Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Cláudio Siervi Mota Junior

Avaliação de Modelos Polinomiais para Representar a Curva de Eficiência de Turbinas Hidro-Geradoras

Curitiba
2020

Cláudio Siervi Mota Junior

Avaliação de Modelos Polinomiais para Representar a Curva de Eficiência de Turbinas Hidro-Geradoras

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Dr. Abel Soares Siqueira

Curitiba
2020

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Avaliação de Modelos Polinomiais para Representar a Curva de Eficiência de Turbinas Hidro-Geradoras

Cláudio Siervi Mota Junior[1]
Prof. Dr. Abel Soares Siqueira[2]

Resumo

Este artigo mostra os resultados de uma pesquisa exploratória que buscou determinar uma forma geral para a função de rendimento de turbinas hidrogeradoras a partir do ajuste de modelos polinomiais à quatro amostras de curva colina. Ao todo foram avaliados dez modelos polinomiais, sendo a seleção de modelos feita pelo algoritmo *Best Subsets* com as métricas de avaliação: R^2 ajustado, PRESS e AIC. A partir deste algoritmo, selecionou-se uma média de quatro modelos com melhor ajuste para cada amostra. Em seguida, foi realizada uma análise dos resíduos destes modelos, por último, selecionou-se o modelo com resíduos válidos e menor número de parâmetros. A pesar dos bons ajustes, concluí-se que os modelos não podem ser generalizados devido à forma de coleta de dados e ao número de amostras. No entanto, a pesquisa mostra que é possível encontrar modelos polinomiais com bons ajustes e com um pequeno número de variáveis para representar as curvas de rendimento. Ao final, sugere-se então expandir o estudo para um conjunto maior de turbinas e usando outros métodos de coletas de dados.
Palavras-chave: rendimento de turbinas, best subsets, usinas hidrelétricas, curva colina, regressão múltipla.

Abstract

This article shows the results of an exploratory research that sought to determine a general form to represent the efficience curve of hydro-generating turbines from the fit of polynomial models to four samples of hill curves. In all, ten polynomial models were evaluated, with the selection of models made by the Best Subsets algorithm with the evaluation metrics: adjusted R2, PRESS and AIC. From this algorithm, an average of four models was selected with the best fit for each sample. Then, an analysis of the residues of these models was carried out. Finally, the model with valid residues and the lowest number of parameters was selected. Despite the good adjustments, it was concluded that the models cannot be generalized due to the form of data collection and the number of samples. However, the research shows that it is possible to find polynomial models with good fits and a small number of variables to represent the efficience curves. In the end, it is then suggested to expand the study to a larger set of turbines and using other data collection methods.
***Keywrods**: hydraulic turbines efficiency, best subsets, hydropower plants, hill curve, multiple regression.*

1 Introdução

A geração de energia em uma hidrelétrica é um processo sintêmico que consiste em transformar a energia potencial da água no reservatório em energia cinética da água nos condutos forçados, que por sua vez vira energia mecânica no eixo da turbina e por final é transformada em energia elétrica no gerador. A função que modela a produção de energia do conjunto turbina-gerador de uma hidrelétrica é dada pela Equação 1. Desta equação, tem-se que, a potência ativa p em uma unidade geradora depende diretamente da altura de queda do reservatório h e da vazão de engolimento da turbina q, onde, η_t e η_g são fatores de eficiência da turbina e do gerador que limitam a energia produzida pelo conjunto e, g e ρ são constantes que representam a força gravitacional e a massa específica da água do reservatório.

$$p(h,q) = \rho.g.\eta_t.\eta_g.h.q. \tag{1}$$

Deste modo, para se obter a potência ótima de um conjunto turbina-gerador é preciso que tanto a turbina quanto o gerador operem com máxima eficiência no respectivo cenário operativo, pois o rendimento da turbina e do gerador limitam a capacidade de geração de energia do conjunto de máquinas.

A determinação da eficiência máxima da turbina começa pela escolha do modelo de turbina a ser instalada na usina. Como exemplificado na Figura 1, existem diferentes modelos (tipos) de turbinas hidráulicas disponíveis no mercado, de modo que o modelo de turbina escolhido para uma hidrelétrica é determinado por uma série de estudos de viabilidade técnico-econômica contratados pelo agente detentor da outorga da usina [1], depois então esses estudos de viabilidade são avaliados pelos órgãos ambientais e pela ANEEL.

As perdas de rendimento de uma turbina estão associadas às perdas de geração verificadas em diferentes

[1]Aluno do programa de Especialização em Data Science & Big Data, claudio.siervi@gmail.com.
[2]Professor do Departamento de Matemática - DMAT/UFPR.

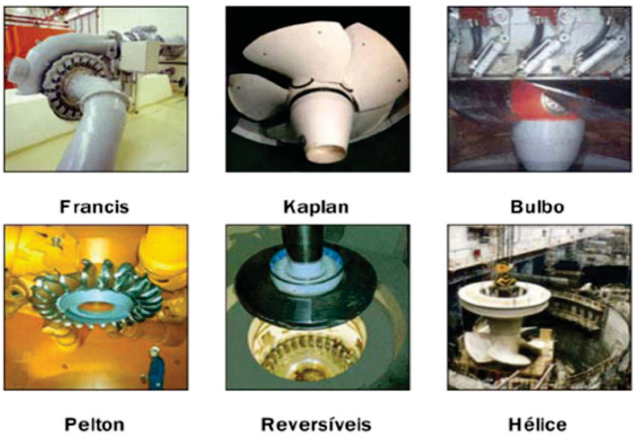

Figura 1: Modelos de turbinas hidráulicas mais comuns. Figura retirada de [2].

condições de operação, onde a potência disponível no eixo da turbina é dissipada em perdas internas e externas na própria turbina[3]. Para minimizar essas perdas, cada turbina da usina é projetada para atender a certos valores prefixados (cenários) de forma a retornar o rendimento máximo em cada cenário.

