadrático médio (RASE). Como conclusão, as diferenças no RASE entre os três tipos de modelo são bastante pequenas, indicando que as três técnicas de modelagem são geralmente comparáveis na previsão do consumo de energia.

Em 2017, Ahmad, Mourshed & Rezgui [1] compararam a performance de uma feed-forward back-propagation Artificial Neural Network (ANN) com um algoritmo de florestas

aleatórias (RF) para prever o consumo horário de energia de um sistema de climatização *Heating, Ventilating and Air Conditioning* (HVAC). A seleção do modelo é baseada na raiz do erro quadrático médio (RMSE). No geral, a ANN performou levemente melhor do que o algoritmo de florestas aleatórias, porém sua aplicabilidade é menos viável. Chegou-se à conclusão que ambos os modelos têm poder preditivo comparável e são quase igualmente aplicáveis em aplicações energéticas de edifícios.

Em 2014, Ahmad et. al [2] revisaram o método de previsão de energia elétrica em edifícios utilizando métodos de inteligência artificial, como Máquina de Vetores de Suporte (SVM) e Redes Neurais Artificiais (RNA). De acordo com o artigo, a hibridização dos dois métodos de previsão tem o potencial de ser aplicada para resultados mais precisos. Os resultados de previsão foram avaliados utilizando análise de erro, como o Raiz do Erro Quadrático Médio (RMSE) e o coeficiente de correlação. Neste estudo, o método híbrido proposto foi comparado com outros métodos individuais, como *Group method of data handling* (GMDH), *Least-squares Support Vector Machine* (LSSVM) e ANN, para validar o desempenho do método híbrido.

1.3 Metodologia da ANEEL

1.3.1 Extração dos Atributos

A extração dos dados dos atributos relativos ao setor elétrico é realizada a partir das Bases de Dados Geográficas das Distribuidoras (BDGD) encaminhadas anualmente pelas empresas para o Sistema de Informações Geográficas Regulatório (SIG-R) da ANEEL. Além dos atributos relativos ao setor elétrico, foram utilizados atributos socioeconômicos, climáticos, infraestruturais e vegetativos/meio ambiente, vindo de órgãos como Instituto Chico Mendes de Conservação da Biodiversidade (ICMBIO), Instituto Nacional de Pesquisas Espaciais (INPE), Instituto Brasileiro de Geografia e Estatística (IBGE), entre outros. Para a obtenção das informações, foram realizadas diversas operações com o auxílio dos softwares ArcMap™ do pacote ArcGIS 10.1 da Esri® e Statistical Analysis System (SAS) Enterprise Guide, visando obter os atributos relativos a cada conjunto de unidades consumidoras. Ao todo foram reunidos 146 atributos para um total de 2610 conjuntos das distribuidoras de energia elétrica.

1.3.2 Seleção dos Atributos

Inicialmente, foi realizado um cálculo de correlações de Pearson e Spearman entre os 146 atributos para retirar os dados que têm relações lineares e monótonas, respectivamente. Atributos que tenham correlações maiores ou iguais a 0.9 devem ser filtrados. Dessa forma, é escolhido o atributo que possui maior correlação com as variáveis DEC e FEC. Em seguida,

para que o atributo fizesse parte da análise seguinte, realizou-se correlação de Pearson e Spearman entre os atributos restantes e os parâmetros DEC e FEC. Foi adotado como limite uma correlação mínima de 0.2 (Pearson ou Spearman).

Nesta fase, restaram 69 atributos para análise. A seleção de atributos foi realizada através de uma técnica estatística de regressão chamada *Stepwise Regression* [19]. Esse procedimento serve para escolha dos atributos mais relevantes, tomando-se como variável dependente o DEC. O mesmo procedimento foi realizado para o FEC. Vale a pena ressaltar que foi usado os dados de DEC e FEC médio dos 3 anos antecedentes (2011, 2012 ,2013). Após todos os ciclos e condições do método iterativo, chegou-se a um resultado de 9 atributos para o DEC e 15 atributos para o FEC, levando em consideração que foi escolhido o modelo sem variáveis socioeconômicos (O nível de explicação do modelo aumentava de forma não significante ao adicionar as variáveis socioeconômicas). As figuras 1.1 e 1.2 mostram as variáveis selecionadas pelo modelo e o [1]coeficiente R^2 relacionados a cada uma delas. As figuras 1.3 e 1.4 mostram a curva de explicação dos modelos de DEC e FEC, respectivamente. Nota-se que à partir da 6 variável, o modelo não melhora significativamente. Portanto, foi decidido usar 6 atributos para o modelo final de DEC e FEC.

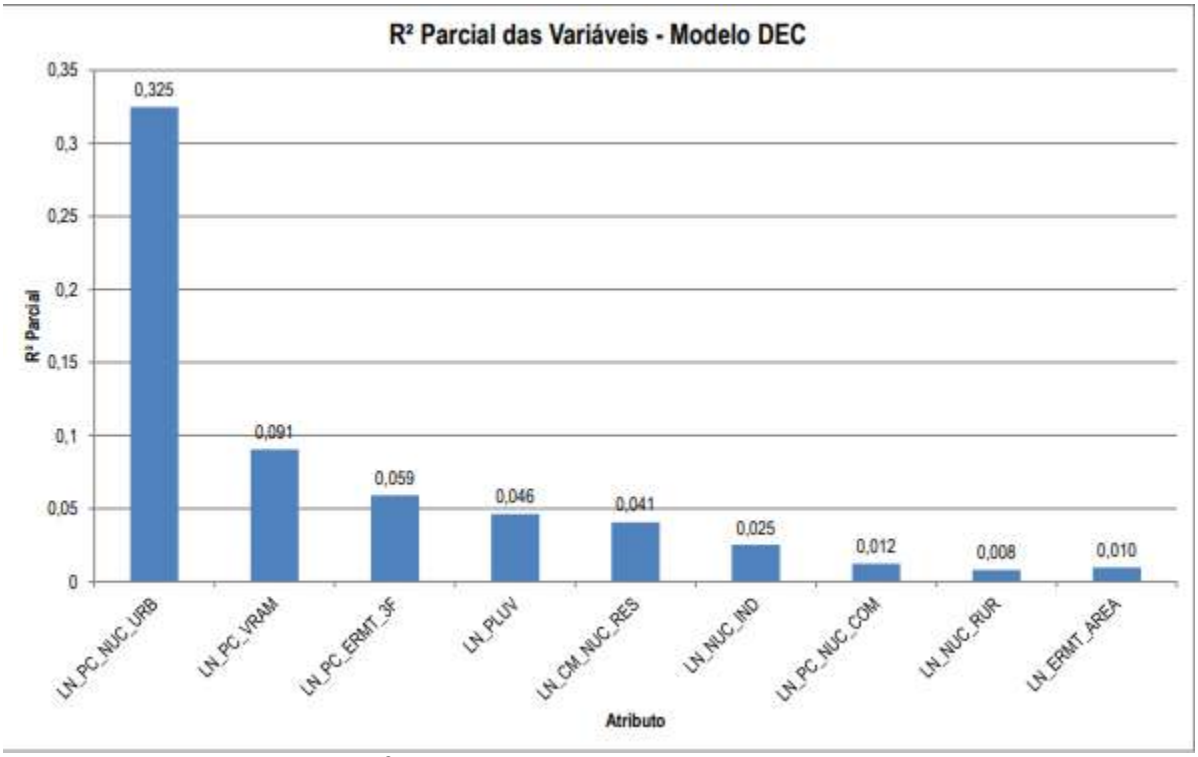


Figura 1.1 – R^2 parcial das variáveis acrescidas ao modelo DEC.
Fonte: Nota Técnica 0102/2014-SRD/ANEEL[6]

[1] O coeficiente R^2 é uma medida que representa o poder adicional de explicação de uma variável dado que outras variáveis já foram inseridas no modelo.

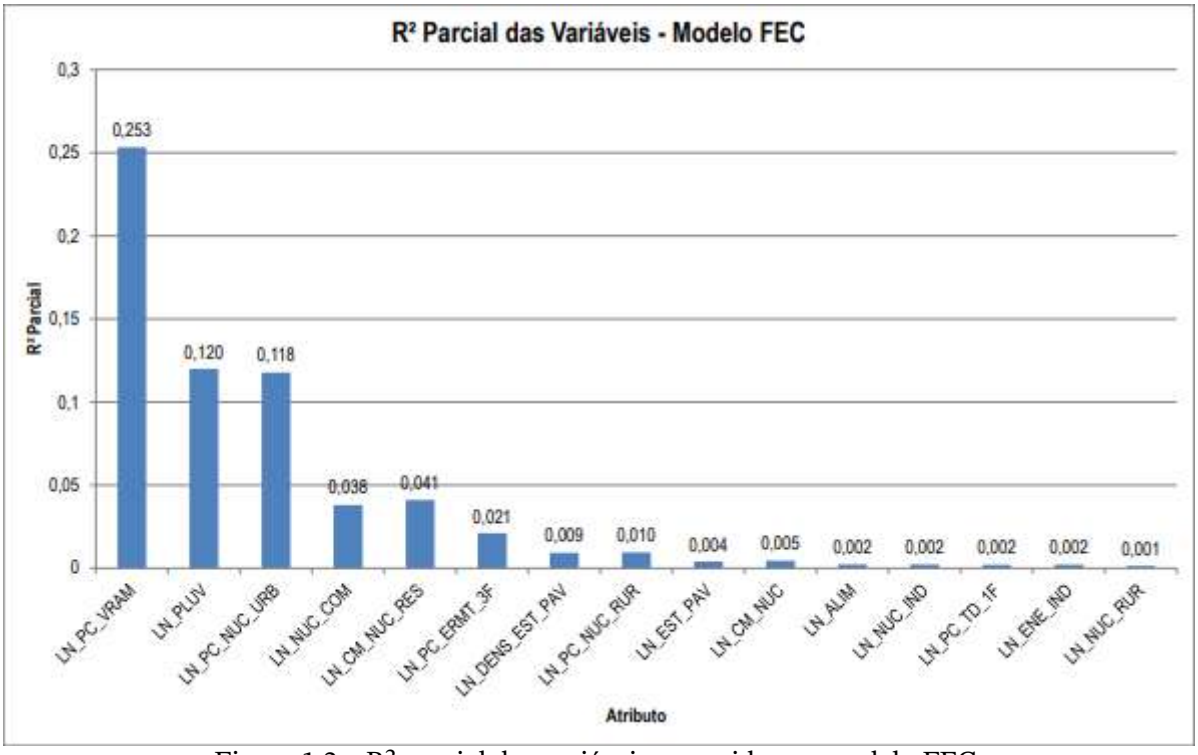


Figura 1.2 – R^2 parcial das variáveis acrescidas ao modelo FEC.
Fonte: Nota Técnica 0102/2014-SRD/ANEEL[6]

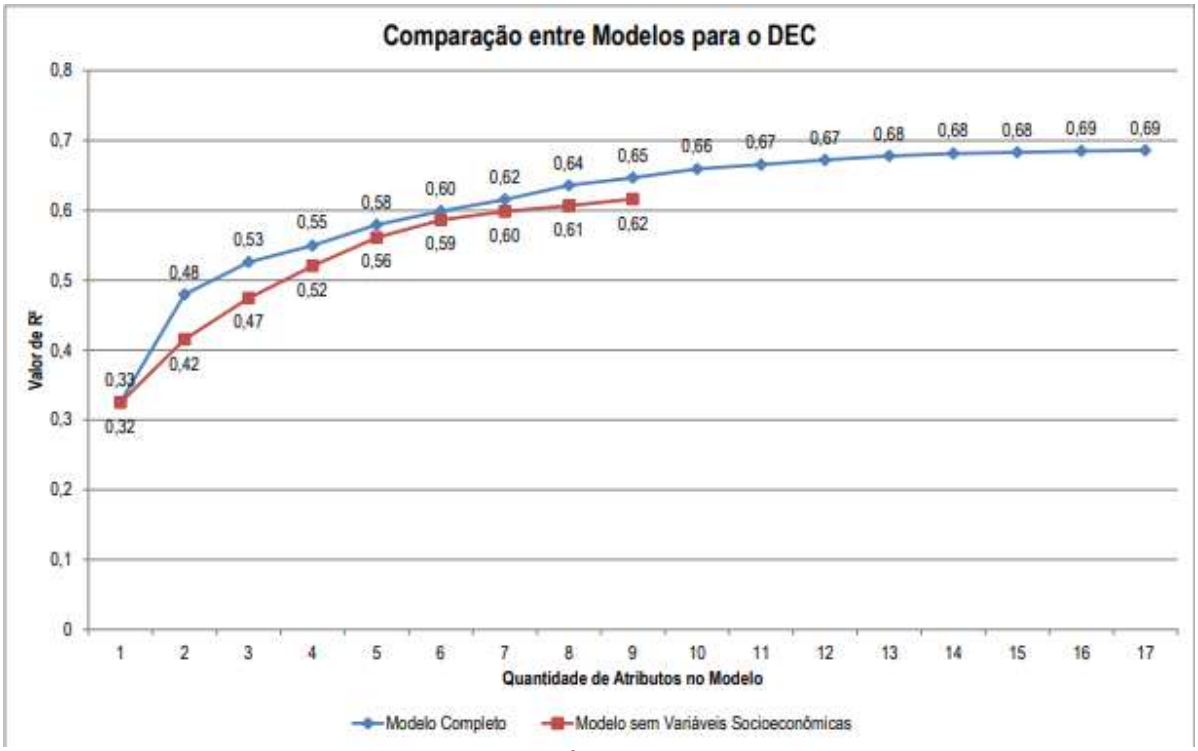


Figura 1.3 – Comparação da variável R^2 entre os dois modelos obtidos para o DEC.
Fonte: Nota Técnica 0102/2014-SRD/ANEEL[6]

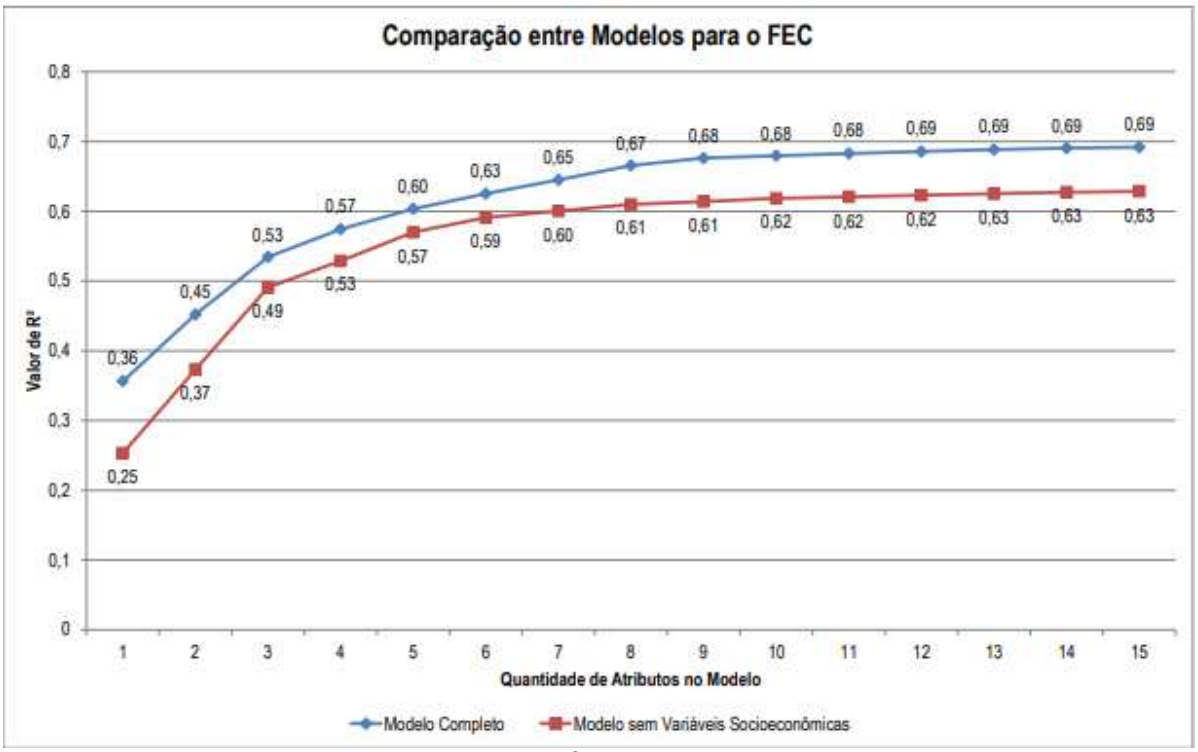


Figura 1.4 – Comparação da variável R^2 entre os dois modelos obtidos para o FEC.
Fonte: Nota Técnica 0102/2014-SRD/ANEEL[6]

DEC		***FEC***	
Sigla	***Atributo***	***Sigla***	***Atributo***
PC_NUC_AD	PERCENTUAL DE NUC EM ÁREAS DE ALTA DENSIDADE (%)	PC_VRAM	PERCENTUAL DE ÁREA COM VEGETAÇÃO REMANESCENTE ALTA OU MÉDIA (%)
PC_VRAM	PERCENTUAL DE ÁREA COM VEGETAÇÃO REMANESCENTE ALTA OU MÉDIA (%)	PLUV	PRECIPITAÇÃO PLUVIOMÉTRICA MÉDIA ANUAL (mm)
PC_ERMT_3F	PERCENTUAL DE REDES MT TRIFÁSICAS (%)	PC_NUC_AD	PERCENTUAL DE NUC EM ÁREAS DE ALTA DENSIDADE (%)
PLUV	PERCENTUAL DE ÁREA COM VEGETAÇÃO REMANESCENTE ALTA OU MÉDIA (%)	NUC_COM	NUC DA CLASSE COMERCIAL
CM_NUC_RES	CONSUMO MÉDIO POR UC DA CLASSE RESIDENCIAL (MWh)	CM_NUC_RES	CONSUMO MÉDIO POR UC DA CLASSE RESIDENCIAL (MWh)
NUC_IND	NUC DA CLASSE INDUSTRIAL	PC_ERMT_3F	PERCENTUAL DE REDES MT TRIFÁSICAS (%)

Tabela 1.1 – Tabela de atributos dos modelos finais de DEC e FEC

Nota-se que os atributos para os modelos de DEC e FEC são relativamente parecidos. A maior mudança é percebida nos coeficientes R^2 dos mesmos atributos para diferentes modelos.

1.3.3 Aplicação do Método Dinâmico

O método dinâmico visa comparar cada conjunto com os conjuntos mais semelhantes a eles. O método é aplicado separadamente para os indicadores DEC e FEC, visto que seus atributos são diferentes. Como os 7 atributos que são utilizados para os dois indicadores possuem grandezas e dimensões totalmente diferentes, foi utilizado uma normalização *Z-Score* dos dados, assim, transformando o dado em uma variável com média zero e desvio padrão unitário.

$$x_{ij} = \frac{x^*_{ij} - m_l}{s_l}, i = 1,...,N; l = 1,...d, \tag{1.1}$$

onde x^*_{ij} é o dado original, m_l é a média amostral, s_l é o desvio padrão amostral, N é o número de conjuntos e d é o número de atributos.

Após normalizar todos os dados, é utilizado a distância euclidiana entre os conjuntos, medida de similaridade que foi chamada de "Método Dinâmico".

$$D(x_i,x_j) = \sqrt{\sum_{l=1}^{d}(x_{il} - x_{jl})^2}, \tag{1.2}$$

onde x_i e x_j são os conjuntos de unidade consumidoras de d (número de atributos) matrizes com dimensões NxN.