Para isto, são realizados ensaios de laboratório ou de campo e estudos da similaridade geométrica e hidrodinâmica que auxiliam na determinação do comportamento da turbina quando se variam as grandezas: queda, vazão, abertura do distribuidor e rotações por minuto. Destes estudos são então construídas as curvas de colina das turbinas. De modo que, uma curva colina é um diagrama que indica a máxima eficiência da turbina em relação à variação de cada grandeza [2]. Os limites operativos da turbina restringem a faixa operativa da curva colina, na Figura 2 tem-se que as linhas vermelhas limitam a altura de queda enquanto que a vazão é limitada pelas bordas superior e inferior do gráfico.

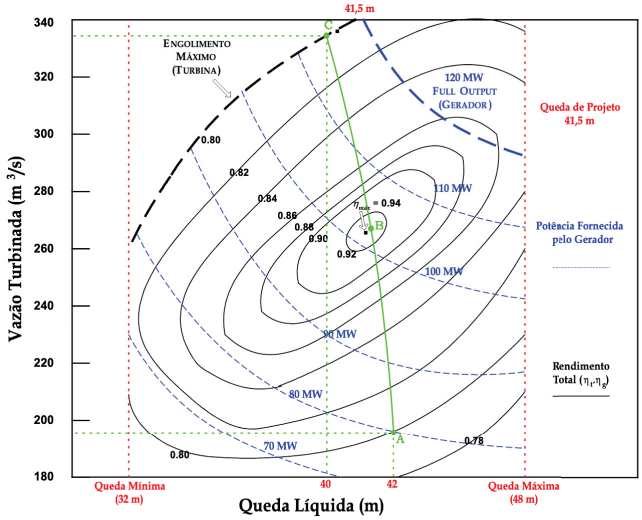

Figura 2: Curva colina ilustrada. Figura retirada de [2].

Em geral, a curva de colina não é usada para determinar o rendimento da turbina na função de produção de energia. Mas sim, utilizam-se fatores de eficiência média η_t ou modelos polinomiais definidos em função da vazão e da queda líquida [4] [5]. Há estudos que calculam o rendimento da turbina em função de variáveis como potência, nível de montante e nível de jusante [6]. E outros estudos que calculam o rendimento em função apenas da potências de entrada e da potência de saída da turbina, descontando as perdas [3].

Nos principais modelos de simulação usados nos estudos de planejamento energético do setor elétrico brasileiro (NEWAVE, DECOMP e SUISHI) a função de produção das hidrelétricas tem representação simplificada de diferentes maneiras, dentre as quais destaca-se a utilização de um parâmetro fixo para representar a função de rendimento das turbinas. Este parâmetro, denominado rendimento médio turbina-gerador, é calculado pela Empresa de Pesquisa Energética (EPE) com base na metodologia definida na Nota Técnica EPE-DEE-RE-037/2011-r2 [5]. Nesta metodologia a EPE propõe um polinômio de segundo grau para representar a função de rendimento utilizada no processo de cálculo do rendimento médio.

Em estudos privados que utilizaram amostras de curva colina de diferentes usinas para calcular o rendimento médio turbina - gerador conforme a metodologia proposta pela EPE, observou-se que o modelo proposto pela EPE para representar a curva de rendimento das turbinas hidráulicas de usinas hidrelétricas conectadas ao Sistema Interligado Nacional (SIN) não retornava bons ajustes na maioria dos estudos. O modelo retornava baixos erros de ajuste, mas com acentuada correlação cíclica dos resíduos com a vazão de engolimento.

Desta forma, este estudo exploratório investiga se há uma forma geral para representar a curva de rendimento de turbinas hidráulicas de modo que o modelo retorne bons ajustes em todas as amostras analisadas.

2 Materiais e Métodos

Nesta seção estão descritas as etapas da pesquisa.

2.1 Modelos avaliados

Este estudo avalia diferentes modelos polinomiais para representar as curvas de eficiência de quatro turbinas hidrogeradoras. Para isto, seguiu-se a abordagem adotada por Diniz et al [4] no artigo *A Mathematical Model for the Efficiency Curves of Hydroeletric units*, onde a curva de rendimento (η_t) é representada por modelos polinomiais que explicam o rendimento das turbinas pelas variáveis altura de queda (h) e vazão de engolimento (q).

O Modelo 1 é o polinômio incompleto de quarta ordem (Equação 2) proposto por Diniz et al. [4] com base no experimento conduzido com uma única curva colina.

$$\eta_t(h,q) = \beta_0 + \beta_1.h + \beta_2.q + \beta_3.h^2 + \beta_4.h.q + \beta_5.q^2 + \beta_6.h^2.q + \beta_7.h.q^2 + \beta_8.h^2.q^2. \quad (2)$$

O Modelo 2 é polinômio de segunda ordem (Equação 3) empregado pela EPE no cálculo dos fatores de rendimento médio turbina-gerador e perda hidráulica média das usinas hidrelétricas conectadas ao SIN [5]. Não foram encontrados estudos que embasassem a escolha deste polinômio de segunda ordem pela EPE, apenas uma referencia ao artigo científico de Diniz et al [4].

$$\eta_t(h,q) = \beta_0 + \beta_1.h + \beta_2.q + \beta_3.h^2 + \beta_4.h.q + \beta_5.q^2. \quad (3)$$

Neste contexto, buscou-se representar a curva de rendimento das turbinas hidrogeradoras a partir de funções polinomiais de ordem igual e superior a dois. Inicialmente foram avaliados polinômios de ordem até vinte, mas se observando *overfitting* em modelos de ordem inferior a dez, decidiu-se por limitar o espaço de busca. Disto, avaliam-se neste artigo modelos polinomiais de ordem até dez, incluindo os dois modelos de referência descritos acima. Deste modo, tem-se que:

- Modelo 1 –> ordem 4 incompleto (Equação 2)
- Modelo 2 –> ordem 2 (Equação 3)
- Modelo 3 –> ordem 3
- Modelo 4 –> ordem 4
- Modelo 5 –> ordem 5
- Modelo 6 –> ordem 6
- Modelo 7 –> ordem 7
- Modelo 8 –> ordem 8
- Modelo 9 –> ordem 9
- Modelo 10 –> ordem 10.