Desse modo, calcula-se a distância de cada conjunto para todos os conjuntos do Brasil. Ressalta-se, porém, que separa-se os conjuntos aéreos dos subterrâneos na etapa de comparação (ou seja, conjuntos aéreos só são comparados com conjuntos aéreos, e conjuntos subterrâneos só se comparam com subterrâneos).

O próximo passo consiste em ordenar as distâncias de cada conjunto para os demais, obtendo-se assim os conjuntos mais próximos a cada conjunto da análise. Adotou-se o número de 100 conjuntos, assim, os conjuntos são comparados aos 100 conjuntos mais próximos, desde que respeitado um limite de homogeneidade entre eles. Para a obtenção desse limite, criou-se a grandeza denominada heterogeneidade percentual, calculada de acordo com a equação (1.3). Definiu-se o limite de heterogeneidade em 20%.

$$Heterogeneidade_i^P = \frac{Max(Dist_i^j)}{3\sqrt{k}} \tag{1.3}$$

Assim, se o valor de heterogeneidade calculado para os 100 conjuntos mais próximos ao conjunto em análise superar 20%, elimina-se o conjunto mais distante da análise e a heterogeneidade é recalculada. Esse processo continua até que a heterogeneidade seja inferior a 20% ou o número de conjuntos comparaveis chegue a 50. Caso a heterogeneidade seja inferior a marca de 20%, o conjunto é classificado como homogêneo. No entanto, se ele chegar a 50 conjuntos comparáveis, é classificado como heterogêneo e recebe um tratamento particular explicado na seção seguinte.

Definidos os conjuntos semelhantes ao conjunto em análise, parte-se para a definição dos limites. Utilizando-se da técnica denominada yardstick competition, a ANEEL define o valor de referência para cada agrupamento, o qual definirá o limite objetivo a ser alcançado pelo conjunto em análise.

Para conjuntos interligados aéreos, define-se o percentil 20 do agrupamento como limite objetivo. Para conjuntos aéreos isolados, adota-se o percentil 50. No caso dos conjuntos subterrâneos, que são comparados apenas entre si, também se adota o percentil 50 como limite objetivo. O percentil é obtido ordenando-se os conjuntos de acordo com a média dos desempenhos observados (valores apurados de DEC ou FEC dos conjuntos) para os três últimos anos civis disponíveis. Assim, no caso de um agrupamento com 100 conjuntos, o percentil 20 será o valor do indicador obtido pelo 20º colocado (partindo-se do melhor para o pior desempenho) entre os conjuntos.

Define-se que o período de transição para que o conjunto atinja o limite objetivo é de 8 anos (equivalente a 2 revisões tarifárias). Desse modo, o limite é construído para ser atingido 8 anos à frente, porém, eles são calculados toda revisão tarifária (de 3 a 5 anos). A fórmula de redução dos limites é linear e dada por[6]:

$$Limite^t = Limite^0 - t \times \left(\frac{Limite^0 - Limite^*}{T} \right), Limite^0 > Limite^* \tag{1.4}$$

$$Limite^t = Limite^0, Limite^0 \leq Limite^* \tag{1.5}$$

onde:

T - período de transição de 8 anos;

t - ano que se deseja calcular o limite;

$Limite^t$ - limite a ser calculado para o ano t;

$Limite^0$ - limite atual do conjunto;

$Limite^*$ - Limite obtido do percentil 20.

1.3.4 Avaliação de Conjuntos Heterogêneos

Para os conjuntos heterogêneos, quais sejam aqueles cuja heterogeneidade percentual excede 20%, é aplicada uma métrica denominada Score ANI, visando definir se o conjunto possui características mais favoráveis ou menos favoráveis que os conjuntos de seu agrupamento. Como diz a Nota Técnica n° 0102/2014 [6]: "Para o cálculo do Score ANI, primeiramente normaliza-se individualmente cada atributo de cada conjunto do Brasil, de modo que o conjunto com o atributo mais complexo receba para esse atributo o valor 100%, enquanto o conjunto com o atributo menos complexo receba o valor 0%. Denomina-se essa grandeza normalizada como “Atributo Normalizado Individual – ANI”. Evidentemente, a complexidade na prestação do serviço de distribuição pode aumentar ou diminuir quando se eleva o valor de um atributo, de acordo com a natureza da grandeza em questão. Para definir se a complexidade aumenta ou diminui com o valor de um atributo, foi utilizado o

sinal da correlação de Pearson entre o atributo e os indicadores DEC e FEC. Assim, tem-se que, para atributos cuja complexidade aumenta conforme se eleva o valor do atributo, o valor do ANI é calculado como:"

$$ANI_{ij} = \frac{X_{ij} - X_{jMin}}{X_{jMax} - X_{jMin}} \times 100[\%] \tag{1.6}$$

"Para atributos cuja complexidade diminui com o aumento do valor do atributo, o valor do ANI é dado por:"

$$ANI_{ij} = 100 - \frac{X_{ij} - X_{jMin}}{X_{jMax} - X_{jMin}} \times 100[\%] \tag{1.7}$$

onde:

i - índice do conjunto;

j - índice do atributo;

X_{ij} - valor do atributo j do conjunto i;

X_{jMin} - valor mínimo do atributo j na base de dados;

X_{jMax} - valor máximo do atributo j na base de dados (excetuando-se os outliers).

Em seguida, é calculado o Score ANI:

$$ScoreANI = \frac{\sum_{j=1}^{n}(ANI_{hj} - \overline{ANI_J})}{n}[\%] \tag{1.8}$$

onde:

ANI_{hj} - ANI do conjunto heterogêneo para o atributo j;

$\overline{ANI_J}$ - ANI médio dos conjuntos do agrupamento do conjunto heterogêneo, para o atributo j;

j - índice do atributo;

n - quantidade de atributos.

Referenciando novamente a Nota Técnica 0136/2021 [7], a explicação dada para o Score ANI é: "O Score ANI determina se um conjunto heterogêneo possui, em média, atributos com maior complexidade (quando positivo) ou menor complexidade (quando negativo), quando comparado aos conjuntos de seu agrupamento. Desse modo, utiliza-se o Score ANI para modificar o percentil do conjunto heterogêneo."A Tabela 1.2 mostra o funcionamento desse mecanismo:

Score ANI	*Variação do Percentil*
$\leq -3\%$	-10%
$\geq -3\% e < 3\%$	0%
$\geq 3\% e < 6\%$	+10%
$\geq 6\% e < 9\%$	+20%
$\geq 9\%$	+30%

Tabela 1.2 – Variação nos percentis a serem utilizados para conjuntos heterogêneos com base no Score ANI.

1.3.5 Avaliação das Trajetórias de Redução Intensas

A Avaliação das Trajetórias de Redução Intensas servem para tornar a descida de redução dos limites menos abrupta (gradual e factível). Assim, definiu-se um limitador para a trajetória de redução a ser estabelecida para os conjuntos. Para tanto, analisou-se o desempenho dos conjuntos no período de 2011 a 2013, verificando-se aqueles que obtiveram os melhores resultados. Selecionou-se então o percentil 90 dos conjuntos que melhoraram como o limitador, resultando em 8 horas para o DEC e 5 interrupções para o FEC como valores máximos de redução anual. Assim, caso a trajetória de redução obtida da comparação entre conjuntos possua diferenças superiores a 8 horas e 5 interrupções em anos sucessivos, limita-se a redução a esses valores.

1.4 Objetivos Gerais

Além de atualizar os dados relativos aos conjuntos das distribuidoras da antiga metodologia de 2014, a proposta geral e central desse trabalho é poder entregar mais poder de explicação para a metodologia de definição dos limites dos indicadores DEC e FEC. Em adição à isso, tornar mais justa a comparação entre conjuntos que destoam da maioria em relação aos atributos, que podem não justificar a situação individual do conjunto.

1.5 Objetivos Específicos

Os objetivos específicos desse trabalho são:

- Extração de Novos Atributos
- Teste e Comparação de Métodos para predição de valores ou para seleção dos atributos

A extração de novos atributos será importante para atualizar os dados que serão utilizadas nesse estudo. Vale ressaltar que, desde 2014, as bases de dados utilizadas no estudo da ANEEL passaram por evoluções no quesito quantidade e qualidade dos dados. Dados vindos de órgãos como: Instituto Brasileiro de Geografia e Estatística (IBGE), Instituto Nacional de

Meteorologia (INMET), Agência Nacional de Águas e Saneamento Básico (ANA), Base de Dados Geográfica da Distribuidora (BDGD), são alguns exemplos de dados que darão mais confiabilidade no nosso estudo, trazendo bases de dados ricas em informações.

A comparação de métodos para a seleção de atributos será essencial para decidir qual método melhor performa de acordo com a quantidade de dados que serão inseridos. A ideia principal será comparar o método *Stepwise Regression* com *Decision Tree Regression* (DTR) e *Artificial Neural Network* (ANN) [1] [8] [10] [22].

1.6 Apresentação do Manuscrito

No capítulo 2 é feita uma revisão sobre os fundamentos teóricos que serão utilizados, além de explicar a metodologia aplicada nesse trabalho. Em seguida, o capítulo 3 é composto pela explicação da extração dos dados e como ficaram as tabelas finais dos dados e seus formatos. No capítulo 4, há a explicação das escolhas dos modelos e seus resultados. Por fim, no capítulo 5 está a conclusão desse trabalho.

2 Fundamentação Teórica

2.1 *Stepwise Regression*

Segundo Martins & Domingues [17], a Regressão *Stepwise* é um método de seleção de variáveis que, em cada etapa, analisa a contribuição de cada variável independente na formação da significância da probabilidade F. O modelo se inicia apenas com a constante e vai adicionando ou retirando, em cada etapa, as variáveis num certo nível de confiança. O método é interrompido quando não houver mais variáveis elegíveis para inclusão e exclusão. Esse método é útil quando lidamos com um grande número de variáveis e desejamos construir um modelo apenas com aquelas que são estatisticamente relevantes, ou seja, variáveis que contribuem significativamente para a formação do modelo. Existem três principais variantes desse método:

- *Forward Selection*: É um método de seleção passo a passo que insere as variáveis independentes sequencialmente no modelo. A primeira variável que se inclui é a que apresenta maior correlação com a variável dependente e assim sucessivamente, até que não haja mais variáveis que atendam ao critério de significância adotado pelo modelo.
- *Backward Elimination*: É um método em que as variáveis independentes começam todas dentro do modelo e, em seguida, são sequencialmente removidas, uma a uma, na medida em que apresentam baixa correlação parcial com a variável dependente. O fim do processo ocorre quando não há mais variáveis abaixo do critério de significância adotado pelo modelo.
- *Stepwise forward-backward*: É uma abordagem que combina os métodos forward e backward, permitindo a adição e remoção de variáveis em cada etapa, desde que os critérios de significância sejam seguidos.

2.1.1 Formato dos Dados

O formato dos dados utilizados na regressão linear múltipla consiste em uma coluna para a varíavel dependente Y e múltiplas colunas para as variáveis independentes $X_1, X_2 ... X_n$. Abaixo está um exemplo fictício do formato dos dados:

Observação	Y	X1	X2	X3	...	Xn
1	50	3.2	7.1	12.5	...	30.8
2	65	2.8	6.5	11.8	...	32.5
3	72	3.5	7.3	12.0	...	29.7
4	58	2.9	6.8	11.2	...	30.4
5	80	3.8	7.5	12.8	...	28.9
6	67	3.1	6.2	11.5	...	31.6

Tabela 2.3 – Dados fictícios de autoria própria.

2.1.2 Modelo

Abaixo está um diagrama esquemático da *Stepwise Regression* bidirecional:

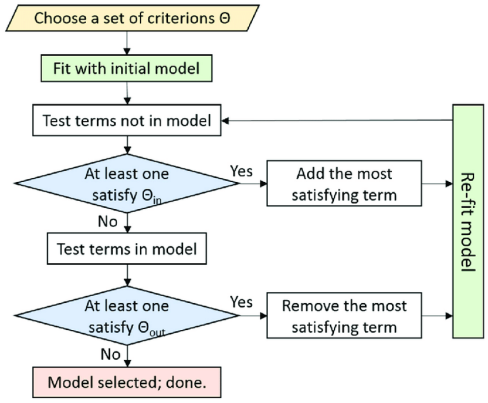


Figura 2.5 – Diagrama esquemático da *Stepwise Regression*.
Fonte: Artigo [24]

Para decidir qual variável entra e qual sai, é utilizado um critério de significância para entrada e saída da variável do modelo. A métrica utilizada para cálculo da significância é o p-valor da tabela de análise de variância.

Ao fim desse processo iterativo e da escolha das variáveis, um modelo de Regressão Linear Múltipla de Primeira Ordem é gerado com as variáveis selecionadas. Vale ressaltar que a ideia principal desse modelo é selecionar um número pequeno de variáveis, fazendo com que as relações entre as variáveis tendam a ser lineares e seja apropriado o uso da regressão linear múltipla de primeira ordem. Abaixo é determinada a equação da regressão linear múltipla:

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + ... + \beta_n \cdot X_n + \epsilon \tag{2.1}$$

onde:

Y - é a variável dependente;

$X_1, X_2, ..., X_n$ - são as variáveis independentes;

$\beta_1, \beta_2, ..., \beta_n$ - determinam a contribuição da variável independente;

ϵ - é o erro aleatório componente do modelo;

α - intercepto (valor quando todos X_i são 0).

Para o cálculo dos β, temos:

$$\beta = (X^T X)^{-1} X^T Y \tag{2.2}$$

onde:

X - é a matriz de design, que inclui todas as variáveis independentes;

Y - é o vetor de variáveis dependentes.

Dessa forma, o vetor de coeficientes β é calculado.

2.2 *Decision Tree Regression*

Segundo Russel & Norvig [21], uma árvore de decisão representa uma função que toma como entrada um vetor de valores de atributos e retorna um valor de saída único. Uma árvore de decisão alcança a decisão realizando uma sequência de testes, começando na raiz e seguindo o ramo apropriado até que uma folha seja alcançada. Cada nó interno na árvore corresponde a um teste do valor de um dos atributos de entrada, os ramos do nó são rotulados com os possíveis valores do atributo, e os nós da folha especificam qual valor deve ser retornado pela função.

O algoritmo de árvores de decisão é um método de aprendizado supervisionado não paramétrico muito utilizado em problemas de classificação e regressão [20]. As árvores de decisão possuem algumas vantagens a seu favor: facilidade de compreensão, escalabilidade para grandes conjuntos de dados e versatilidade no tratamento de entradas discretas e contínuas. Para esse trabalho em específico, as árvores de decisão serão utilizadas para regressão, pois é um modelo de aprendizado de máquina utilizado para prever valores contínuos. Ao contrário das árvores de decisão para classificação, que preveem classes discretas, as árvores de decisão para regressão estimam valores numéricos.

2.2.1 Formato dos Dados

O formato dos dados utilizados na *Decision Tree Regression* é igual aos dados da regressão linear múltipla Tabela 2.3. Uma coluna para a varíavel dependente Y e múltiplas colunas para as variáveis independentes $X_1, X_2 ... X_n$. Além disso, o conjunto de dados é dividido entre treinamento e teste, com uma proporção de 4:1, normalmente.

2.2.2 Treinamento do Modelo

Para o treinamento do modelo, a árvore é construída de maneira recursiva, dividindo o conjunto de dados em subconjuntos menores. A cada ramificação criada pela árvore, o algoritmo escolhe o melhor atributo para poder dividir os dados.

O critério para o fim do algoritmo pode ser definido pelo usuário. Isso pode incluir uma profundidade máxima da árvore, um número mínimo de amostras em uma folha, ou quando a redução na métrica de avaliação não é significativa.

Abaixo está uma representação do diagrama esquemático da *Decision Tree Regression*:

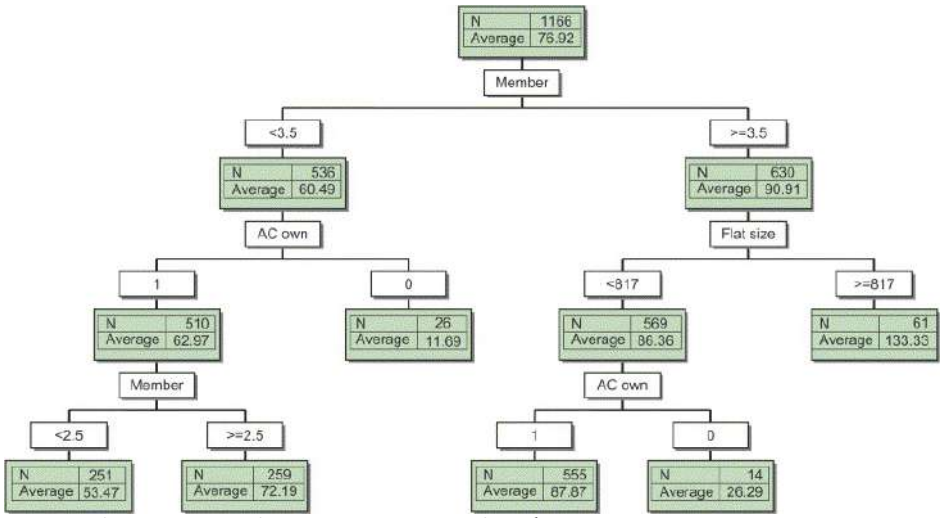


Figura 2.6 – Diagrama esquemático da *Decision Tree Regression*.
Fonte: Artigo [22]

Como os dados têm valores contínuos, consideremos que um nó pode ser representado pela letra "m". Critérios comums para minimizar e determinar qual atributo e ponto será usado para ramificar a árvore são: *Mean Squared Error* (MSE), *Half Poisson Deviance* e *Mean Absolute Error* (MAE). Essas métricas são relacionadas a redução de variância.

Valor médio de y_m:

$$\bar{y}_m = \frac{1}{n_m} \sum_{y \epsilon Q_m} y \tag{2.3}$$

Mean Squared Error:

$$H(Q_m) = \frac{1}{n_m} \sum_{y \epsilon Q_m} \left(y - \bar{y}_m\right)^2 \tag{2.4}$$

Half Poisson deviance:

$$H(Q_m) = \frac{1}{n_m} \sum_{y \epsilon Q_m} \left(y log \frac{y}{\overline{y}_m} - y + \overline{y}_m \right) \tag{2.5}$$

Mean Absolute Error:

$$H(Q_m) = \frac{1}{n_m} \sum_{y \epsilon Q_m} |y - median(y)_m| \tag{2.6}$$

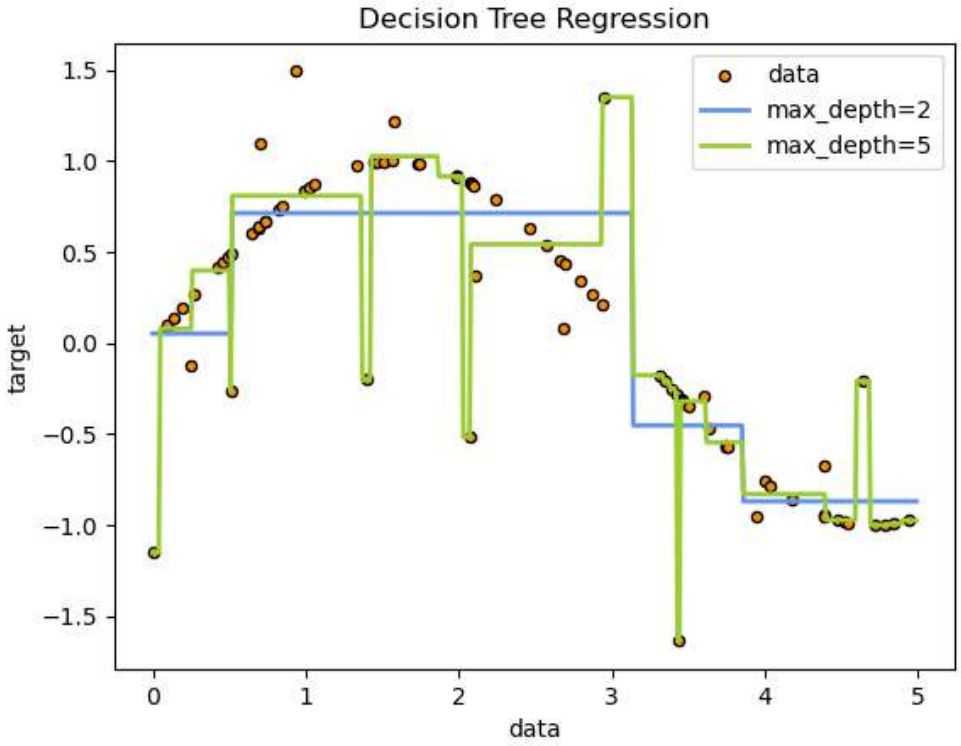


Figura 2.7 – *Decision Tree Regression*. Exemplo de *Underfitting* e *Overfitting*.
Fonte: Scikit-Learn [20]

2.2.3 Teste do Modelo

Ao término do treinamento do modelo, cada folha contém um valor que representa a predição para as instâncias de dados que contém essas características. Para fazer uma predição para uma nova instância de dados, ela percorre a árvore seguindo as decisões tomadas nos nós internos até atingir uma folha. O valor na folha (Valor médio entre os pontos do treinamento que caíram nessa folha) é então usado como a predição para a instância.