2.2 Coleta de Dados

Devido à escassez de dados disponíveis sobre curvas de colina, as amostras utilizadas nos experimentos deste estudo foram extraídas de diferentes fontes de dados, de forma que cada amostra corresponde à uma turbina distinta de hidrelétricas diferentes.

A primeira amostra analisada (Turbina A) foi retirada do artigo de Diniz et al [4], a qual é coletada discretizando os valores de queda líquida e vazão de forma uniforme na curva colina, de modo que os fatores de rendimento η_t são obtidos por interpolação linear da grade de valores de queda e vazão.

As demais amostras analisadas (Turbinas B, C e D) são de hidrelétricas reais operam no SIN, as quais foram cedidas pelos agentes de geração detentores da outorga destas usinas. Estas amostras foram coletadas pelos agentes de geração com o intuito de calcular o rendimento médio turbina-gerador dos modelos de simulação utilizados nos estudos de planejamento da operação. De acordo com Nota Técnica EPE-DEE-RE-037/2011-r2 da EPE [5], o processo de coleta destas amostras pressupõe que: sejam escolhidos valores de queda líquida compreendendo da mínima à máxima queda líquida permitida para operação da turbina. Para cada valor de queda escolhido devem então serem selecionados valores de vazão que compreendem da mínima à máxima vazão turbinada permitida para a operação da turbina na correspondente queda.

Ou seja, apenas a amostra da Turbina A foi obtida a partir de um experimento planejado para determinar uma forma geral para a função de rendimento de turbinas hidrogeradoras. As demais amostras das Turbinas B, C e D foram obtidas a partir de experimentos planejados para o cálculo do rendimento médio turbina-gerador dessas usinas.

A Figura 3 mostra a dispersão dos valores de rendimento em função da queda líquida e da vazão de engolimento da turbina, indicando que realmente existe uma relação não linear entre essas variáveis.

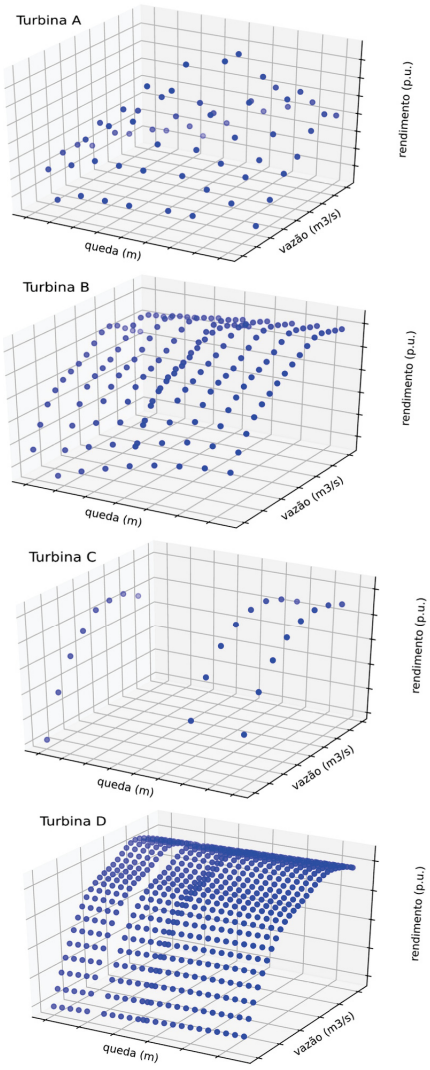

Figura 3: Eficiência em função da queda e do engolimento.

2.3 Análise de Regressão

A análise de regressão é uma técnica estatística utilizada para determinar o relacionamento entre uma única variável dependente (predita) e uma ou mais variáveis independentes (variáveis exploratórias)[7]. Na análise de regressão linear a variável predita é expressa como

uma combinação linear das variáveis exploratórias enquanto que na análise de regressão não linear a varável predita é expressa como uma combinação não linear das variáveis exploratórias. Uma análise de regressão linear é dita simples se é realizada observando apenas uma variável exploratória e é denominada múltipla (ou geral) quando expressa a variável predita por duas ou mais variáveis exploratórias (além do termo constante) [8]. A regressão linear múltipla é uma extensão natural da regressão linear simples e é especificada como:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i \qquad (4)$$

Onde, β_i são os coeficientes; X_{ij} são as variáveis exploratória; ϵ_i é o termo de erro; com $i = \{0, \ldots, n\}$.

Um caso particular dos modelos de regressão linear múltipla é a análise de regressão polinomial. Estes modelos contem termos quadráticos e de ordem superior das variáveis explicativas [9]. O seguinte modelo representa uma regressão polinomial de ordem N e com k variáveis exploratórias:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + ... + \beta_k X_{ki}^N + \epsilon_i \qquad (5)$$

Para ajustar modelos de regressão linear múltipla (da Equação 5) pelo método dos mínimos quadrados é preciso então verificar as seguintes proposições :

1. As variáveis explanatórias tem valores fixos (não estocásticas).
2. Não existe relação linear exata entre duas ou mais variáveis explanatórias.
3. O erro tem esperança matemática zero e variância constante para todas as observações.
4. O termo de erro tem distribuição normal.