2.3 *Artificial Neural Network*

2.3.1 Neurônio

O neurônio é a unidade básica de processamento de uma rede neural[3]. Um neurônio recebe como sua entrada, a saída de neurônios de camadas anteriores. Abaixo está a representação gráfica de um neurônio.

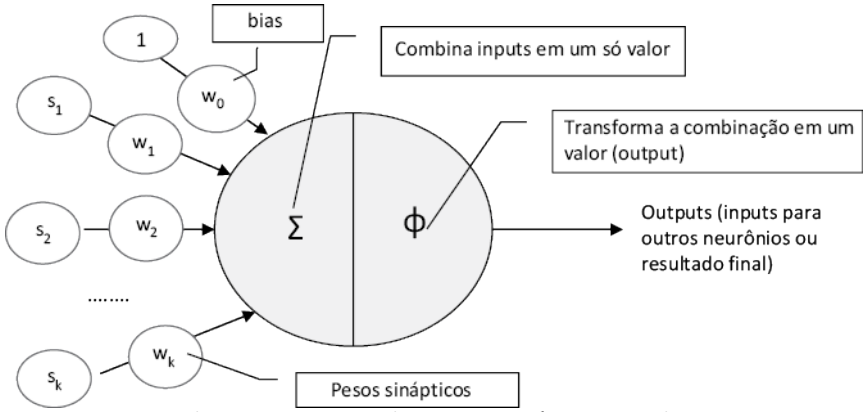


Figura 2.8 – Funcionamento do Neurônio.
Fonte: Livro [3]

O perceptron recebe várias entradas $(x_1, x_2, ..., x_n)$, cada uma multiplicada por um peso correspondente $(w_1, w_2, ..., w_n)$. $\sum$ é a função de combinação dos inputs, transformando-os em um único valor de w. A equação para o cálculo de w é:

$$w = w_0 + w_1 x_1 + ... + w_k x_k \tag{2.7}$$

O termo w_0 é denominado *Bias* e tem papel equivalente ao intercepto na regressão, fazendo com que w seja diferente de zero caso todas as inputs sejam zero. Os pesos $w_0, w_1, ..., w_k$ serão determinados durante o treinamento da rede. Treinar uma rede neural significa determinar os pesos dos neurônios que permite prever o output Y com o menor erro possível.

A variável Φ é denominada como função de ativação. Ela é responsável por transformar o valor calculado de w na saída do neurônio. Nomeando a saída do neurônio como s, temos $s = \Phi(w)$. Há algumas funções de ativação e cada uma possuí uma saída:

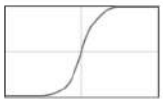
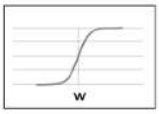
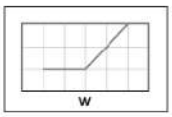
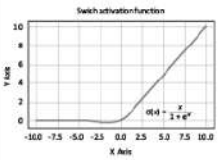
Função	Forma	Comentários	Gráfico
Função linear	$\Phi(w) = b\,w$		
Função logística (sigmoide)	$\Phi(w) = \frac{1}{1+e^{-aw}}$	A saída Φ(w) varia entre 0 e 1	
Função tangente hiperbólica (tanh)	$\Phi(w) = \frac{2}{1+e^{-2w}} - 1$	A saída Φ(w) varia entre -1 e 1	
Função ReLU (*Rectified linear unit*)	$\Phi(w) = \max(0, w)$		
Softmax (logística generalizada)	$\Phi(w_i) = \frac{\exp(w_i)}{\sum_i^n \exp(w_i)}$	Outputs positivos Soma de todos os outputs é igual a 1	
SWISH	$\Phi(w) = \frac{w}{1+\exp(-w)}$	Propostas por cientistas da Google. Segundo eles, é melhor que ReLU	

Figura 2.9 – Funções de Ativação.
Fonte: Livro [3]

De acordo com Lima [16], regras de aprendizagem são esquemas de atualização dos valores dos pesos das sinapses de uma rede neural, de forma a obter da rede um padrão de processamento desejado. Geralmente, este método se baseia na minimização do erro entre o valor desejado de saída e o valor computado pela rede. Baseado nessa diferença, os pesos da rede são ajustados a cada apresentação de um padrão de treinamento, de forma a minimizar o erro na saída para o conjunto total das informações fornecidas. O processo é repetido diversas vezes até que o procedimento tenha incorporado o conhecimento relativo às informações fornecidas.

A regra de atualização dos pesos no perceptron é dada por:

$$w_i = (y_{desejado} - y_{computado}) \cdot x_i \tag{2.8}$$

Vale notar que o perceptron é eficaz apenas para problemas linearmente separáveis, ou seja, problemas nos quais é possível traçar uma linha de decisão que separa as classes no espaço de entrada. Para problemas mais complexos, redes neurais mais sofisticadas, como Multilayer Perceptrons (MLPs), são necessárias.

A Equação (2.8) pode ser generalizada, a fim de poder ser aplicada a redes de multi-camadas. Esta generalização chama-se regra de retropropagacão (*backpropagation*).

2.3.2 *Multi-Layer Perceptron*

Segundo Lima [16], de forma genérica, para qualquer sistema de dimensão N é possível achar uma superfície separadora para as classes do problema usando uma rede de múltiplas camadas. Entretanto, encontrar os valores dos pesos necessários ou mesmo a arquitetura adequada constitui outro problema. A dificuldade configura-se em não saber o valor desejado nas saídas dos nós da camada oculta e, portanto, não se poder usar a regra de atualização do *Perceptron* para atualizar os pesos que ligam a camada de entrada à camada oculta.

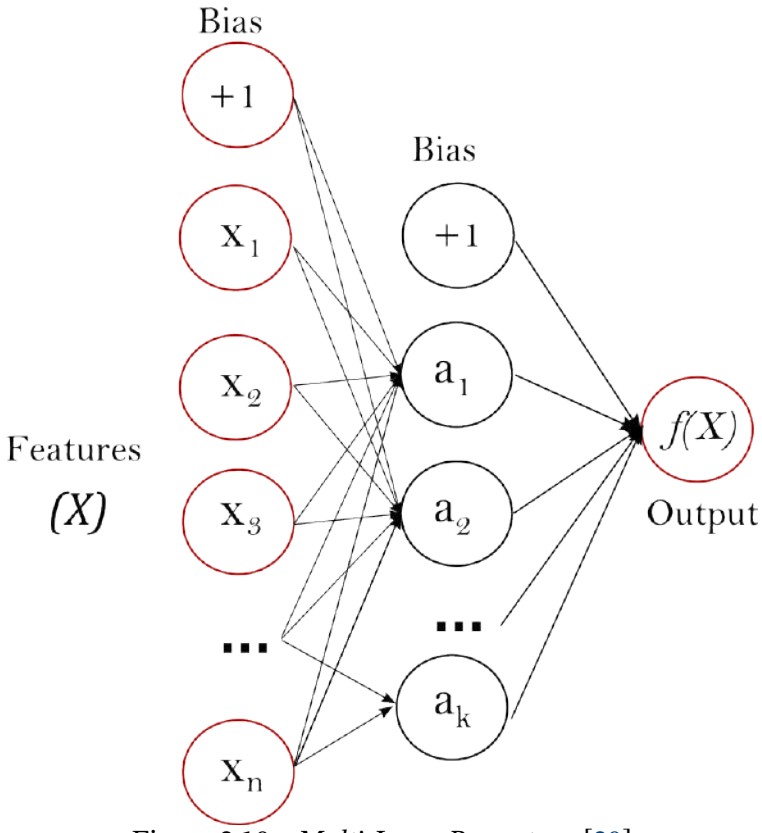


Figura 2.10 – *Multi-Layer Perceptron*.[20]

2.3.3 Treinamento do Modelo

A retropropagação (*backpropagation*) é um algoritmo usado para treinar redes neurais, especialmente redes com múltiplas camadas (MLPs). A ideia principal é ajustar os pesos da rede de acordo com o gradiente da função de perda em relação aos pesos. A fórmula para o treinamento dos *Perceptrons* multicamadas pode ser descrita em 4 passos:

- *Feedforward*:

  Para um exemplo de treinamento dado, a propagação direita é feita para calcular a saída da rede neural. Equação (2.7)

- Cálculo do Erro:

O erro entre as saídas previstas e as saídas desejadas é calculado usando alguma função de perda, como por exemplo o erro quadrático médio (MSE).

- *Backpropagation*:

  No sentido contrário da propagação direta, calcula-se os gradientes locais dos pesos em relação à função de perda na última camada e os propaga.

$$\frac{\partial Loss}{\partial w_{ij}^{(L)}} = a_i^{(L-1)} \cdot \delta_j^{(L)} \tag{2.9}$$

  onde:

  $w_{ij}^{(L)}$ - é o peso entre o neurônio i na camada $L-1$ e o neurônio j na camada L;

  $a_i^{(L-1)}$ - é a saída do neurônio i na camada $L-1$;

  $\delta_j^{(L)}$ - é o gradiente local do neurônio j na camada L.

- Pesos:

  Por fim, os pesos devem ser atualizados usando o gradiente.

$$w_{ij}^{(L)} \leftarrow w_{ij}^{(L)} - \alpha \frac{\partial Loss}{\partial w_{ij}^{(L)}} \tag{2.10}$$

  onde:

  α - é a taxa de aprendizado.

2.3.4 Teste do Modelo

De forma análoga com as árvores de decisão, o modelo é testado com o conjunto de teste, o qual foi separado previamente do conjunto de treinamento.

2.4 Metodologia

2.4.1 Extração dos atributos

Na primeira parte do trabalho, foi feita uma extração de diversos atributos que foram julgados importantes a priori, com características do tipo: facilidade na obtenção do dado, confiabilidade da informação, reprodutibilidade e, aparentemente, mínima correlação com DEC/FEC. Várias bases de dados de múltiplos órgãos governamentais foram consultadas.

2.4.2 Formatação dos dados

Posteriormente, os dados serão submetidos a um processo abrangente de tratamento, que envolve a remoção de imperfeições e erros. Isso incluiu correções na estrutura da base de dados, bem como correções de registros incompletos ou datados de maneira incorreta.

A forma final da tabela com os dados deve ter dimensões $(N + 5)x(NC)$.

N - é o número de atributos extraídos;

NC - é o número de conjuntos das distribuidoras.

ID Conjunto	Distribuidora	Nome do Conjunto	Atributo 1	Atributo 2	...	Atributo N	DEC	FEC
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...

Tabela 2.4 – Formato final da tabela de atributos.
Fonte: Autoria própria.

2.4.3 Transformação

Em seguida, os dados irão passar por algum tipo de transformação estatística, afim de melhorar o desempenho dos três modelos. Um bom exemplo seria algum tipo de normalização dos dados.

2.4.4 Regressões

Na quarta etapa, dois códigos (um para o modelo com a variável dependente DEC e outro para o FEC) *Python* irão rodar os três modelos de regressão escolhidos: *Sequential Feature Selector* aliado a uma Regressão Linear, *Decision Tree Regression* e *MLP Regressor*. Além disso, também será aplicado a Análise de Componentes Principais (PCA), que é um método de redução de dimensionalidade frequentemente utilizado para diminuir a dimensionalidade de conjuntos de dados extensos. Isso é feito transformando um grande conjunto de variáveis em um conjunto menor que ainda contém a maior parte das informações do conjunto original.

O PCA é particularmente útil para lidar com multicolinearidade entre variáveis, identificar padrões subjacentes nos dados e reduzir a complexidade do modelo. Ao realizar a decomposição das variáveis originais em componentes principais, o PCA permite uma representação mais compacta dos dados, destacando as direções de maior variância. Dessa forma, acredita-se que a aplicação do PCA pode contribuir significativamente para as variáveis dos modelos, identificando aquelas que capturam a maior quantidade de informação essencial.

A versão do Python utilizada foi a 3.11, a versão utilizada do framework foi a scikit-learn 1.3.2 e a versão utilizada do Jupyter Notebook foi 7.0.6.

2.4.5 Métricas comparativas

Em uma avaliação preliminar, *Stepwise Regression* faz a seleção de variáveis com base na correlação entre a variável independente e a dependente em uma regressão linear. Assim, as conclusões relacionadas a esse modelo, seguem em linha com a mesma construção teórica da Agência.

No caso da *MLP Regression*, as interações, sejam lineares ou não lineares, e os efeitos de ordem superior são considerados, porém não de forma compreensível. Essa não compreensibilidade traz dificuldades quanto ao alinhamento direto com os requisitos da Agência, como: possibilidade de definição clara no regulamento, facilidade de obtenção de forma sistemática e viabilidade de fiscalização. Entretanto, pode ser interessante a utilização desse método para previsões das variáveis DEC e FEC, além de expor a representatividade dos atributos em relação à essas variáveis.

Por fim, quando uma *Decision Tree Regression* é construída, ela seleciona variáveis com base naquelas que fornecem a melhor divisão nos dados em cada etapa. Isso significa que, em cada nó da árvore, apenas a variável mais informativa é escolhida para realizar a divisão naquele ponto específico. Portanto, pode acontecer que algumas variáveis importantes sejam negligenciadas em determinadas partes da árvore. A grande contribuição desse método se deve à fácil interpretabilidade, porém a identificação de padrões locais e eficiência computacional são características importantíssimas.

Para a comparação dos três modelos, além do que foi dito acima, usaremos o coeficiente de determinação R^2. O método que possuir maior R^2 mostra que o polinômio de estimação dos dados está mais próximo dos dados de teste previamente separados. Além disso, serão calculados também o MSE e o MAE, com o intuito de auxiliar o coeficiente R^2.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{2.13}$$

onde:

y_i - é o valor real da variável dependente;

$\hat{y}_i$ - é o valor previsto da variável dependente;

$\overline{y}$ - é a média dos valores reais da variável dependente;

n - é o número total de observações.

2.4.6 Fluxograma

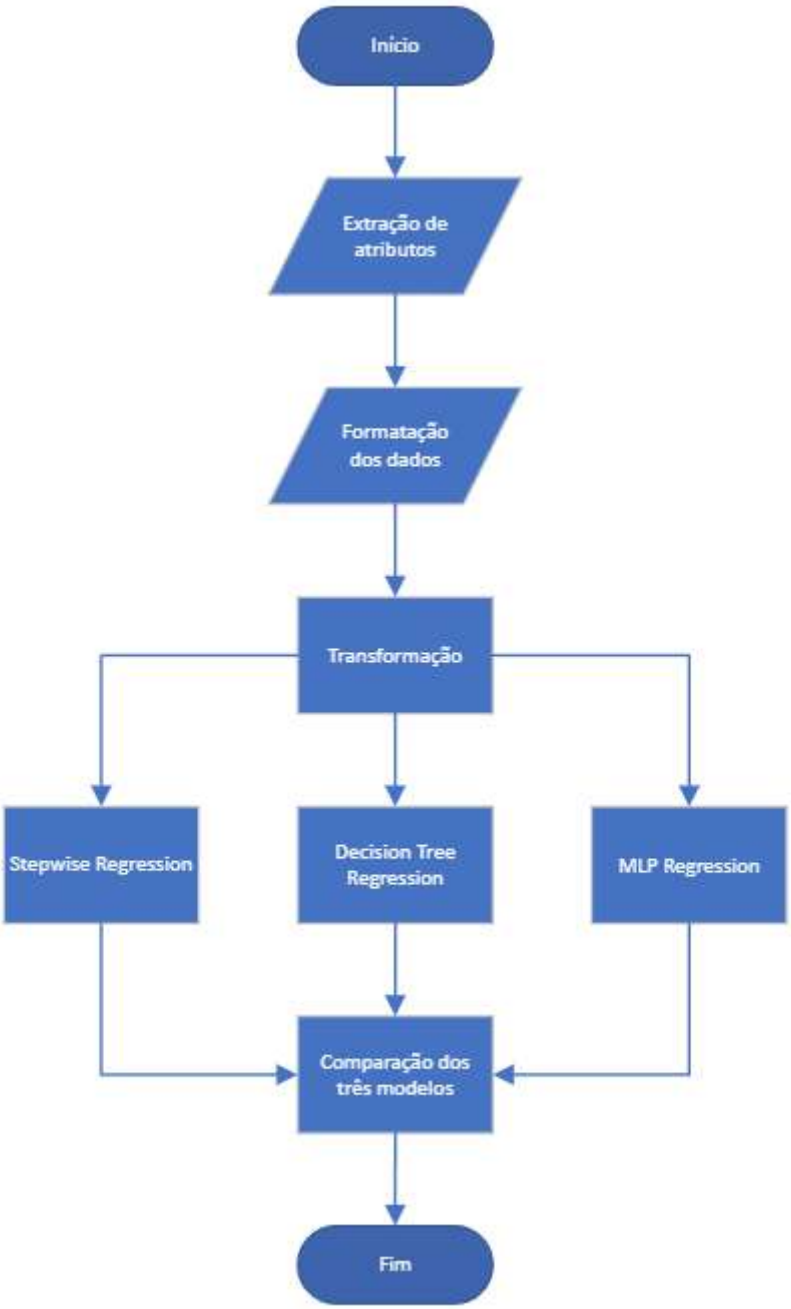


Figura 2.11 – Diagrama da Metodologia.
Fonte: Autoria própria.

3 Extração e Tratamento dos Dados

Esse capítulo será responsável por explicar o processo de extração dos atributos do objeto de estudo. Além disso, irá mostrar as manipulações feitas nas bases de dados e suas formas finais. Ao todo, existem 4 tipos de atributos: atributos relacionados ao terreno, atributos relacionados a BDGD, atributos climáticos e, por fim, atributos socioeconômicos.

Para que os atributos ficassem relativos aos conjuntos das distribuidoras, foi feito um mapa que junta todos os conjuntos de todas as distribuidoras do Brasil. Assim, os dados que fossem extraídos poderiam dar informações relativas aos conjuntos. Vale ressaltar que os conjuntos utilizados no estudo se referem à situação em 31 de dezembro de 2021. Ao todo, totalizam 3061 conjuntos resultantes de um *Merge* feito com os conjuntos das 52 concessionárias de energia elétrica (Dados do Amapá não estão disponíveis na data-base). A seguir está o *Layer* do mapa de conjuntos.