2.4 Algoritmos *Best Subsets*

Os algoritmos *"Best"Subsets* determinam subconjuntos de modelos com melhor ajuste de acordo com uma ou mais métricas e sem realizar seleção de variáveis. Assim, esses algoritmos requerem o cálculo de apenas uma fração de todos os modelos de regressão possíveis.

Kutner et al [9] explicam que esses algoritmos não apenas fornecem os melhores subconjuntos de acordo com o critério especificado, mas também identificam subconjuntos "bons"que podem ser refinados a partir de uma seleção final de variáveis para determinar um subconjunto de variáveis exploratórias a serem empregadas no modelo de regressão.

A seleção de variáveis dos modelos com melhores ajustes pode feita por exemplo com métodos como *forward stepwise selection* e *backward stewise selection* [10].

A principal diferença entre os métodos *stepwise* e o algoritmo *"best"subsets* é que o *stepwise* termina seu procedimento de busca retornando um único "melhor"modelo final. Em contra partida, o algoritmo *"best"subsets* retorna vários modelos considerados "melhores"por diferentes métricas. Isto torna o algoritmo *"best"subsets* a melhor opção para o experimento deste estudo, pois a identificação de um único modelo de regressão pode ocultar o fato de que vários outros modelos de regressão também podem ser "bons". Finalmente, a "qualidade"de um modelo de regressão só pode ser estabelecida por um exame completo usando uma variedade de diagnósticos [9].

No entanto, como este trabalho busca identificar uma função geral para representar a curva de rendimento das turbinas, optou-se por selecionar o modelo com o melhor ajuste, no subconjunto de melhores modelos, por eliminação dos modelos que não satisfazem os pressupostos do método dos mínimos quadrados e, então, dentre estes, selecionar o modelo com o menor número de variáveis explicativas.

► Critérios de Seleção

A seleção do subconjunto de "melhores"modelos é feita com base nos critérios: R^2 ajustado, PRESS e AIC.

O coeficiente de determinação ajustado, ou R^2 ajustado, é uma estatística que indica o quanto o modelo em analise explica a variância da amostra. Este coeficiente considera o número de variáveis exploratórias do modelo de regressão por meio dos graus de liberdade e penaliza os modelos com grande número de variáveis exploratórias [8, 9]. De modo que, quando a variância explicada pelo modelo é máxima, adicionar novas variáveis explicativas diminui o valor deste coeficiente.

A estatística PRESS (*Predicted Residual Error Sum of Squares*) é uma técnica de validação cruzada de modelos de regressão linear que avalia a capacidade preditiva do modelo e também pode ser usada para comparar diferentes modelos de regressão. Esta técnica divide uma amostra de tamanho n em dois subconjuntos usados no treino e na validação dos modelos, onde o conjunto de validação tem apenas 1 observação e o conjunto de treino tem n−1 observações. Para cada um dos n conjuntos de validação calcula-se então o resíduo entre a variável preditiva e a estimativa da variável preditiva no conjunto de validação $(y_i - \hat{y}_{(i)})$ e, ao final, calcula-se a soma dos resíduos quadrados [9, 11].

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2.$$

A estatística AIC (*Akaike's Information Criterion*) compara diferentes modelos de regressão combinando informações sobre o a soma dos resíduos quadrados, número de parâmetros no modelo e o tamanho da amostra. Um valor baixo de AIC em comparação com aos valores de outros modelos possíveis indica um bom ajuste [9, 11].

► Validação dos Modelos

Os modelos selecionados como melhor ajuste em cada amostra são então validados por uma análise de resíduos, a qual buscou verificar se as estimativas dos modelos são não tendenciosas e satisfazem os pressupostos do

Como os dados utilizados não são séries temporais, neste estudo não se verificará se há ou não correlação serial entre os erros de observações diferentes.

	Turbinas	min	25%	50%	75%	max
A	queda	32	36	40	44	48
	vazão	180	220	260	300	340
	rend.	0.77	0.81	0.83	0.85	0.91
B	queda	50	58	67	74	82
	vazão	250	325	400	475	550
	rend.	0.69	0.86	0.90	0.93	0.96
C	queda	14	14	17	18	18
	vazão	30	47	65	84	108
	rend.	0.84	0.90	0.92	0.93	0.94
D	queda	58	61	65	69	72
	vazão	40	61	84	100	120
	rend.	0.65	0.86	0.92	0.93	0.95

Tabela 1: Frequência de ocorrência das variáveis queda líquida, vazão e rendimento das Turbinas A, B, C e D.

método dos mínimos quadrados: resíduos com média zero, variância constante e distribuição normal.

E, ao final, seleciona - se o modelo mais parcimonioso em cada amostra, que é o modelo com menor número parâmetros dentre os modelos que cumprem os pressupostos dos mínimos quadrados.

2.5 Softwares

Os experimentos deste estudo foram implementados na linguagem Python utilizando os pacotes: StatsModels, Sklearn, Pandas, ScyPy, Matlibplot e Seaborn.

3 Resultados

Uma vez obtidas as amostras das curvas de colina das Turbinas A, B, C e D (seção 3) e definido um conjunto de modelos a serem analisados (seção 2.1), selecionou-se então um subconjunto de "melhores"modelos de acordo com os critérios especificados na seção 2.4. Em seguida, todos os modelos de cada subconjunto foram avaliados de acordo com os pressupostos do método dos mínimos quadrados (seção 2.3).