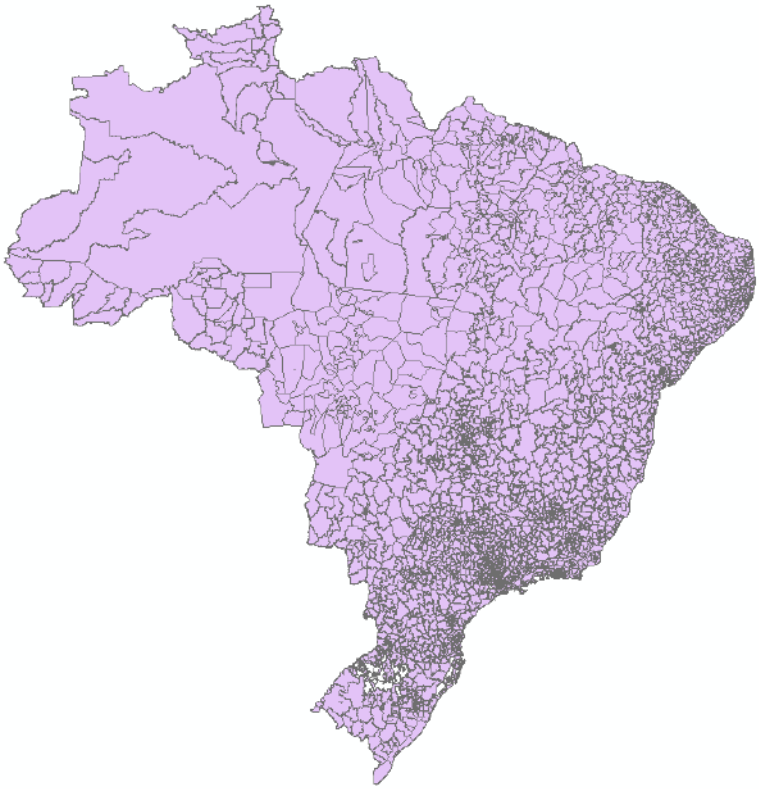


Figura 3.12 – Mapa dos conjuntos das distribuidoras.
Fonte: Autoria própria

3.1 Terreno

Os dados relacionados a terreno foram retirados de diversas fontes. Fontes como: Sistema de Informação Geográfico Regulatório SIG-R[5], Instituto Nacional de Colonização e Reforma Agrária[12] (INCRA), Agência Nacional de Águas e Saneamento Básico [4] (ANA), Instituto Brasileiro de Geografia e Estatística (IBGE), Fundação Nacional dos Povos Indígenas [9](FUNAI) e Ministério do Meio Ambiente [18](MMA).

Praticamente todos os dados vieram no formato desejado para que pudesse ser manipulado e transformado em dados para conjuntos, exceto o dado de vegetação. Nesse caso, o layer de vegetação foi dividido em 3 dados: Vegetação de Alto Porte, Vegetação de Médio Porte, Vegetação de Baixo Porte.

Analisando a base de dados de vegetação do IBGE, notou-se uma variedade maior em relação a classificação dos biomas. Tendo isso em vista, foi decidido utilizar a variável CLAS DOMI, que se refere à Vegetação/Área natural ou antropismo do principal componente, para separar todos os possíveis tipos de vegetação. Com o uso do software estatístico SAS® Enterprise Guide 8.2 foram descobertos 144 classificações diferentes de vegetação. Assim, o processo de filtragem da vegetação em relação ao seu porte foi feito no software ArcMap™ 10.7.1.

A classificação do bioma em relação ao porte de sua vegetação foi feita de forma similar à base de dados do PROBIO, base de dados elaborada em 2008 e atualmente usada pela ANEEL. Como a base de dados do IBGE possui mais classificações, nomes de biomas relativamente próximos foram adicionados à mesma categoria em relação à base passada. Entretanto, biomas totalmente novos foram analisados de forma singular e individualmente encaixados em suas classificações.

A aplicação dos filtros das vegetações foram feitos com o software ArcMap™ 10.7.1. Vale ressaltar que algumas classificações foram retiradas por não terem influência no fenômeno em análise. As classificações retiradas à priori foram: Massa d'água (MAGUA), Influência urbana (Iu) e Área indiscriminada (Ai).

LEGENDA	DESCRIÇÃO DA VEGETAÇÃO/ANTROPISMO
Ac	Agricultura/Alta
Acc	Agricultura com Culturas Cíclicas/Alta
Acp	Agricultura com Culturas Permanentes/Alta
Ag	Agropecuária/Alta
Aa	Floresta Ombrófila Aberta Aluvial
Aab	Floresta Ombrófila Aberta Aluvial com bambus
Aac	Floresta Ombrófila Aberta Aluvial com cipós
Aap	Floresta Ombrófila Aberta Aluvial com palmeiras

Ab	Floresta Ombrófila Aberta das Terras Baixas
Abb	Floresta Ombrófila Aberta das Terras Baixas com bambus
Abc	Floresta Ombrófila Aberta das Terras Baixas com cipós
Abp	Floresta Ombrófila Aberta das Terras Baixas com palmeiras
As	Floresta Ombrófila Aberta Submontana
Asb	Floresta Ombrófila Aberta Submontana com bambus
Asc	Floresta Ombrófila Aberta Submontana com cipós
Asp	Floresta Ombrófila Aberta Submontana com palmeiras
Ass	Floresta Ombrófila Aberta Submontana com sororocas
Ca	Floresta Estacional Decidual Aluvial
Cb	Floresta Estacional Decidual das Terras Baixas
Cbe	Floresta Estacional Decidual das Terras Baixas com dossel emergente
Cm	Floresta Estacional Decidual Montana
Cmu	Floresta Estacional Decidual Montana com dossel uniforme
Cs	Floresta Estacional Decidual Submontana
Cse	Floresta Estacional Decidual Submontana com dossel emergente
Csu	Floresta Estacional Decidual Submontana com dossel uniforme
Da	Floresta Ombrófila Densa Aluvial
Dae	Floresta Ombrófila Densa Aluvial com dossel emergente
Dau	Floresta Ombrófila Densa Aluvial com dossel uniforme
Db	Floresta Ombrófila Densa das Terras Baixas
Dbe	Floresta Ombrófila Densa das Terras Baixas com dossel emergente
Dbu	Floresta Ombrófila Densa das Terras Baixas com dossel uniforme
Dl	Floresta Ombrófila Densa Alto-Montana
Dm	Floresta Ombrófila Densa Montana
Dme	Floresta Ombrtfila Densa Montana com dossel emergente
Dmu	Floresta Ombrtfila Densa Montana com dossel uniforme
Ds	Floresta Ombrófila Densa Submontana
Dse	Floresta Ombrófila Densa Submontana com dossel emergente
Dsu	Floresta Ombrófila Densa Submontana com dossel uniforme
Fa	Floresta Estacional Semidecidual Aluvial
Fae	Floresta Estacional Semidecidual Aluvial com dossel emergente
Fau	Floresta Estacional Semidecidual Aluvial com dossel uniforme
Fb	Floresta Estacional Semidecidual das Terras Baixas
Fbe	Floresta Estacional Semidecidual das Terras Baixas com dossel emergente

Fm	Floresta Estacional Semidecidual Montana
Fme	Floresta Estacional Semidecidual Montana com dossel emergente
Fs	Floresta Estacional Semidecidual Submontana
Fse	Floresta Estacional Semidecidual Submontana com dossel emergente
Fsu	Floresta Estacional Semidecidual Submontana com dossel uniforme
Ha	Floresta Estacional Sempre Verde Aluvial
Hae	Floresta Estacional Sempre Verde Aluvial com dossel emergente
Hb	Floresta Estacional Sempre Verde das Terras Baixas
Hbe	Floresta Estacional Sempre Verde das Terras Baixas com dossel emergente
Hbu	Floresta Estacional Sempre Verde das Terras Baixas com dossel uniforme
Hs	Floresta Estacional Sempre Verde Submontana
Hse	Floresta Estacional Sempre Verde Submontana com dossel emergente
Hsu	Floresta Estacional Sempre Verde Submontana com dossel uniforme
La	Campinarana Arborizada
Lap	Campinarana Arborizada com palmeiras
Las	Campinarana Arborizada sem palmeiras
Ldp	Campinarana Florestada com palmeiras
Lds	Campinarana Florestada sem palmeiras
LOt	Contato Campinarana/Floresta Ombrófila - Ecótono
Ml	Floresta Ombrófila Mista Alto-Montana
Mm	Floresta Ombrófila Mista Montana
ONt	Contato Floresta Ombrófila/Floresta Estacional - Ecótono
R	Florestamento/Reflorestamento
Ra	Florestamento/Reflorestamento com Acácias
Re	Florestamento/Reflorestamento com Eucaliptos
Rp	Florestamento/Reflorestamento com Pinus
Rs	Florestamento/Reflorestamento com Seringueiras
SNt	Contato Savana/Floresta Estacional - Ecótono
SOt	Contato Savana/Floresta Ombrófila – Ecótono
TNt	Contato Savana-Estépica/Floresta Estacional – Ecótono

Tabela 3.5 – Filtro do *Layer* de Vegetação Alto Porte

LEGENDA	DESCRIÇÃO DA VEGETAÇÃO/ANTROPISMO
Ac	Agricultura/Média
Acc	Agricultura com Culturas Cíclicas/Média
Acp	Agricultura com Culturas Permanentes/Média
Ag	Agropecuária/Média
Lb	Campinarana Arbustiva
Lbp	Campinarana Arbustiva com palmeiras
Lbs	Campinarana Arbustiva sem palmeiras
Lg	Campinarana Gramíneo-Lenhosa
Lgs	Campinarana Gramíneo-Lenhosa sem palmeiras
Pa	Formação Pioneira com influência fluvial e/ou lacustre
Paa	Formação Pioneira com influência fluvial e/ou lacustre arbustiva
Paap	Formação Pioneira com influência fluvial e/ou lacustre arbustiva com palmeiras
Paas	Formação Pioneira com influência fluvial e/ou lacustre arbustiva sem palmeiras
Pah	Formação Pioneira com influência fluvial e/ou lacustre herbácea
Pahp	Formação Pioneira com influência fluvial e/ou lacustre herbácea com palmeiras
Pahs	Formação Pioneira com influência fluvial e/ou lacustre herbácea sem palmeiras
Pap	Formação Pioneira com influência fluvial e/ou lacustre palmeiral
Pf	Formação Pioneira com influência fluviomarinha
Pfh	Formação Pioneira com influência fluviomarinha herbácea
Pfm	Formação Pioneira com influência fluviomarinha arbórea
Pm	Formação Pioneira com influência marinha
Pma	Formação Pioneira com influência marinha arbórea
Pmb	Formação Pioneira com influência marinha arbórea
Pmh	Formação Pioneira com influência marinha herbácea
Sa	Savana Arborizada
Saf	Savana Arborizada com floresta-de-galeria
Sas	Savana Arborizada sem floresta-de-galeria
Sd	Savana Florestada
Sp	Savana Parque
Spf	Savana Parque com floresta-de-galeria

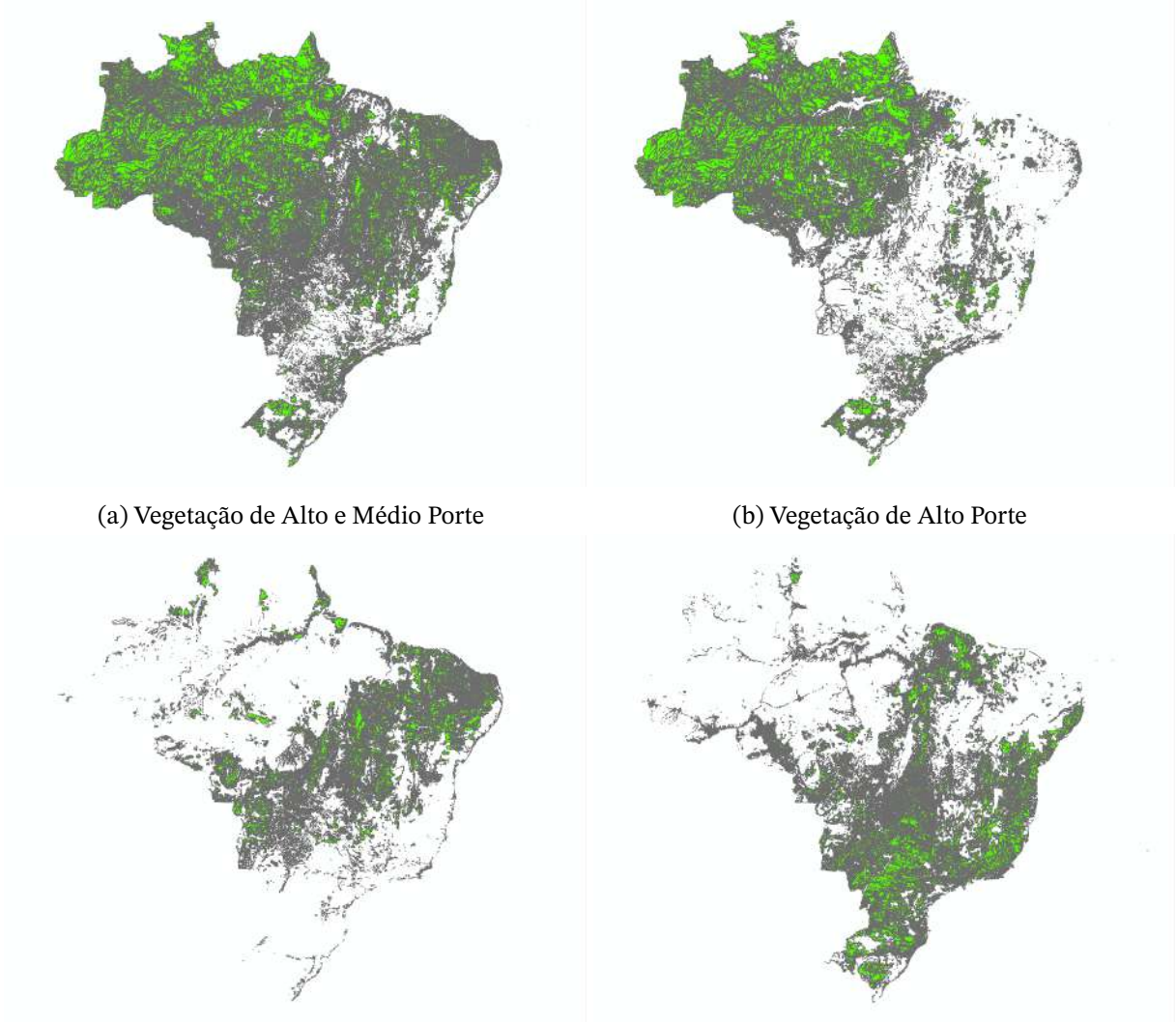
Sps	Savana Parque sem floresta-de-galeria
SPt	Contato Savana/Formações Pioneiras - Ecótono
STNt	Contato Savana/Savana-Estépica/Floresta Estacional – Ecótono
STt	Contato Savana/Savana-Estépica – Ecótono
Ta	Savana-Estépica Arborizada
Tas	Savana-Estépica Arborizada sem palmeiras e sem floresta-de-galeria
Td	Savana-Estépica Florestada
Tdp	Savana-Estépica Florestada com palmeiras
Tds	Savana-Estépica Florestada sem palmeiras
Tp	Savana-Estépica Parque
Tpf	Savana-Estépica Parque com floresta-de-galeria
Tpp	Savana-Estépica Parque com palmeiras
Tps	Savana-Estépica Parque sem palmeiras e sem floresta-de-galeria
TPt	Contato Savana-Estépica/Formações Pioneiras - Ecótono

Tabela 3.6 – Filtro do *Layer* de Vegetação Médio Porte

LEGENDA	**DESCRIÇÃO DA VEGETAÇÃO/ANTROPISMO**
Ac	Agricultura/Baixa
Acc	Agricultura com Culturas Cíclicas/Baixa
Acp	Agricultura com Culturas Permanentes/Baixa
Ag	Agropecuária/Baixa
Ap	Pecuária (pastagens)
Ar	Afloramento Rochoso
Dn	Dunas
Eaf	Estepe Arborizada com floresta-de-galeria
Eg	Estepe Gramíneo-Lenhosa
Egf	Estepe Gramíneo-Lenhosa com floresta-de-galeria
Egs	Estepe Gramíneo-Lenhosa sem floresta-de-galeria
Epf	Estepe Parque com floresta-de-galeria
rl	Refúgio Vegetacional Alto-Montano
rlh	Refúgio Vegetacional Alto-Montano herbáceo
rm	Refúgio Vegetacional Montano
rmb	Refúgio Vegetacional Montano arbustivo
rmh	Refúgio Vegetacional Montano herbáceo
rsb	Refúgio Vegetacional Submontano arbustivo
rsh	Refúgio Vegetacional Submontano herbáceo
Sg	Savana Gramíneo-Lenhosa

Sgf	Savana Gramíneo-Lenhosa com floresta-de-galeria
Sgs	Savana Gramíneo-Lenhosa sem floresta-de-galeria
Tgf	Savana-Estépica Gramíneo-Lenhosa com floresta-de-galeria
Tgp	Savana-Estépica Gramíneo-Lenhosa com palmeiras

Tabela 3.7 – Filtro do *Layer* de Vegetação Baixo Porte



(a) Vegetação de Alto e Médio Porte

(b) Vegetação de Alto Porte

(c) Vegetação de Médio Porte

(d) Vegetação de Baixo Porte

Figura 3.13 – Vegetação IBGE
Fonte: Autoria própria.

Assim, todos os atributos de terreno estavam prontos para entrar no formato do mapa de conjuntos. Duas funções do *software* ArcMap™ foram utilizadas para extrair os dados:

- *Clip*

O mapa de conjuntos é cortado no formato do atributo. Assim, calcula-se a área do atributo em km^2 presente no respectivo conjunto.

- *Spatial Join + Dissolve*

  O Spatial Join era responsável por ligar áreas do atributo com áreas do conjunto. Ao final, usa-se a função *Dissolve* com a funcionalidade "Sum"para que se pudesse somar a característica dentro dos conjuntos.

Nº	Nome do Atributo	Sigla	Código	Fonte
1	Área do Conjunto em km2	AREA_CONJ	T01	SIG-R
2	Área de Assentamentos Incra em km2	INCRA	T02	INCRA
3	Área de Massa D'Água em km2	MAGUA	T03	IBGE
4	Áreas vulneráveis a inundação em km	INUN	T04	ANA
5	Área de Quilombolas	QUI	T05	INCRA
6	Rodovias em km	RODO	T06	IBGE
7	Rodovias por área do conjunto km/km2	RODO_AREA_CONJ	T07	-
8	Área de Terra Indígenas	INDIGENA	T08	FUNAI
9	Unidades de Conservação	UNI_CONS	T09	MMA
10	Vegetação de alto porte	VEG_ALTA	T10	IBGE
11	Vegetação de médio porte	VEG_MED	T11	IBGE
12	Vegetação de baixo porte	VEG_BAIXA	T12	IBGE
13	Vegetação de alto e médio porte	VEG_AM	T13	IBGE
14	Percentual de Vegetação de alto porte	PC_VEG_ALTO	T14	-
15	Percentual de Vegetação de médio porte	PC_VEG_MED	T15	-
16	Percentual de Vegetação de baixo porte	PC_VEG_BAIXA	T16	-
17	Percentual de Vegetação de alto e médio porte	PC_VEG_AM	T17	-

Tabela 3.8 – Tabela de atributos relacionados a terreno.