▶ **Análise das Amostras**

A Tabela 1 mostra a frequência de ocorrência dos valores queda, vazão e rendimento das amostras. Estas amostras são de hidrelétricas distintas e têm tamanhos diferentes com respectivas 63, 160, 24 e 466 observações.

Nos gráficos de caixa da Figura 4, tem-se que a amostra da Turbina A tem distribuição simétrica nas três variáveis enquanto que as amostras das Turbinas B, C e D mostram assimetrias em relação à variável rendimento. De modo que, essas assimetrias negativas indicam uma concentração de valores mais altos de rendimento nessas amostras. Isto se confirma pelos valores da Tabela 1, onde cada coluna representa um nível do gráfico de caixa, pois os rendimentos da amostra A se concentram em valores entre 0.8 (p.u.) e 0.9 (p.u.) enquanto que nas outras amostras (B, C e D) os rendimentos se concentram em valores acima de 0.9 (p.u.).

Ainda em relação à Figura 4, observa-se que a Turbina B apresenta uma alta queda e uma grande vazão de engolimento indicando que esta turbina está instalada em uma hidrelétrica de grande porte com um grande reservatório. Observa-se também que a Turbina C apresenta uma baixa altura de queda e uma pequena vazão, assim indicando que esta turbina está instalada em uma hidrelétrica pequena e com reservatório sem capacidade de armazenamento (fio d'água).

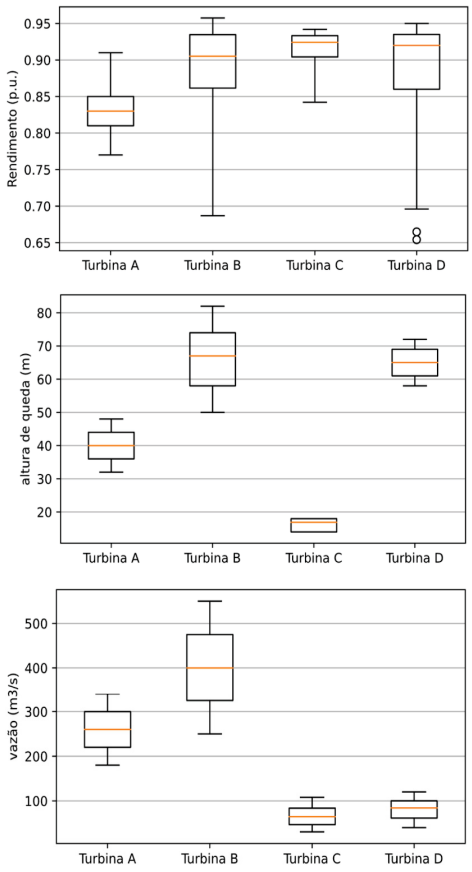

Figura 4: Gráfico de caixa das variáveis das amostras das Turbinas A, B, C e D.

A Figura 5 mostra a dispersão e a distribuição das variaríeis amostrais das quatro turbinas.

Esta figura indica que a amostra da Turbina A foi coletada de forma distinta das amostras das Turbinas B, C e D. Pois na Turbina A o rendimento não tem uma relação direta com a vazão enquanto que nas amostras B, C e D se observa uma forte relação não linear entre essas variáveis. Além disso, a dispersão das variáveis queda versus vazão mostram algum tipo de censura nos dados da Turbina A enquanto que nas Turbinas B, C e D mostram um padrão diferente. A censura dos dados da Turbina A indica que a coleta desta amostra foi restrita aos limites operativos da turbina, enquanto que as amostras das Turbinas B, C e D indicam que estas amostras foram coletadas em todo o domínio das curvas de colina, ou seja, dentro e fora dos limites operativos

da turbina.

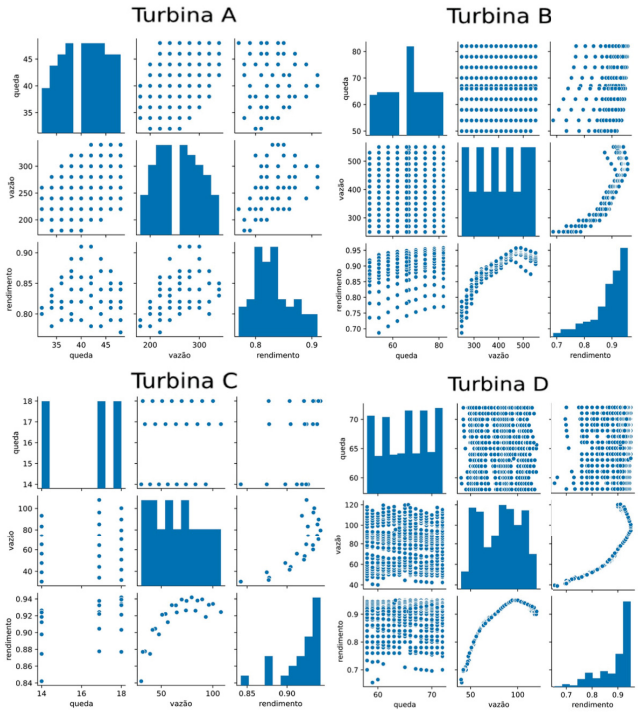

Figura 5: Dispersão e histogramas das amostras das Turbinas A, B, C e D.

► Seleção de modelos

Nesta etapa foi realizada a seleção dos modelos com melhores ajustes de acordo com cada métrica.

A Figura 6 mostra o cálculo das três métricas propostas para selecionar os subconjuntos de "melhores"modelos nas amostras. De modo que, na primeira coluna de gráficos tem-se as estimativas de R^2 ajustado, na segunda coluna tem-se os valores da estatística PRESS e na terceira coluna tem-se os valores de AIC.

Nestes gráficos da Figura 6 não são mostrados os valores de PRESS acima 0.002, por isto os gráficos de PRESS não exibem as estatísticas calculadas para todos os modelos. Também não são exibidos todos os valores de R^2 ajustado para a amostra da Turbina C, pois os Modelos 9 e 10 não foram calculados devido ao número de graus de liberdade ser menor que número de coeficientes destes polinômios.

A Tabela 2 mostra um resumo dos subconjuntos de modelos selecionados conforme os critérios da Figura 6.

Nesta tabela observa-se que, em relação à Turbina A, o R^2 ajustado é otimizado pelos Modelos 2, 3 e 4 enquanto que PRESS e AIC são mínimos com o modelo 5. Em relação à Turbina B, o R^2 ajustado é ótimo nos Modelos 2 e 3 enquanto que PRESS e AIC são mínimos nos Modelos 5 e 6. Em relação à Turbina C, o R^2 ajustado é otimizado pelos Modelos 1, 2, 3 e 4 enquanto que o PRESS é mínimo nos Modelos 2 e 3 e AIC é mínimo no Modelo 4. Em relação à Turbina D, o R^2 ajustado é otimizado pelos

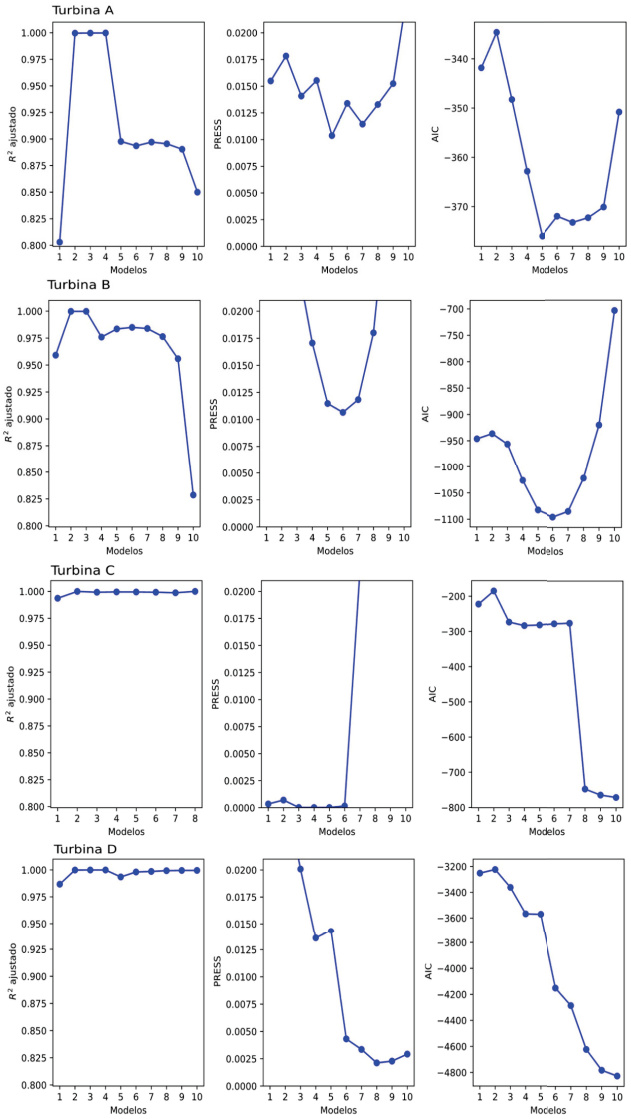

Figura 6: Seleção de modelos nas amostras das Turbinas A, B, C e D. Aplicação do algoritmo best subsets com as métricas R^2 ajustado, PRESS e AIC.

Modelos 2, 3 e 4 enquanto que o PRESS é mínimo com o Modelo 8 e o AIC é mínimo com o modelo 10.

► Validação

Nesta etapa, os modelos selecionados em cada subconjunto mostrado na Tabela 2 são avaliados em relação à distribuição dos resíduos de ajuste. Os gráficos das Figuras 7, 8, 9 e 10 mostram a dispersão dos resíduos em relação à altura de queda (na primeira coluna), a dispersão dos resíduos em relação à vazão (na segunda coluna) e o gráfico quantil-quantil (na terceira coluna) que forma uma linha aproximadamente reta de pontos quando estes têm distribuição normal.

Como pode ser observado na Figura 7, os resíduos dos modelos selecionados na Turbina A indicam que o Modelo 3 é aquele com o menor número de variáveis que resulta em média zero, variância constante e resíduos normais, enquanto que os resíduos do Modelo 2 indicam

Modelo	Turbina A	Turbina B	Turbina C	Turbina D
1			√	
2	√	√	√	√
3	√	√	√	√
4	√		√	√
5	√	√		
6		√		
7				
8				√
9				
10				√

Tabela 2: Subconjuntos de modelos com melhores ajustes nas amostras das Turbinas A, B, C e D.

média diferente de zero e não normalidade, e os resíduos dos Modelos 4 e 5 apresentam as mesmas características do Modelo 3.

Na Figura 8, tem-se que os resíduos dos modelos selecionados na Turbina B indicam que o Modelo 5 é aquele com o menor número de variáveis que resulta em média zero, variância constante e resíduos normais, enquanto que os resíduos dos Modelos 2 e 3 indicam média diferente de zero e variância com acentuada correlação cíclica, o Modelo 6 apresenta características parecidas com o Modelo 5.

Na Figura 9, tem-se que os resíduos dos modelos selecionados na Turbina C indicam que o Modelo 3 é aquele com o menor número de variáveis que resulta em média zero, variância constante e resíduos normais, enquanto que os resíduos dos Modelos 1 e 2 indicam não normalidade e o Modelo 4 apresenta características parecidas com o Modelo 3.