3.2 Base de Dados Geográfica das Distribuidoras

Para a extração de dados da BDGD[5], as principais tabelas acessadas foram: Circuito de Média Tensão (CTMT), Equipamento Regulador (EQRE), Ramal de ligação (RAMLIG), Segmento do Sistema de Distribuição de Baixa Tensão (SSDBT), Segmento do Sistema de Distribuição de Média Tensão (SSDMT), Unidade Consumidora de Baixa Tensão (UCBT) e Unidade Transformadora de Média Tensão (UNTRMT).

A partir dessas tabelas e da utilização do *software* SAS Enterprise Guide com suas diversas funcionalidades na parte estatística, os dados foram montados com sua última operação sendo um *left join* com a base de conjuntos.

Nº	Nome do Atributo	Sigla	Código
1	Número Total de UCBT	NUC_BT	BDGD01
2	Número de UCBT por km de Rede BT	NUC_BT/kmBT	BDGD02

Nº	Nome do Atributo		Sigla	Código
3	Número de UCBT por Transformador MT/BT		NUC_BT/ TrafoMTBT	BDGD03
4	Número de UCBT por Área do Conjunto		NUC_BT/km2	BDGD04
5	Total de Energia Consumida		TEC	BDGD05
6	Total de Energia Consumida por km de Rede BT		TEC/kmBT	BDGD06
7	Número de UCBT por Classe	Residencial	NUC_BT_RES	BDGD07
8		Comercial	NUC_BT_COM	BDGD08
9		Industrial	NUC_BT_IND	BDGD09
10		Rural	NUC_BT_RUR	BDGD10
11		Poder Público	NUC_BT_PP	BDGD11
12		Consumo Próprio	NUC_BT_CP	BDGD12
13	Total de Energia Consumida por Classe	Residencial	TEC_RES	BDGD13
14		Comercial	TEC_COM	BDGD14
15		Industrial	TEC_IND	BDGD15
16		Rural	TEC_RUR	BDGD16
17		Poder Público	TEC_PP	BDGD17
18		Consumo Próprio	TEC_CP	BDGD18
19	Número de PIP		NPIP	BDGD19
20	Número PIP por km de Rede BT		NPIP/kmBT	BDGD20
21	Número de PIP por Área do Conjunto		NPIP/km2	BDGD21
22	Número de PIP por Transformador MT/BT		NPIP/TrafoMTBT	BDGD22
23	Total de Energia dos PIP		TEPIP	BDGD23
24	Total de Energia dos PIP por km de Rede BT		TEPIP/kmBT	BDGD24
25	Extensão de Rede BT em km	Sem Neutro	EXT_BT_SN	BDGD25
26		Com Neutro	EXT_BT_CN	BDGD26
27		Monofásica	EXT_BT_1F	BDGD27
28		Bifásica	EXT_BT_2F	BDGD28
29		Trifásica	EXT_BT_3F	BDGD29
30	Extensão de Rede BT por Área do Conjunto	Sem Neutro	EXT_BT_SN/km2	BDGD30
31		Com Neutro	EXT_BT_CN/km2	BDGD31
32		Monofásica	EXT_BT_1F/km2	BDGD32
33		Bifásica	EXT_BT_2F/km2	BDGD33
34		Trifásica	EXT_BT_3F/km2	BDGD34
35	Extensão de Rede MT em km	Monofásica	EXT_MT_1F	BDGD35
36		Bifásica	EXT_MT_2F	BDGD36
37		Trifásica	EXT_MT_3F	BDGD37

Nº	Nome do Atributo		Sigla	Código
38		Trifásica com isolamento	EXT_MT_3F_C_ISO	BDGD38
39		Trifásica sem isolamennto	EXT_MT_3F_S_ISO	BDGD39
40	Número de Transformadores MT	Monofásico	TRAFO_1F	BDGD40
41		Bifásico	TRAFO_2F	BDGD41
42		Trifásico	TRAFO_3F	BDGD42
43		Total	TRAFO	BDGD43
44	Soma da Potência Nominal	Monofásico	POT_NOM_1F	BDGD44
45		Bifásico	POT_NOM_2F	BDGD45
46		Trifásico	POT_NOM_3F	BDGD46
47		Total	POT_NOM	BDGD47
48	Transformadores MT por km de Rede BT	Monofásico	TRAFO_kmBT_1F	BDGD48
49		Bifásico	TRAFO_kmBT_2F	BDGD49
50		Trifásico	TRAFO_kmBT_3F	BDGD50
51		Total	TRAFO_kmBT	BDGD51
52	Potência Nominal por km de Rede BT	Monofásico	POT_NOM_kmBT_1F	BDGD52
53		Bifásico	POT_NOM_kmBT_2F	BDGD53
54		Trifásico	POT_NOM_kmBT_3F	BDGD54
55		Total	POT_NOM_kmBT	BDGD55
56	Potência Nominal Média de Trafos MT		AVG_POT_NOM	BDGD56
57	Ramal de ligação em km	Monofásico	RAMLIG_1F	BDGD57
58		Bifásico	RAMLIG_2F	BDGD58
59		Trifásico	RAMLIG_3F	BDGD59
60	Equipamento Regulador		EQRE	BDGD60

Tabela 3.9 – Tabela de atributos relacionados a rede elétrica.

3.3 Climáticos

Os dados climáticos extraídos foram retirados de duas fontes: Agência Nacional de Águas e Saneamento Básico [4](ANA) e Instituto Nacional de Metereologia[13](INMET).

As manipulações desses dados foram feitas nos dois *softwares* já citados: SAS Enterprise Guide e ArcMap™.

Da extração dos dados a construção do atributo, houveram algumas etapas a serem seguidas:

1. Dados

   As duas bases de dados extraídas desses órgãos possuem dados climáticos. Da base da ANA foram primariamente extraídos dados pluviométricos, pois a quantidade dos demais dados climáticos é ínfima. Já da base do INMET foram extraídos os dados de Precipitação total, Temperatura do Ar, Vento Rajada Máxima e Vento Velocidade Horária.

2. Formato dos Dados

   Os dados de pluviometria vindo da ANA estão num formato de 3.398.481 registros mensais de aproximadamente 4000 estações climáticas durante diversos anos. Os dados vindo do INMET totalizam 46.204.056 registros horários de aproximadamente 627 estações durante os 10 anos (2012-2021).

3. Tratamento dos Dados

   Nessa etapa, dados duplicados, dados faltantes, dados enviesados foram retirados da análise. Além disso, optou-se por deixar os dados das duas bases no formato diário, para que pudesse ser calculado os percentis. Para o atributo de pluviometria, juntou-se as estações das bases da ANA com as bases do INMET, formando assim 13.528.771 registros diários de 4768 estações. Para os dados de Temperatura e Ventos, manteve-se o número de 627 estações.

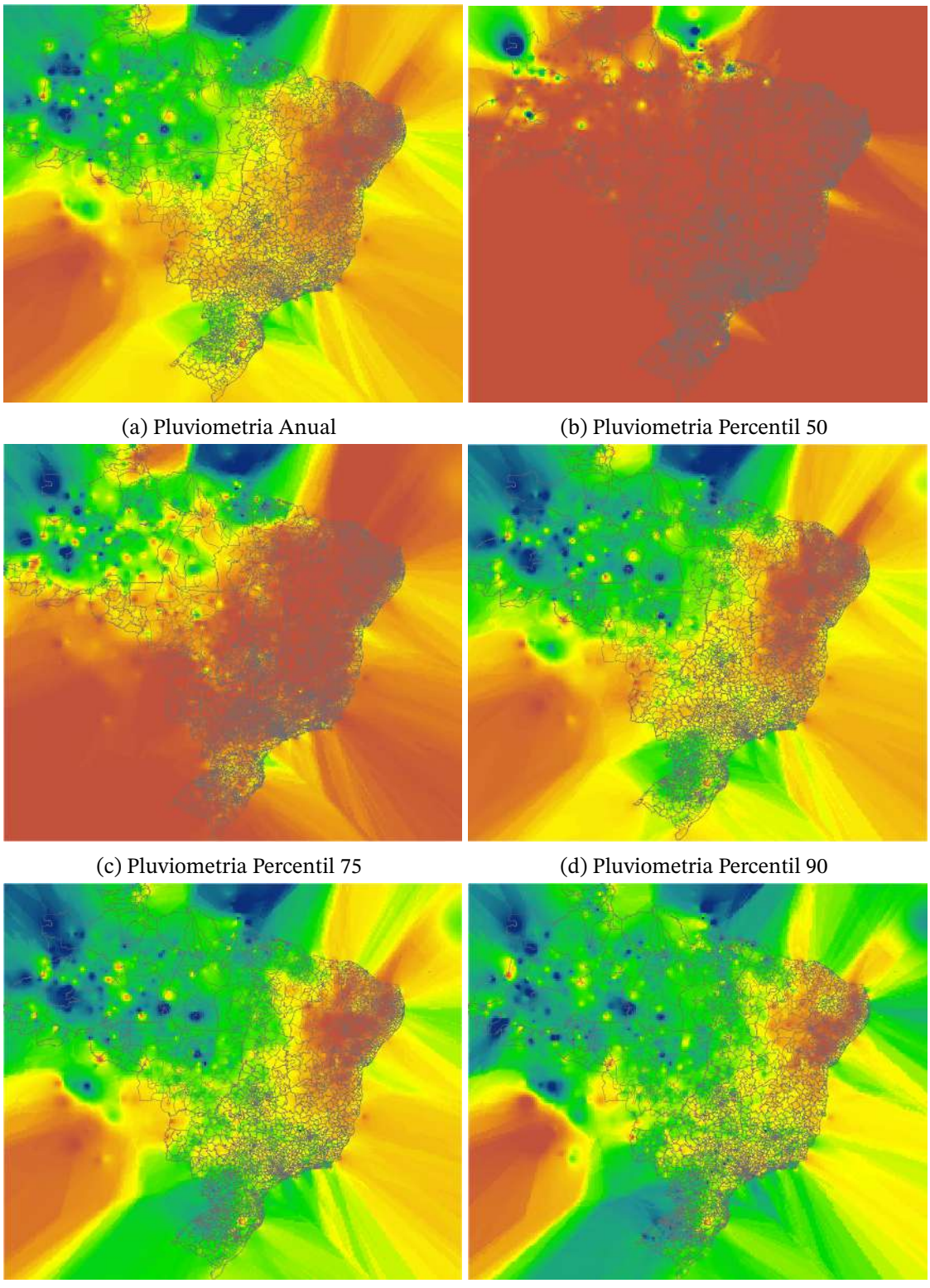
   Para formar o dado de Pluviometria Anual, optou-se por criar uma estimativa do comportamento mensal decenal, ou seja, entender o comportamento de todos os meses durante os 10 anos de dados. Isso foi realizado porque havia discrepância entre o número de anos que as estações estavam funcionando, devido a algumas estações serem mais recentes.

   Assim, usou-se a ferramenta *Summary Statistics* e foram criados os atributos de pluviometria anual e seus percentis com registros diários utilizando 4768 estações. Os atributos de temperatura média, temperatura máxima, temperatura mínima, vento médio, vento rajada máxima e seus devidos percentis foram criados a partir dos dados díarios utilizando 627 estações.

4. Inverse Distance Weighting

   Logo em seguida, com os dados pontuais das estações, usou-se o ArcMap™ para espalhá-las sobre o mapa do Brasil. Para realizar a interpolação, foi usado um método

chamado: *Inverse Distance Weighting*[23], que é um método de interpolação determinista responsável por entender o comportamento do atributo entre os espaços onde não há estações. Abaixo estão as figuras *Raster* dos atributos selecionados.



(a) Pluviometria Anual

(b) Pluviometria Percentil 50

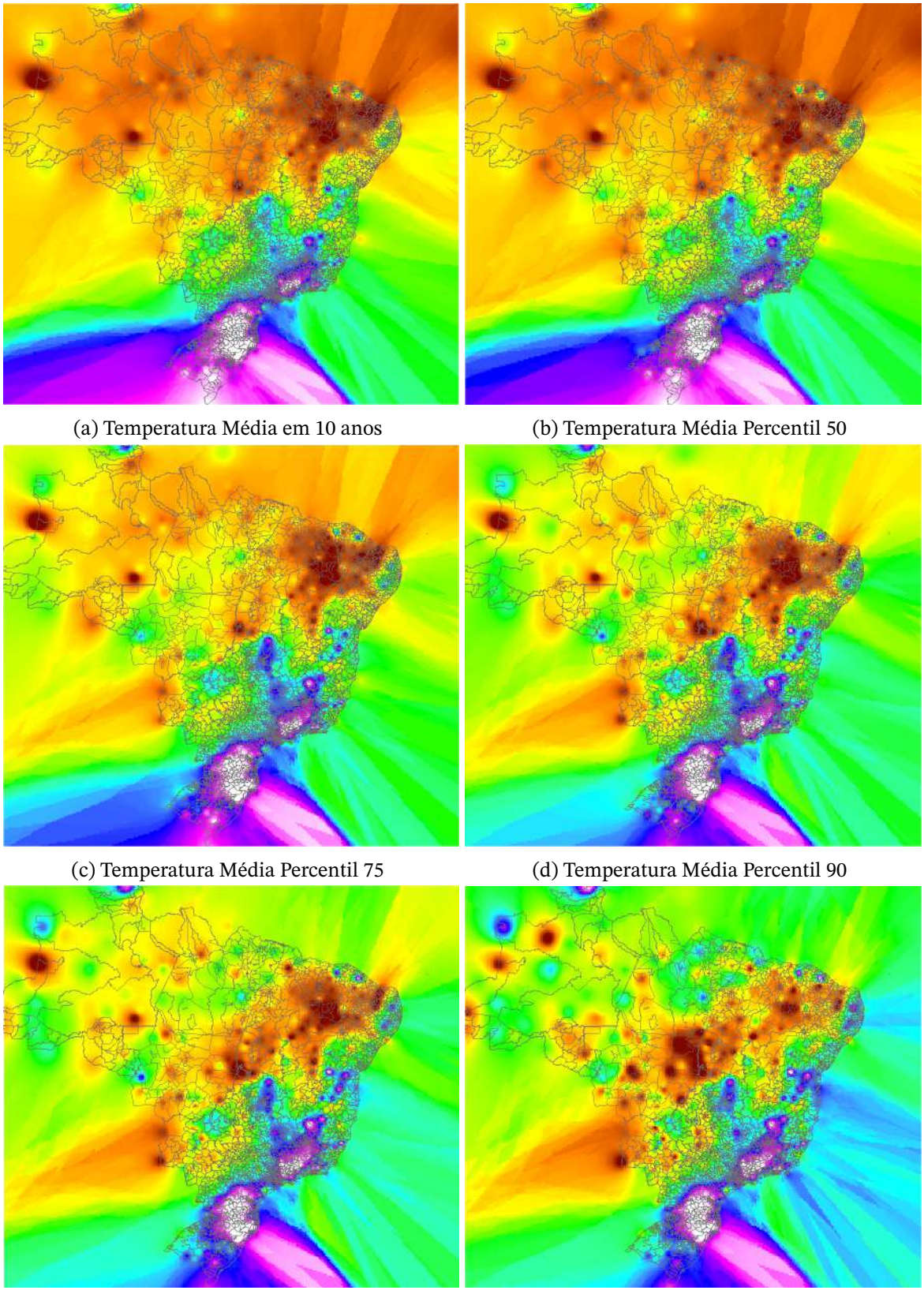
(c) Pluviometria Percentil 75

(d) Pluviometria Percentil 90

(e) Pluviometria Percentil 95

(f) Pluviometria Percentil 99

Figura 3.14 – Pluviometrias
Fonte: Autoria própria.



(a) Temperatura Média em 10 anos

(b) Temperatura Média Percentil 50

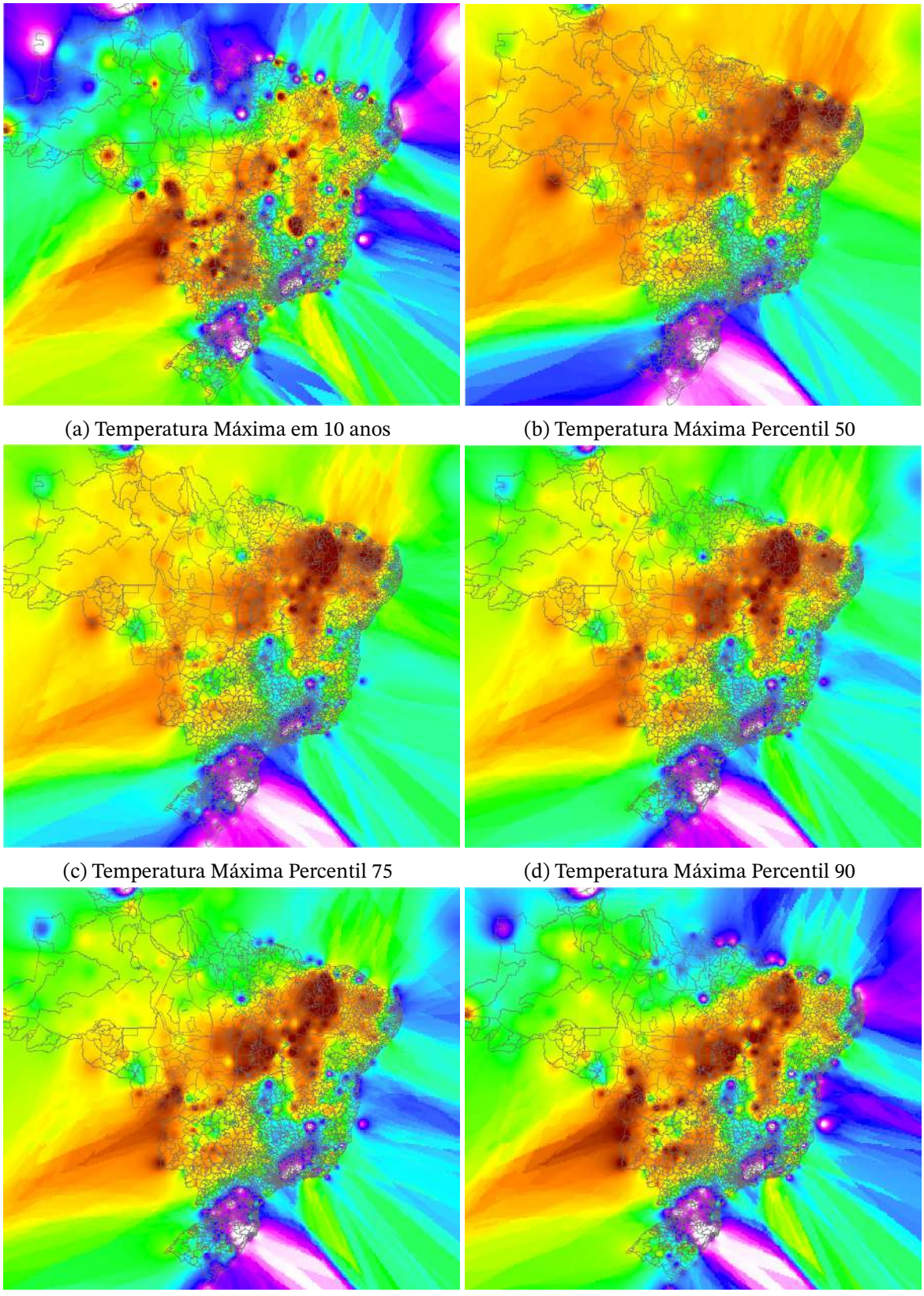
(c) Temperatura Média Percentil 75

(d) Temperatura Média Percentil 90

(e) Temperatura Média Percentil 95

(f) Temperatura Média Percentil 99

Figura 3.15 – Temperaturas Médias
Fonte: Autoria própria.