A Figura 10 mostra que na amostra da Turbina D o Modelo 3 é aquele com o menor número de variáveis que resulta em média zero, variância constante e resíduos normais. O resíduo do Modelo 2 indica variância não constante e média diferente de zero. E os Modelos 8 e 10 tem resíduos com as características semelhantes ao Modelo 3. Os gráficos de resíduos do Modelo 10 não foram incluídos nesta figura pois não acrescentavam informações relevantes.

Deste modo, os modelos selecionados como melhor ajuste nas amostras das Turbinas A, B, C e D são os Modelos 3, 5, 3 e 3.

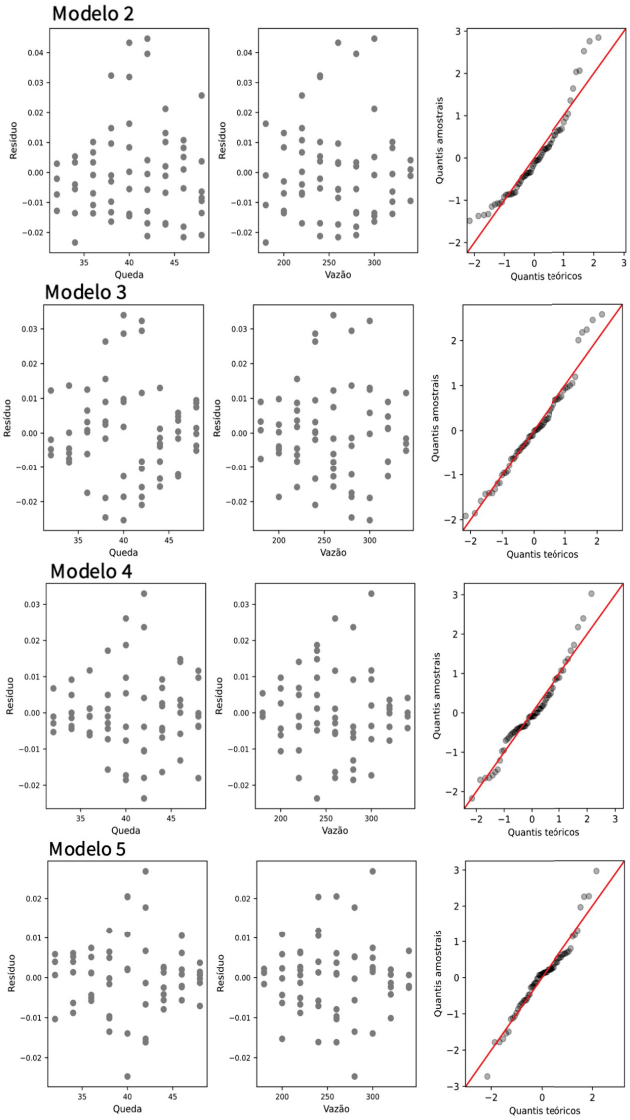

Figura 7: Gráfico de resíduos e de probabilidade normal para o subconjunto de modelos selecionados a partir da amostra da Turbina A pelo algoritmo best subsets.

4 Discussão

O experimento deste estudo foi realizado com poucos dados disponíveis e utilizando quatro amostras de curvas de colina sem catalogação dos tipos de turbina das quais as amostras foram retiradas.

A análise das amostras indica que as curvas de colina avaliadas são provenientes de hidrelétricas bastante diferentes, onde, provavelmente, a amostra da Turbina B é de uma hidrelétrica de grande porte, a amostra da Turbina C é de uma pequena hidrelétrica e as demais amostras (Turbinas A e D) são de hidrelétricas de porte médio. Ou seja, aparentemente, as amostras utilizadas neste experimento são de tipos (modelos) de turbinas diferentes [1], pois cada tipo de turbina é fabricado para operar em condições de vazão e altura de queda distintas.

E isto é algo positivo para este estudo, que busca determinar uma forma geral para as curvas de rendimento de turbinas hidrogeradoras, pois aumenta a representatividade das amostras. Mas, ainda assim, quatro amostras é uma quantidade muito pequena para se ter alguma representatividade da população de turbinas instaladas nas hidrelétricas do SIN, tão pouco representa os sistemas elétricos de outros países.

Além disso, conforme descrito anteriormente, as amostras foram coletadas de formas diferentes e para finalidades distintas, o que impede sejam feitas quais quer relações entre os modelos selecionados a partir de cada amostra. Assim sendo, a maior ocorrência de polinô-