(a) Temperatura Máxima em 10 anos

(b) Temperatura Máxima Percentil 50

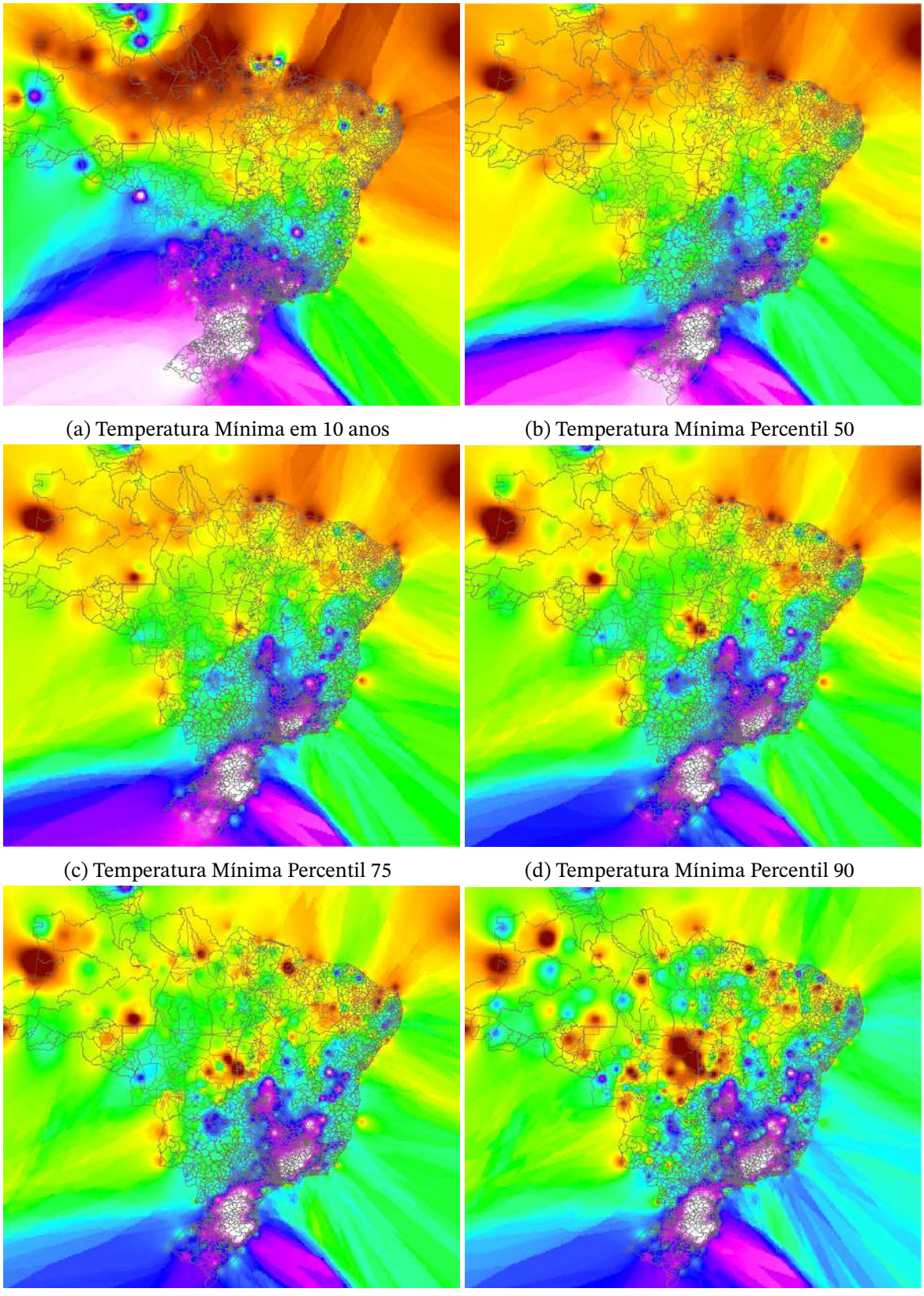
(c) Temperatura Máxima Percentil 75

(d) Temperatura Máxima Percentil 90

(e) Temperatura Máxima Percentil 95

(f) Temperatura Máxima Percentil 99

Figura 3.16 – Temperaturas Máximas
Fonte: Autoria própria.



(a) Temperatura Mínima em 10 anos

(b) Temperatura Mínima Percentil 50

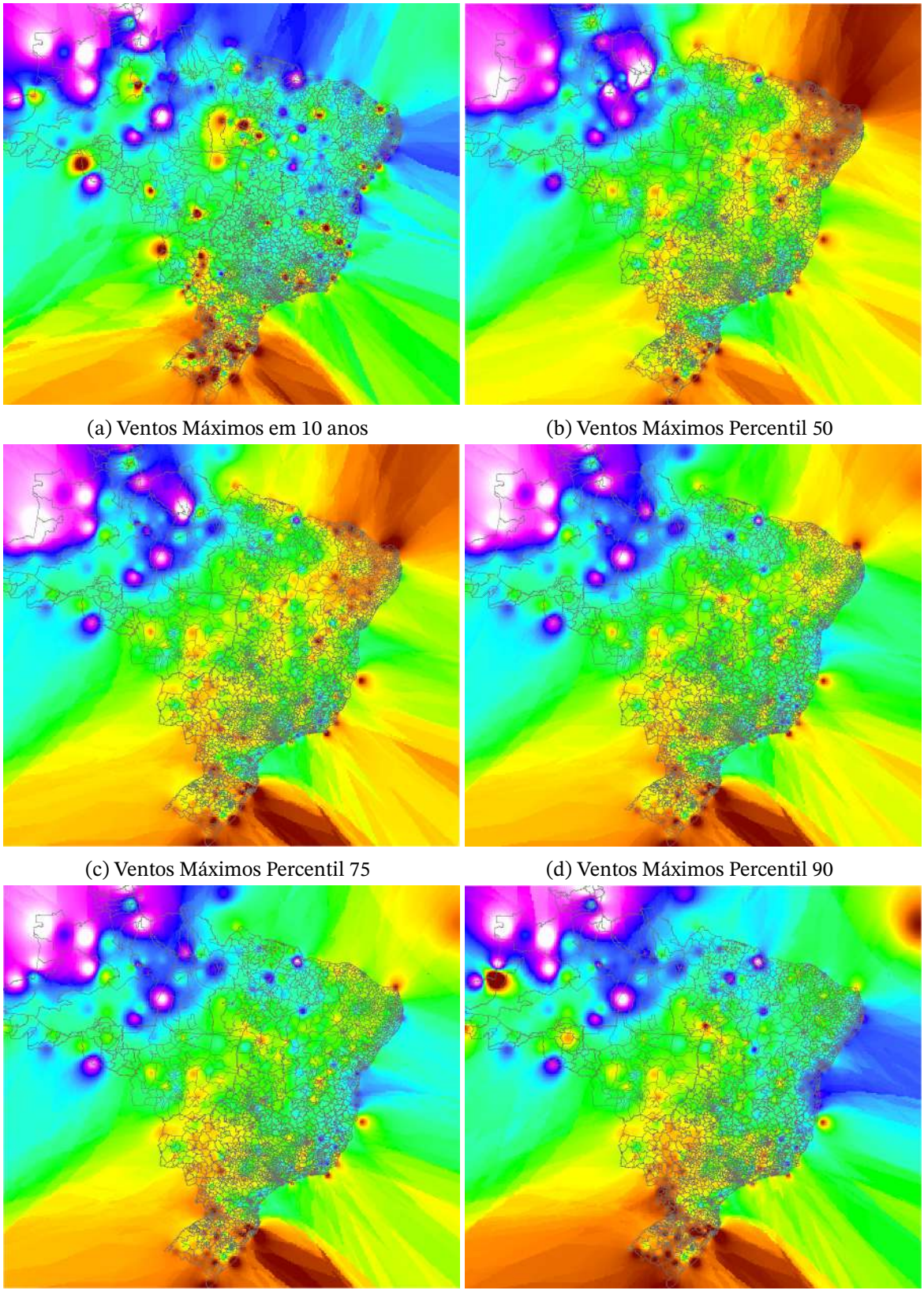
(c) Temperatura Mínima Percentil 75

(d) Temperatura Mínima Percentil 90

(e) Temperatura Mínima Percentil 95

(f) Temperatura Mínima Percentil 99

Figura 3.17 – Temperaturas Mínimas
Fonte: Autoria própria.



(a) Ventos Máximos em 10 anos

(b) Ventos Máximos Percentil 50

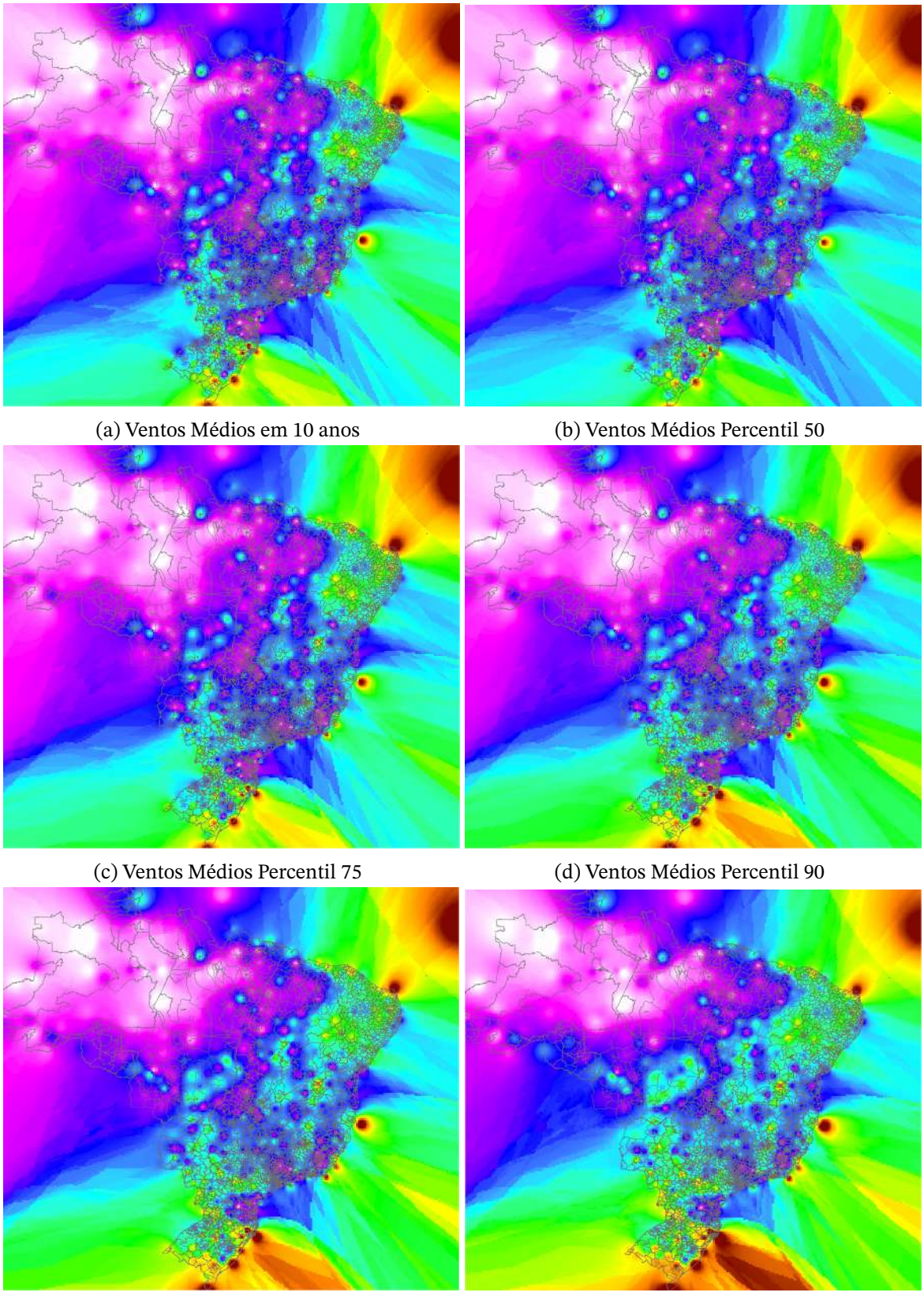
(c) Ventos Máximos Percentil 75

(d) Ventos Máximos Percentil 90

(e) Ventos Máximos Percentil 95

(f) Ventos Máximos Percentil 99

Figura 3.18 – Ventos Máximos
Fonte: Autoria própria.



(a) Ventos Médios em 10 anos

(b) Ventos Médios Percentil 50

(c) Ventos Médios Percentil 75

(d) Ventos Médios Percentil 90

(e) Ventos Médios Percentil 95

(f) Ventos Médios Percentil 99

Figura 3.19 – Ventos Médios
Fonte: Autoria própria.

5. *Raster to Polygon*

Por fim, usa-se uma função do ArcMap chamada *Raster to Polygon*, que é responsável por transformar o *Raster* em Polígonos que contenham os valores dos atributos. Assim, basta realizar um *Spatial Join* e um *Dissolve* no mapa de conjuntos e o atributo estará em função dele.

Esta é a tabela final dos atributos relacionados aos conjuntos das distribuidoras.

Nº	**Nome do Atributo**		**Sigla**	**Código**
1	Pluviometria	Anual	PLUV_ANUAL	C01
2		Percentil 50	PLUV_P50	C02
3		Percentil 75	PLUV_P75	C03
4		Percentil 90	PLUV_P90	C04
5		Percentil 95	PLUV_P95	C05
6		Percentil 99	PLUV_P99	C06
7	Temperatura Média	10 anos	TEMP_MED	C07
8		Percentil 50	TEMP_MED_P50	C08
9		Percentil 75	TEMP_MED_P75	C09
10		Percentil 90	TEMP_MED_P90	C10
11		Percentil 95	TEMP_MED_P95	C11
12		Percentil 99	TEMP_MED_P99	C12
13	Temperatura Máxima	10 anos	TEMP_MAX	C13
14		Percentil 50	TEMP_MAX_P50	C14
15		Percentil 75	TEMP_MAX_P75	C15
16		Percentil 90	TEMP_MAX_P90	C16
17		Percentil 95	TEMP_MAX_P95	C17
18		Percentil 99	TEMP_MAX_P99	C18
19	Temperatura Mínima	10 anos	TEMP_MIN	C19
20		Percentil 50	TEMP_MIN_P50	C20
21		Percentil 75	TEMP_MIN_P75	C21
22		Percentil 90	TEMP_MIN_P90	C22
23		Percentil 95	TEMP_MIN_P95	C23
24		Percentil 99	TEMP_MIN_P99	C24
25	Vento Máximo	10 anos	VENTO_MAX	C25
26		Percentil 50	VENTO_MAX_P50	C26
27		Percentil 75	VENTO_MAX_P75	C27
28		Percentil 90	VENTO_MAX_P90	C28
29		Percentil 95	VENTO_MAX_P95	C29
30		Percentil 99	VENTO_MAX_P99	C30
31	Vento Médio	10 anos	VENTO_MED	C31
32		Percentil 50	VENTO_MED_P50	C32
33		Percentil 75	VENTO_MED_P75	C33
34		Percentil 90	VENTO_MED_P90	C34
35		Percentil 95	VENTO_MED_P95	C35
36		Percentil 99	VENTO_MED_P99	C36

Tabela 3.10 – Tabela de atributos referentes ao clima.

3.4 Socioeconômicos

Por fim, a última categoria de dados, que são os socioeconômicos, estavam relacionados aos municípios do Brasil. Assim, de forma suscinta, criou-se 52 mapas (1 para cada distribuidora), atráves de uma função chamada *Intersect*, que relacionam espacialmente a área dos conjuntos com os municípios do país.

Em seguida, o *Spatial Join* foi utilizado para juntar o mapa de Conjuntos e Municípios com o número de unidades consumidoras que estavam inseridas nessa região. Assim, como ressalta a Nota técnica 0102/2014[6]: "a melhor aproximação consiste em ponderar as variáveis de acordo com a quantidade de unidades consumidoras que o conjunto possui em cada um dos municípios que possuem intersecção com a área do conjunto".

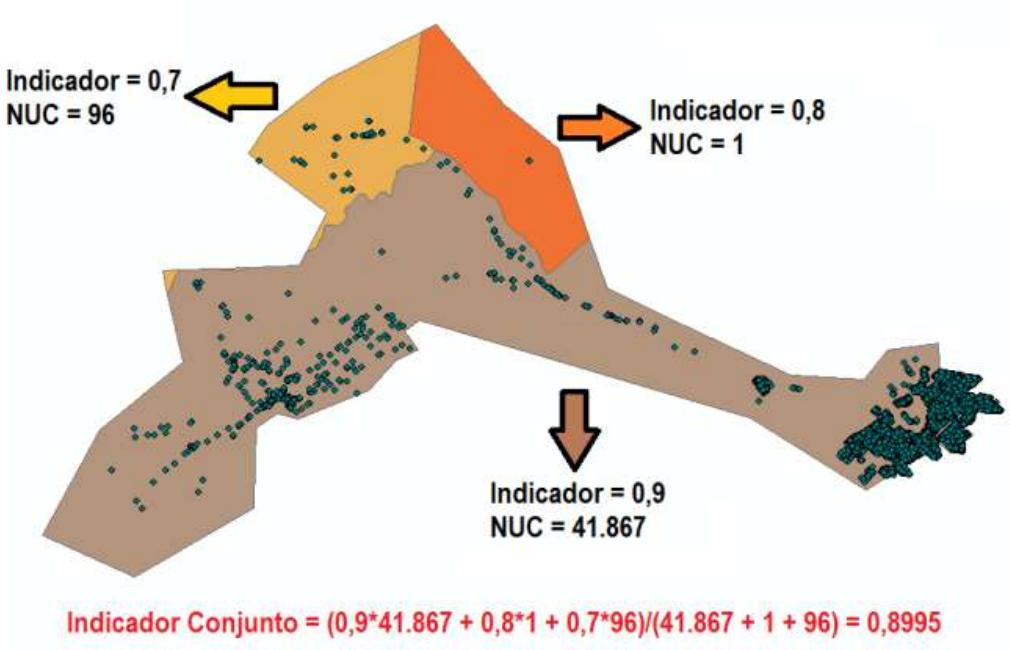


Figura 3.20 – Transformação da variável de município para conjunto
Fonte: Nota técnica ANEEL [6]

Assim, utilizando o *software* Excel, foram obtidos os atributos socioeconômicos. Por motivos de formatação e extensão dos nomes, foi decidido deixar apenas as siglas dos atributos socioeconômicos. As fontes dos atributos podem ser encontrados anexos a esse trabalho.

Nº	Sigla	Código
1	ESPVIDA	SE001
2	MORT1	SE002
3	T_ENV	SE003
4	E_ANOSESTUDO	SE004

Nº	Sigla	Código
5	CORTE1	SE005
6	CORTE2	SE006
7	CORTE3	SE007
8	CORTE4	SE008
9	CORTE9	SE009
10	GINI	SE010
11	PIND	SE011
12	PMPOB	SE012
13	PPOB	SE013
14	PREN10RICOS	SE014
15	PREN20	SE015
16	RDPCT	SE016
17	RDPC1	SE017
18	RDPC10	SE018
19	THEIL	SE019
20	T_AGUA	SE020
21	T_BANAGUA	SE021
22	T_DENS	SE022
23	T_LIXO	SE023
24	T_LUZ	SE024
25	AGUA_ESGOTO	SE025
26	PAREDE	SE026
27	IDHM	SE027
28	IDHM_ E	SE028
29	IDHM_L	SE029
30	IDHM_R	SE030
31	HOMICIDIOS	SE031
32	MORT_VIOL	SE032
33	TAX_HOMIC	SE033
34	TAX_MVCI	SE034
35	OBITOS_AGR	SE035
36	VIOLENCIA	SE036
37	VIOLENCIA_FIS	SE037
38	DOM_URB_TOT	SE038
39	DOM_URB_PROP	SE039
40	DOM_URB_ALUG	SE040
41	DOM_URB_CED	SE041

Nº	Sigla	Código
42	DOM_URB_OUT_OCUP	SE042
43	M_DOM_URB_TOT	SE043
44	M_DOM_URB_PROP	SE044
45	M_DOM_URB_ALUG	SE045
46	M_DOM_URB_CED	SE046
47	M_DOM_URB_OUT	SE047
48	DOM_U_RED_AGUA	SE048
49	DOM_U_AGUA_POÇO_NAS	SE049
50	DOM_U_AGUA_OUT	SE050
51	M_DOM_U_RED_AGUA	SE051
52	M_DOM_U_AGUA_POÇO_NAS	SE052
53	M_DOM_U_AGUA_OUT	SE053
54	DOM_U_BANH_TOT	SE054
55	DOM_U_BANH_RED_ESG_PLUV	SE055
56	DOM_U_BANH_FOSSA	SE056
57	DOM_U_BANH_ESG_OUT	SE057
58	DOM_U_BANH_SEM_ESG	SE058
59	M_DOM_U_BANH_TOT	SE059
60	M_DOM_U_BANH_RED_ESG_PLUV	SE060
61	M_DOM_U_BANH_FOSSA	SE061
62	M_DOM_U_BANH_ESG_OUT	SE062
63	M_DOM_U_BANH_SEM_ESG	SE063
64	DOM_U_LIXO_COL_TOT	SE064
65	DOM_U_LIXO_SERV_LIMP	SE065
66	DOM_U_LIXO_CAÇAMBA	SE066
67	DOM_U_SEM_LIXO_COL	SE067
68	M_DOM_U_LIXO_COL_TOT	SE068
69	M_DOM_U_LIXO_SERV_LIMP	SE069
70	M_DOM_U_LIXO_CAÇAMBA	SE070
71	M_DOM_U_SEM_LIXO_COL	SE071
72	DOM_U_ADEQ	SE072
73	DOM_U_SEMI_ADEQ	SE073
74	DOM_U_INADEQ	SE074
75	M_DOM_U_ADEQ	SE075
76	M_DOM_U_SEMI_ADEQ	SE076
77	M_DOM_U_INADEQ	SE077
78	RDPC_ATÉ_1/4_SM	SE078

Nº	Sigla	Código
79	RDPC_+1/4_A_1/2_SM	SE079
80	RDPC_+1/2_A_1_SM	SE080
81	RDPC_+1_A_2_SM	SE081
82	RDPC_+2_SM	SE082
83	RDPC_SEM_REND	SE083
84	M_RDPC_ATÉ_1/4_SM	SE084
85	M_RDPC_+1/4_A_1/2_SM	SE085
86	M_RDPC_+1/2_A_1_SM	SE086
87	M_RDPC_+1_A_2_SM	SE087
88	M_RDPC_+2_SM	SE088
89	M_RDPC_SEM_REND	SE089
90	POP	SE090
91	%_POP_URB	SE091
92	%_POP_RUR	SE092
93	DOM	SE093
94	%_DOM_SAN_AD	SE094
95	%_DOM_SAN_SEMI-AD	SE095
96	%_DOM_SAN_INAD	SE096
97	RDPC	SE097
98	RDPC_Q1	SE098
99	RDPC_Q2	SE099
100	RDPC_Q3	SE100
101	POP_DOM	SE101
102	%_POP_DOM_RDPC_R$70	SE102
103	%_POP_DOM_RDPC_R$127,5	SE103
104	%_POP_DOM_RDPC_R$255	SE104
105	%_POP_DOM_RDPC_R$225	SE105
106	TX_ANALF_>15	SE106
107	PIB	SE107
108	PIB_PC	SE108
109	POP14	SE109
110	POP15	SE110
111	POP16	SE111
112	POP17	SE112
113	POP18	SE113
114	POP19	SE114
115	POP20	SE115

Nº	Sigla	Código
116	POP21	SE116
117	POP_URB	SE117
118	POP_RUR	SE118
119	PES_>10	SE119
120	PES_>10_ALF	SE120
121	TAX_ALF_>10	SE121
122	DOM_URB	SE122
123	DOM_RUR	SE123
124	M_DOM	SE124
125	M_DOM_URB	SE125
126	M_DOM_RUR	SE126
127	MED_M_DOM	SE127
128	MED_M_DOM_URB	SE128
129	MED_M_DOM_RUR	SE129
130	DOM_BANH	SE130
131	DOM_BANH_RED_ESG	SE131
132	DOM_BANH_FOSSA	SE132
133	DOM_BANH_OUT	SE133
134	DOM_SEM_BANH	SE134
135	DOM_RED_AGUA	SE135
136	DOM_AGUA_POÇO_NAS	SE136
137	DOM_AGUA_OUT	SE137
138	DOM_LIXO_COL_TOT	SE138
139	DOM_LIXO_SERV_LIMP	SE139
140	DOM_LIXO_CAÇAMBA	SE140
141	DOM_SEM_LIXO_COL	SE141
142	PES_>10_ATÉ_1/2_SM	SE142
143	PES_>10_>1/2-1_SM	SE143
144	PES_>10_>1-2_SM	SE144
145	PES_>10_>2-5_SM	SE145
146	PES_>10_>5-10_SM	SE146
147	PES_>10_>10-20_SM	SE147
148	PES_>10_>20_SM	SE148
149	PES_>10_SEM_REND	SE149
150	DOM_ATÉ_1/2_SM	SE150
151	DOM_>1/2-1_SM	SE151
152	DOM_>1-2_SM	SE152

Nº	Sigla	Código
153	DOM_>2-5_SM	SE153
154	DOM_>5-10_SM	SE154
155	DOM_>10-20_SM	SE155
156	DOM_>20_SM	SE156
157	DOM_SEM_REND	SE157
158	DOM_RDPC_ATÉ_1/4_SM	SE158
159	DOM_RDPC_>1/4-1/2_SM	SE159
160	DOM_RDPC_>1/2-1_SM	SE160
161	DOM_RDPC_>1-2_SM	SE161
162	DOM_RDPC_>2-3_SM	SE162
163	DOM_RDPC_>3-5_SM	SE163
164	DOM_RDPC_>5_SM	SE164
165	DOM_RDPC_SEM_REND	SE165

Tabela 3.11 – Tabela de atributos relacionados a dados socioeconômicos.

Dessa forma, tendo todos os atributos reunidos em um mesmo formato, basta juntá-los em uma grande tabela. A tabela final para esse estudo terá o formato da Tabela 2.4 feita no capítulo 2 desse documento. Com dimensões 283 x 3061, onde 278 são as variáveis independentes (atributos) e 2 são as variáveis dependentes (DEC e FEC).

Antes do teste dos modelos propostos, os dados passaram por um tratamento padrão dentro dos estudos de Aprendizado de Máquina. Campos com dados faltantes foram preenchidos com a média do respectivo atributo, com o intuito de não mudar a média e desvio padrão dos dados e melhorar a performance dos modelos. Além disso, as variáveis independentes DEC e FEC foram extraídas como a média dos dois últimos anos (2022 e 2021), para que um ano atípico de alguma distribuidora pudesse ser compensado, de certa forma, pelo comportamento do outro ano. Por fim, os dados foram normalizados com um MinMaxScaler e separados em 75% para treinos e 25% para testes.

4 Resultados

A opção por não incorporar o código Python diretamente na dissertação, mas disponibilizar em um repositório no GitHub, visa otimizar a clareza e a acessibilidade do documento. Portanto, optou-se por essa opção e foi criado um repositório no GitHub com os dois códigos e a base de dados utilizada. A seguir está o link do repositório: https://github.com/giba0604/analysis_decfec

4.1 Resultado MLP Regressor

A sensibilidade dos modelos de Rede Neural Artificial foi testada para diferentes números de camadas ocultas. Vale lembrar que, para as Redes Neurais Artificias, não existe uma regra geral que selecione o número ideal de camadas ocultas. Dessa forma, foram testados modelos com diferentes números de neurônios, camadas, funções de ativação e taxas de aprendizado. Por fim, alguns dos melhores modelos para esse trabalho em específico foram selecionados.

MLP DEC (learning rate init = $10e^{-}5$)		
Função de ativação	Hidden Layers (150,100,50)	Hidden Layers (100,75,50)
'logistic'	$R^2 = 0.653$ MAE = 4.585 MSE = 65.307 Tempo = 350.53s	$R^2 = 0.660$ MAE = 4.423 MSE = 63.987 Tempo = 428.93s
'tanh'	$R^2 = 0.647$ MAE =4.362 MSE = 66.359 Tempo = 250.59s	$R^2 = 0.659$ MAE = 4.324 MSE = 64.222 Tempo = 201.96s
'identity'	$R^2 = 0.544$ MAE =5.862 MSE = 85.917 Tempo = 40.16s	$R^2 = 0.536$ MAE = 5.850 MSE = 87.378 Tempo = 43.96s
'relu'	$R^2 = 0.640$ MAE = 3.973 MSE = 67.698 Tempo = 145.37s	$R^2 = 0.678$ MAE = 3.971 MSE = 60.545 Tempo = 112.92s

Tabela 4.12 – Tabela comparativa do modelo MLP para o indicador DEC

MLP FEC (learning rate init = $10e^{-}5$)		
Função de ativação	Hidden Layers (150,100,50)	Hidden Layers (100,75,50)
'logistic'	$R^2 = 0.643$ MAE = 1.898 MSE = 9.823 Tempo = 568.28s	$R^2 = 0.647$ MAE = 1.885 MSE = 9.711 Tempo = 382.77s
'tanh'	$R^2 = 0.630$ MAE = 1.827 MSE = 10.173 Tempo = 223.85s	$R^2 = 0.622$ MAE = 1.799 MSE = 10.402 Tempo = 158.07s
'identity'	$R^2 = 0.563$ MAE = 2.208 MSE = 12.017 Tempo = 29.96s	$R^2 = 0.558$ MAE = 2.215 MSE = 12.174 Tempo = 31.65s
'relu'	$R^2 = 0.687$ MAE = 1.688 MSE = 8.629 Tempo = 85.67s	$R^2 = 0.653$ MAE = 1.718 MSE = 9.553 Tempo = 85.98s

Tabela 4.13 – Tabela comparativa do modelo MLP para o indicador FEC

Analisando os 8 modelos para a variável independente DEC e os 8 para o FEC, temos resultados satisfatórios, levando em conta que a Metodologia[6] atual possui R^2 (com 6 variáveis no modelo) iguais a 0.59 e 0.6 para DEC e FEC, respectivamente. Os melhores modelos utilizaram a função de ativação ReLU, que tem tido grande sucesso nos problemas preditores de Aprendizado de Máquina.

Percebe-se que para o caso do DEC, a solução com menos neurônios nas camadas ocultas produziu um resultado mais satisfatório, de modo que aumentar a robustez das camadas ocultas começou a causar um leve caminho ao *Overfitting*. Isso pode nos levar a conclusão de que o modelo do DEC precisa de menos variáveis independentes para poder explicar seu comportamento e, além disso, ainda existem variáveis importantes a serem aplicadas dentro desse modelo.

Em compensação, para o caso do FEC, a solução com mais neurônios nas camadas ocultas produziu um melhor resultado. Ou seja, diminuir os neurônios das camadas ocultas levaria a um leve *Underfitting*.

Enfim, esse comportamento foi apresentado no estudo feito em 2014 pela ANEEL[6], onde o modelo de DEC selecionou 9 variáveis independentes e o de FEC 15.

4.2 Resultado *Decision Tree Regressor*

A sensibilidade das árvores de decisão foi testada a partir de duas variáveis: método de redução de variância das equações (2.4)(2.6) e profundidade da árvore. A variável *max depth* é o critério de parada utilizado pelo algoritmo.

Decision Tree DEC		
max depth	'squared error'	'absolute error'
3	$R^2 = 0.509$ MAE = 5.910 MSE = 92.490 Tempo = 0.19s	$R^2 = 0.524$ MAE = 5.604 MSE = 89.583 Tempo = 4.68s
4	$R^2 = 0.504$ MAE = 5.830 MSE = 93.380 Tempo = 0.25s	$R^2 = 0.564$ MAE = 5.285 MSE = 82.078 Tempo = 5.28s
5	$R^2 = 0.492$ MAE = 5.677 MSE = 95.613 Tempo = 0.31s	$R^2 = 0.527$ MAE = 5.334 MSE = 89.110 Tempo = 5.72s

Tabela 4.14 – Tabela comparativa do modelo Decision Tree para o indicador DEC

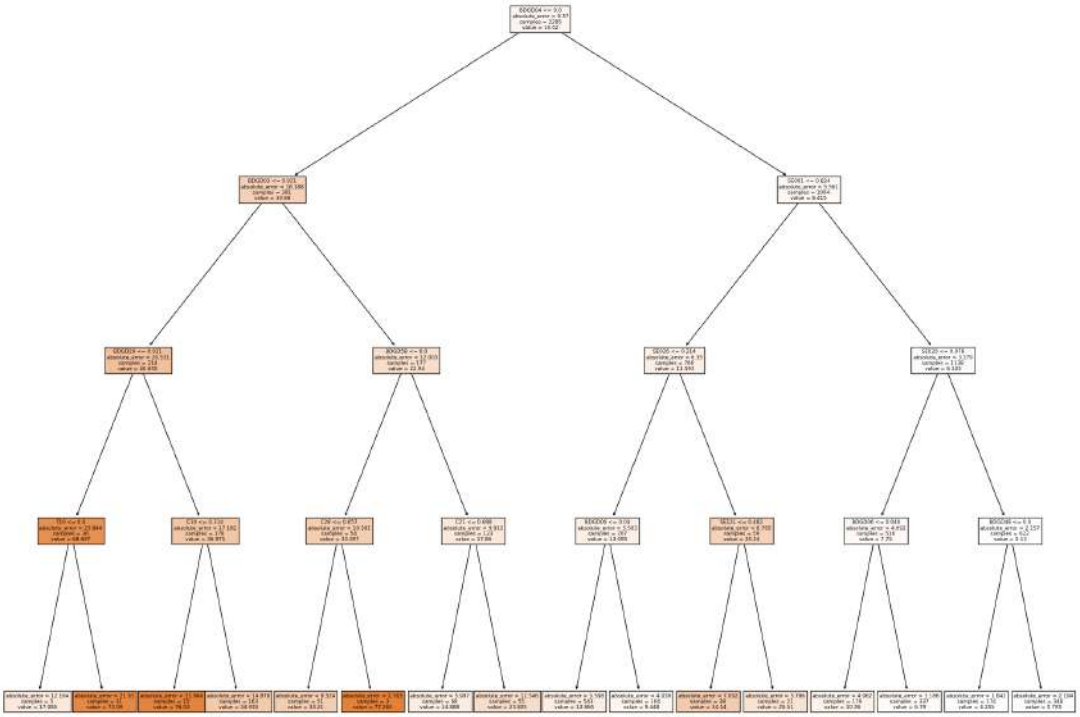


Figura 4.21 – Árvore de Decisão para o modelo DEC

Feature	Gini Importance
T10:	2.0748%
BDGD03:	6.4900%
BDGD04:	41.6036%
BDGD06:	3.6943%
BDGD29:	4.4280%
BDGD48:	0.7999%
BDGD58:	3.6424%
C21:	1.3308%
C28:	1.4090%
C30:	3.3226%
SE020:	3.0426%
SE026:	4.9909%
SE091:	21.7909%
SE121:	1.3793%

Tabela 4.15 – Variáveis do modelo DEC e seus índices de Gini

Através dos testes em diferentes profundidades e diferentes métodos de redução de variância, o resultado mais sofisticado possui profundidade igual a 4 camadas e *absolute error* como método de redução de variância. Notou-se que tentar resumir o problema com 3 camadas causou uma tendência ao *Underfitting* e torná-lo mais robusto com 5 camadas, para o *Overfitting*.

Para o modelo de FEC, temos:

Decision Tree FEC		
max depth	'squared error'	'absolute error'
3	$R^2 = 0.417$ MAE = 2.476 MSE = 16.060 Tempo = 0.20s	$R^2 = 0.525$ MAE = 2.194 MSE = 13.076 Tempo = 4.71s
4	$R^2 = 0.425$ MAE = 2.343 MSE = 15.815 Tempo = 0.26s	$R^2 = 0.541$ MAE = 2.139 MSE = 12.635 Tempo = 5.33s
5	$R^2 = 0.520$ MAE = 2.252 MSE = 13.203 Tempo = 0.32s	$R^2 = 0.507$ MAE = 2.142 MSE = 13.567 Tempo = 5.81s

Tabela 4.16 – Tabela comparativa do modelo Decision Tree para o indicador FEC

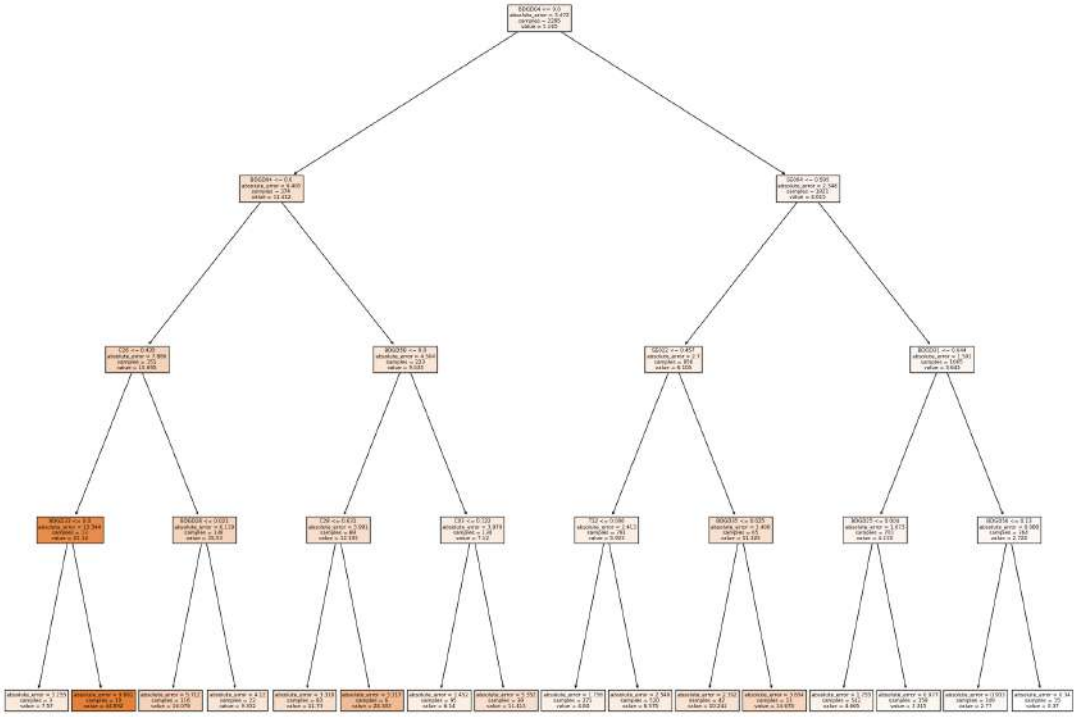


Figura 4.22 – Árvore de Decisão para o modelo FEC

Feature	Gini Importance
T12:	3.6639%
BDGD04:	42.8395%
BDGD25:	2.2318%
BDGD28:	3.0934%
BDGD31:	5.5441%
BDGD33:	3.1715%
BDGD35:	1.2504%
BDGD56:	0.9834%
BDGD58:	3.9395%
C03:	3.1021%
C26:	4.8929%
C28:	2.0318%
SE022:	6.1557%
SE094:	17.0994%

Tabela 4.17 – Variáveis do modelo FEC e seus índices de Gini

Mais uma vez, comparando os modelos de DEC e FEC, percebemos a tendência entre a simplificação das variáveis dar mais explicação para o modelo DEC do que o FEC.

4.3 Resultado SFS Regressor

Por fim, foi testada a sensibilidade do método *Sequential Feature Selector* (SFS) ou *Stepwise*. Para esse teste foram utilizadas as variáveis de tolerância.

SFS DEC		
tol	Linear Regression	n features
0.01	$R^2 = 0.468$ MAE = 6.296 MSE = 100.089 Tempo = 12.45s	6
0.001	$R^2 = 0.521$ MAE = 5.925 MSE = 90.185 Tempo = 73.17s	33
0.0001	$R^2 = 0.546$ MAE = 6.012 MSE = 85.421 Tempo = 234.36s	82

Tabela 4.18 – Tabela comparativa do modelo SFS para o indicador DEC

SFS FEC		
tol	Linear Regression	n features
0.01	$R^2 = 0.325$ MAE = 2.516 MSE = 18.579 Tempo = 12.49s	6
0.001	$R^2 = 0.491$ MAE = 2.290 MSE = 14.001 Tempo = 94.71s	40
0.0001	$R^2 = 0.512$ MAE = 2.242 MSE = 13.439 Tempo = 165.68s	64

Tabela 4.19 – Tabela comparativa do modelo SFS para o indicador FEC

Novamente, a constatação feita sobre os modelos de DEC e FEC foi visualizada. Para uma tolerância de 0.01, temos a seleção de 6 variáveis, onde o modelo DEC consegue obter um R^2 igual a 0.468 e o modelo FEC um R^2 igual a 0.325. Vale ressaltar que a vantagem desse modelo é poder simplificar o estudo e explicar variáveis dependentes com poucas variáveis independentes.

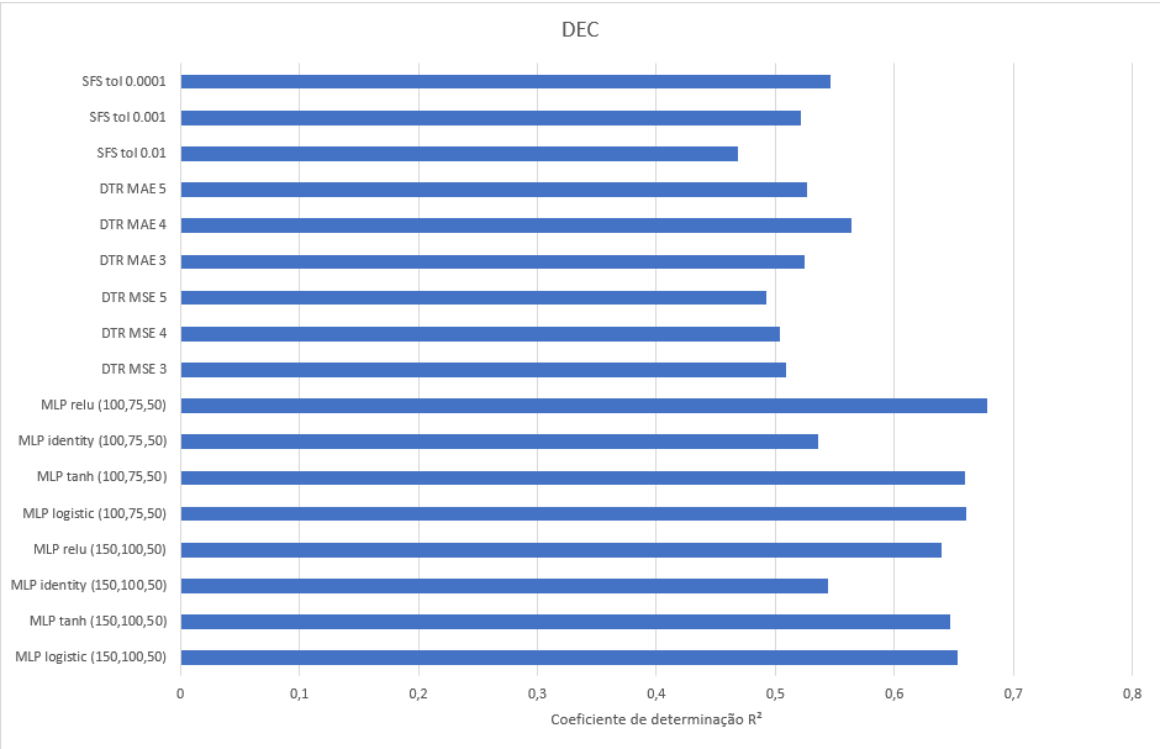


Figura 4.23 – Modelo DEC

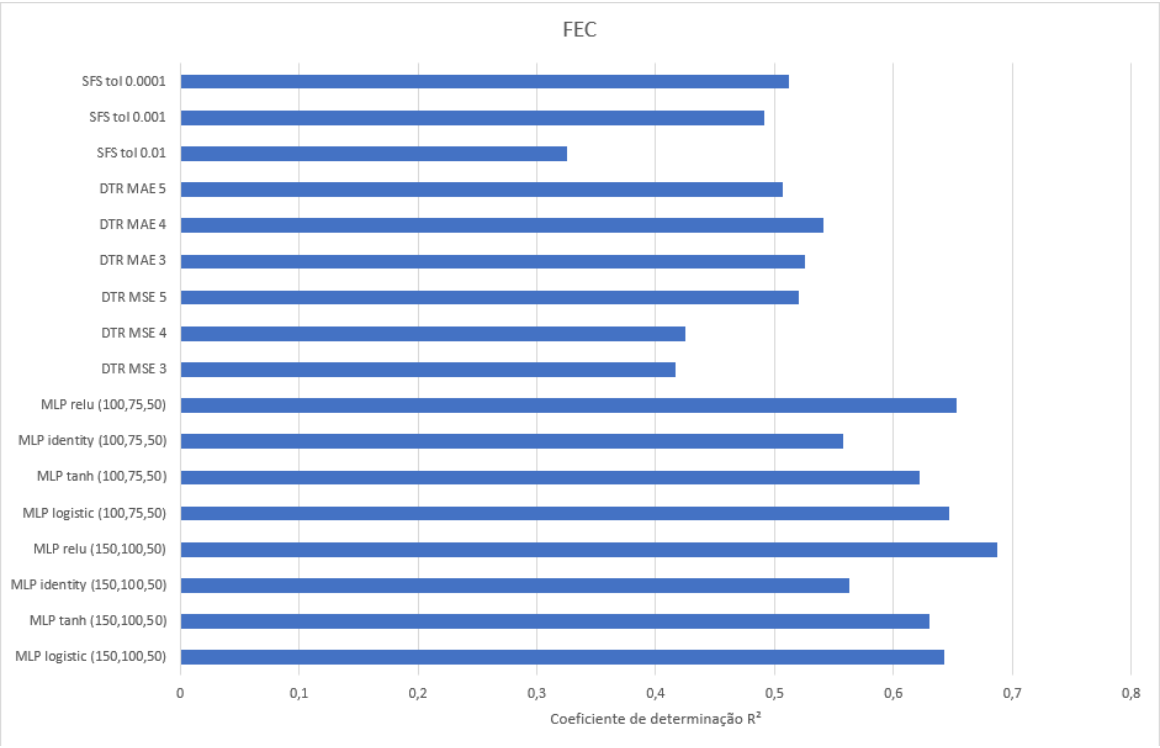


Figura 4.24 – Modelo FEC

Nota-se que os modelos com melhor desempenho em respeito a variável R^2 são os MLP com ativação relu, tanto para o modelo DEC, quanto para o FEC.

4.4 Teste dos modelos com Análise de Componentes Principais

A redução do número de variáveis em um conjunto de dados naturalmente implica em perda de precisão, mas a estratégia na redução de dimensionalidade é tentar trocar um pouco de precisão por simplicidade.

Nesse teste final, foram testados o melhor modelo de cada método. A ideia é transformar as 278 variáveis independentes em 80, 40, 20, 10 variáveis combinadas, respectivamente.

Teste com PCA DEC			
n components	MLP	Decision Tree 'absolute error' max depth=4	SFS tol = 0.0001
80	(40,20,10) R^2 = -1.370 MAE = 11.919 MSE = 446.123 Tempo = 438.22s	R^2 = 0.339 MAE = 6.312 MSE = 124.363 Tempo = 1.89s	R^2 = 0.231 MAE = 8.438 MSE = 144.765 Tempo = 24.21s
40	(20,10,5) R^2 = -2.068 MAE = 12.311 MSE = 577.550 Tempo = 457.99s	R^2 = 0.232 MAE = 6.289 MSE = 144.602 Tempo = 0.95s	R^2 = 0.321 MAE = 7.698 MSE = 127.702 Tempo = 5.47s
20	(10,5,2) R^2 = -0.382 MAE = 9.347 MSE = 260.158 Tempo = 330.93s	R^2 = 0.357 MAE = 6.020 MSE = 120.954 Tempo = 0.40s	R^2 = 0.377 MAE = 7.302 MSE = 117.335 Tempo = 1.42s
10	(5,2) R^2 = 0.441 MAE = 6.262 MSE = 105.118 Tempo = 150.16s	R^2 = 0.406 MAE = 5.876 MSE = 111.870 Tempo = 0.21s	R^2 = 0.434 MAE = 6.665 MSE = 106.507 Tempo = 0.36s

Tabela 4.20 – Tabela comparativa dos modelos com PCA para o indicador DEC

Teste com PCA FEC			
n components	MLP	Decision Tree 'absolute error' max depth=4	SFS tol = 0.0001
80	(40,20,10) R^2 = -0.408 MAE = 3.976 MSE = 38.767 Tempo = 153.48s	R^2 = 0.381 MAE = 2.507 MSE = 17.028 Tempo = 1.51s	R^2 = 0.246 MAE = 3.170 MSE = 20.742 Tempo = 27.01s
40	(20,10,5) R^2 = -0.387 MAE = 3.917 MSE = 38.191 Tempo = 152.27s	R^2 = 0.399 MAE = 2.498 MSE = 16.547 Tempo = 0.79s	R^2 = 0.349 MAE = 2.929 MSE = 17.926 Tempo = 6.04s
20	(10,5,2) R^2 = -0.038 MAE = 3.286 MSE = 28.585 Tempo = 164.85s	R^2 = 0.369 MAE = 2.517 MSE = 17.377 Tempo = 0.40s	R^2 = 0.415 MAE = 2.708 MSE = 16.090 Tempo = 1.55s
10	(5,2) R^2 = 0.444 MAE = 2.531 MSE = 15.293 Tempo = 32.08s	R^2 = 0.382 MAE = 2.452 MSE = 17.000 Tempo = 0.217s	R^2 = 0.439 MAE = 2.577 MSE = 15.445 Tempo = 0.37s

Tabela 4.21 – Tabela comparativa dos modelos com PCA para o indicador FEC

Esse resultado mostrou que, para os modelos de DEC e FEC, a redução de dimensionalidade do PCA retirou informações importantes para os dois modelos. Além disso, percebemos que ao aumentarmos o número de principais componentes, pioramos o desempenho dos três métodos.

5 Conclusões

No contexto específico das distribuidoras de energia elétrica, a metodologia de definição de limites para indicadores de continuidade DEC e FEC é essencial para definir comparações entre os conjuntos das distribuidoras de energia elétrica. No passado, técnicas estatísticas foram aplicadas para poder tentar parametrizar e definir quais limites dos indicadores eram justos de serem aplicados.

Esse estudo foi capaz de comparar três métodos: *Stepwise Regression*, *Decision Tree Regression* e *Multi-Layer Perceptron Regression*. Dos três métodos com diversos modelos, o *MLP Regressor* com ativação relu foi o método que mais satisfez as métricas propostas no estudo. Mesmo sendo o melhor preditor, um R^2 igual a 0.678 para o DEC e 0.687 para o FEC ainda não são suficientes para que os limites da metodologia sejam previstos por esse modelo, sem passar pela etapa seguinte do "Método Dinâmico".

Comparando os modelos com as métricas propostas (MAE, MSE, R^2), em geral, a *MLP Regression* foi o melhor modelo, seguido das árvores de decisão e por último a *Stepwise Regression*. Entretanto, a *Decision Tree Regression* e a *Stepwise Regression* possuem uma vantagem em relação a *MLP Regression*, que é a de entregar as "importâncias das características", o que pode ser um diferencial para modelos que necessitam delas. Ao fim, olhando um contexto geral, o algoritmo de árvores de decisão tem menor custo computacional, maior facilidade de entendimento, bom coeficiente de determinação e ainda consegue selecionar variáveis.

É importante reconhecer as limitações inerentes ao estudo, como, escassez de dados relativos a gestão das distribuidoras. Acredita-se que o modelo apresentado ainda pode ser melhorado a partir da inclusão de novos dados, como: Tempo Médio de Preparação (TMP), Tempo Médio de Deslocamento (TMD), Tempo Médio de Execução (TME), Tempo Médio de Atendimento a Emergências (TMAE) e Percentual do Número de Ocorrências Emergenciais com Interrupção de Energia Elétrica (PNIE). Além desses dados, recomenda-se o teste dos atributos utilizando extensão de rede MT, o que não foi feito nesse trabalho. Portanto, propõe-se como agenda futura a extração desses dados em específico com intuito de trazer maior desempenho para o modelo proposto.

A conclusão desse trabalho representa um ponto de partida para futuros debates acadêmicos. Espera-se que contribua para o avanço do conhecimento na aplicação de métodos de aprendizado de máquina, principalmente no que diz respeito a Metodologia de Definição dos Limites dos indicadores de continuidade.

Referências

[1] Muhammad Waseem Ahmad, Monjur Mourshed e Yacine Rezgui. “Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption”. Em: *Energy and Buildings* 147 (2017), pp. 77–89. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2017.04.038. URL: https://www.sciencedirect.com/science/article/pii/S0378778816313937.

[2] A.S. Ahmad et. al. “A review on applications of ANN and SVM for building electrical energy consumption forecasting”. Em: *Renewable and Sustainable Energy Reviews* 33 (2014), pp. 102–109. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2014.01.069. URL: https://www.sciencedirect.com/science/article/pii/S1364032114000914.

[3] Abraham L. SICSU; et. al. *Técnicas de machine learning*. 1ª ed. Editora Blucher., 2023. ISBN: 9786555063974. URL: https://integrada.minhabiblioteca.com.br/#/books/9786555063974/..

[4] ANA. *Dados Abertos*. Disponível em: https://dadosabertos.ana.gov.br/explore?collection=Dataset. (Acesso em 13/12/2023).

[5] ANEEL. *Base de Dados Geográfica das Distribuidoras*. Disponível em: https://dadosabertos-aneel.opendata.arcgis.com/explore?tags=dist. (Acesso em 13/12/2023).

[6] ANEEL. *Nota Técnica n° 0102/2014-SRD/ANEEL*. Disponível em: https://antigo.aneel.gov.br/web/guest/audiencias-publicas-antigas?p_auth=LeU9r9qY&p_p_id=participacaopublica_WAR_participacaopublicaportlet&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&p_p_col_id=column-2&p_p_col_pos=1&p_p_col_count=2&_participacaopublica_WAR_participacaopublicaportlet_ideParticipacaoPublica=898&_participacaopublica_WAR_participacaopublicaportlet_javax.portlet.action=visualizarParticipacaoPublica. 2014. (Acesso em 13/12/2023).

[7] ANEEL. *Nota Técnica n° 0136/2021-SRD/ANEEL*. Disponível em: https://antigo.aneel.gov.br/web/guest/tomadas-de-subsidios?p_auth=gtjFLK9G&p_p_id=participacaopublica_WAR_participacaopublicaportlet&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&p_p_col_id=column-2&p_p_col_count=1&_participacaopublica_WAR_participacaopublicaportlet_ideParticipacaoPublica=3636&_participacaopublica_WAR_participacaopublicaportlet_javax.portlet.action=visualizarParticipacaoPublica. 2021. (Acesso em 09/01/2023).

[8] Aurélien Geron. *handson-ml3*. Upload de Aurélien Geron. Disponível em: https://github.com/ageron/handson-ml3. 2023. (Acesso em 13/12/2023).

[9] FUNAI. *Geoserver*. Disponível em: https://geoserver.funai.gov.br/geoserver/web/. (Acesso em 13/12/2023).

[10] Aurélien Geron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 1ª ed. O'Reilly Media, Inc., abr. de 2017. ISBN: 1491962291.

[11] Pedro A. González e Jesús M. Zamarreño. "Prediction of hourly energy consumption in buildings based on a feedback artificial neural network". Em: *Energy and Buildings* 37.6 (2005), pp. 595–601. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2004.09.006. URL: https://www.sciencedirect.com/science/article/pii/S0378778804003032.

[12] INCRA. *Base de Dados INCRA*. Disponível em: https://certificacao.incra.gov.br/csv_shp/export_shp.py. (Acesso em 13/12/2023).

[13] INMET. *Dados Históricos*. Disponível em: https://portal.inmet.gov.br/dadoshistoricos. (Acesso em 13/12/2023).

[14] S.A Kalogirou. "Applications of artificial neural networks in energy systems". Em: *Energy Conversion and Management* 40.10 (1999), pp. 1073–1087. ISSN: 0196-8904. DOI: https://doi.org/10.1016/S0196-8904(99)00012-6. URL: https://www.sciencedirect.com/science/article/pii/S0196890499000126.

[15] Soteris A. Kalogirou e Milorad Bojic. "Artificial neural networks for the prediction of the energy consumption of a passive solar building". Em: *Energy* 25.5 (2000), pp. 479–491. ISSN: 0360-5442. DOI: https://doi.org/10.1016/S0360-5442(99)00086-9. URL: https://www.sciencedirect.com/science/article/pii/S0360544299000869.

[16] Isaías. LIMA. *Inteligência Artificial*. 1ª ed. Grupo GEN., 2014. ISBN: 9788595152724. URL: https://integrada.minhabiblioteca.com.br/#/books/9788595152724/..

[17] Osmar. MARTINS Gilberto de A.; DOMINGUES. *Estatística Geral e Aplicada, 6ª edição*. 6ª ed. Grupo GEN., 2017. ISBN: 9788597012682. URL: https://integrada.minhabiblioteca.com.br/#/books/9788597012682/..

[18] MMA. *Dados Geográficos*. Disponível em: http://mapas.mma.gov.br/i3geo/datadownload.htm. (Acesso em 13/12/2023).

[19] Ruengvirayudh P. e Brooks G. P. "Comparing stepwise regression models to the best-subsets models, or, the art of stepwise." Em: *General Linear Model Journal* 42(1) (2016), pp. 1–14.

[20] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. Em: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[21] Peter. RUSSELL Stuart J.; NORVIG. *Inteligência Artificial: Uma Abordagem Moderna*. 4ª ed. Grupo GEN., 2022. ISBN: 9788595159495. URL: https://integrada.minhabiblioteca.com.br/#/books/9788595159495/..

[22] Geoffrey K.F. Tso e Kelvin K.W. Yau. “Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks”. Em: *Energy* 32.9 (2007), pp. 1761–1768. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2006.11.010. URL: https://www.sciencedirect.com/science/article/pii/S0360544206003288.

[23] Alexandre C. Xavier, Carey W. King e Bridget R. Scanlon. “Daily gridded meteorological variables in Brazil (1980–2013)”. Em: *International Journal of Climatology* 36.6 (2016), pp. 2644–2659. DOI: https://doi.org/10.1002/joc.4518. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.4518.

[24] Jian Yang et al. “Feedback System Control Optimized Electrospinning for Fabrication of an Excellent Superhydrophobic Surface”. Em: *Nanomaterials* 7 (out. de 2017), p. 319. DOI: 10.3390/nano7100319.