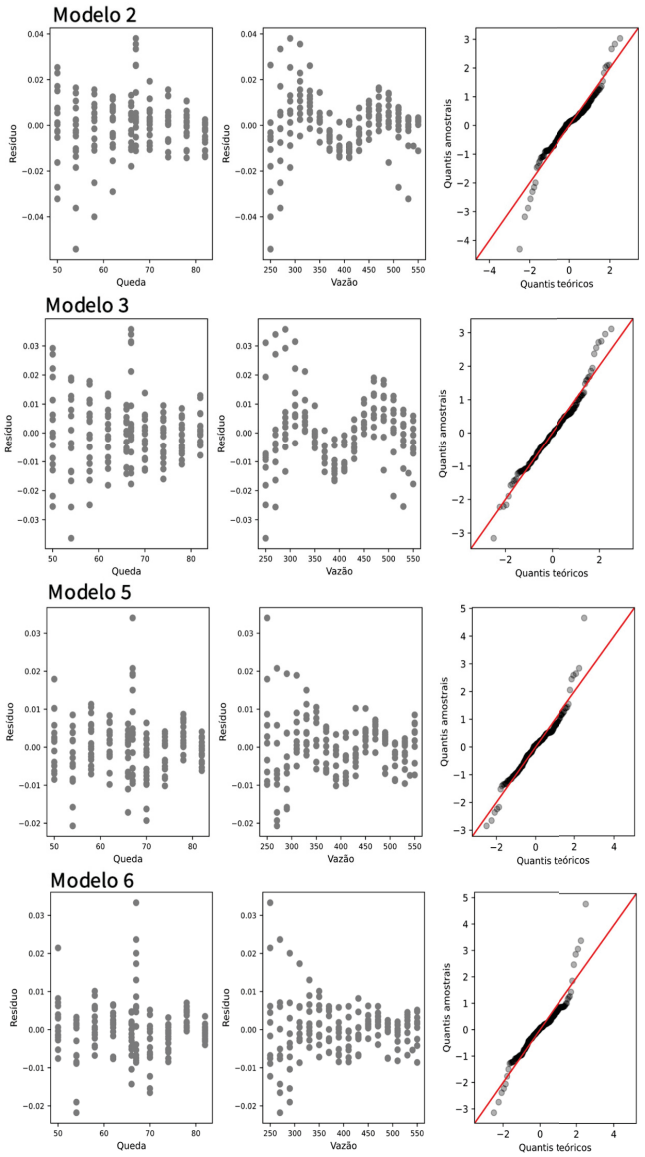

Figura 8: Gráfico de resíduos e de probabilidade normal para o subconjunto de modelos selecionados a partir da amostra da Turbina B pelo algoritmo best subsets.

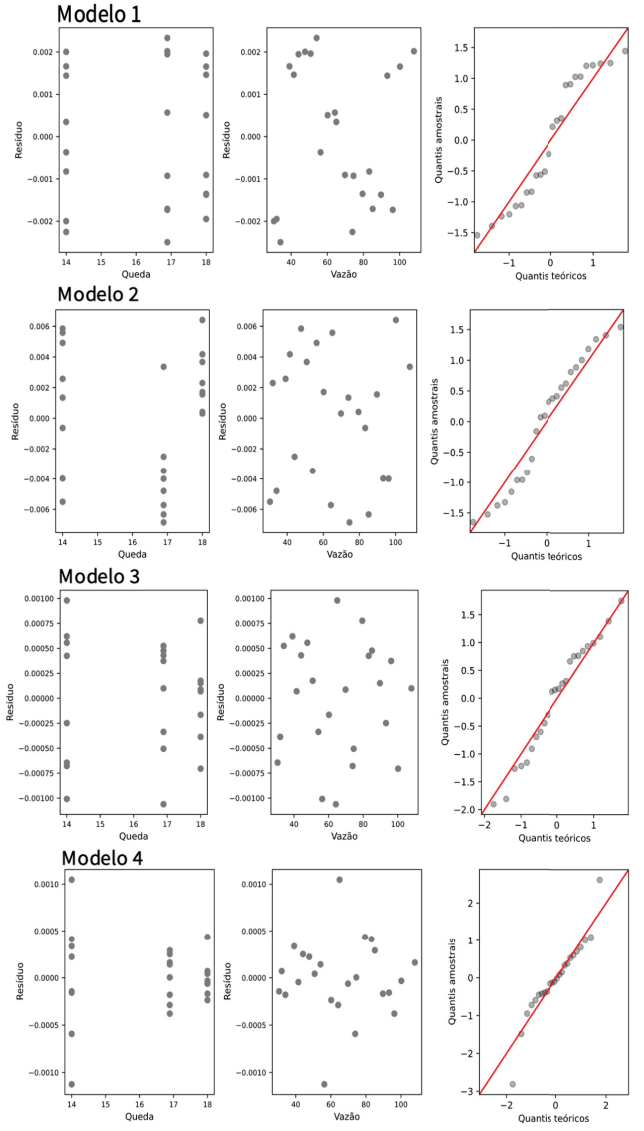

Figura 9: Gráfico de resíduos e de probabilidade normal para o subconjunto de modelos selecionados a partir da amostra da Turbina C pelo algoritmo best subsets.

mios de terceira ordem observada entre os modelos com melhor ajuste em cada amostra é uma coincidência.

Entretanto, a analise de quais modelos apresentaram os melhores ajustes em cada amostra trás evidencias de que a curva de rendimento pode ser bem representada por funções polinomiais, pois os pontos observados nos gráficos de dispersão das amostras foram bem ajustados por polinômios de até quinta ordem.

Por sua vez, o emprego do Algoritmo *Best subsets* mostrou que não é preciso varrer um grande espaço de busca para encontrar polinômios que satisfaçam os pressupostos do método dos mínimos quadrados e retornem bons ajuste nas amostras.

As métricas escolhidas para o experimento resultaram em diferentes modelos com melhor ajuste, sendo que cada métrica contribuiu individualmente para selecionar pelo menos um modelo com o melhor ajuste em uma das quatro amostras.

A estatística R^2 ajustado dos modelos selecionados pelo melhor ajuste indica que estes modelos explicam mais de 95% da variância das amostras analisadas.

O PRESS calculado para estes modelos se mostrou bastante baixo. A pesar dos valores de PRESS estarem fora dos limites definidos nos gráficos da Figura 6, o PRESS da Turbina B - Modelo 3 é de 0.02466 e o PRESS da Turbina D - Modelo 3 é de 0.02008, o que são valores muito próximos dos limites estipulados.

No geral, a estatística AIC selecionou modelos válidos com um maior número de variáveis explicativas. Mas, a maioria destes modelos não foi selecionada como melhor ajuste em detrimento de modelos mais simples com as mesmas características. Apenas na amostra da Turbina B é que esta estatística indicou o modelo mais simples e

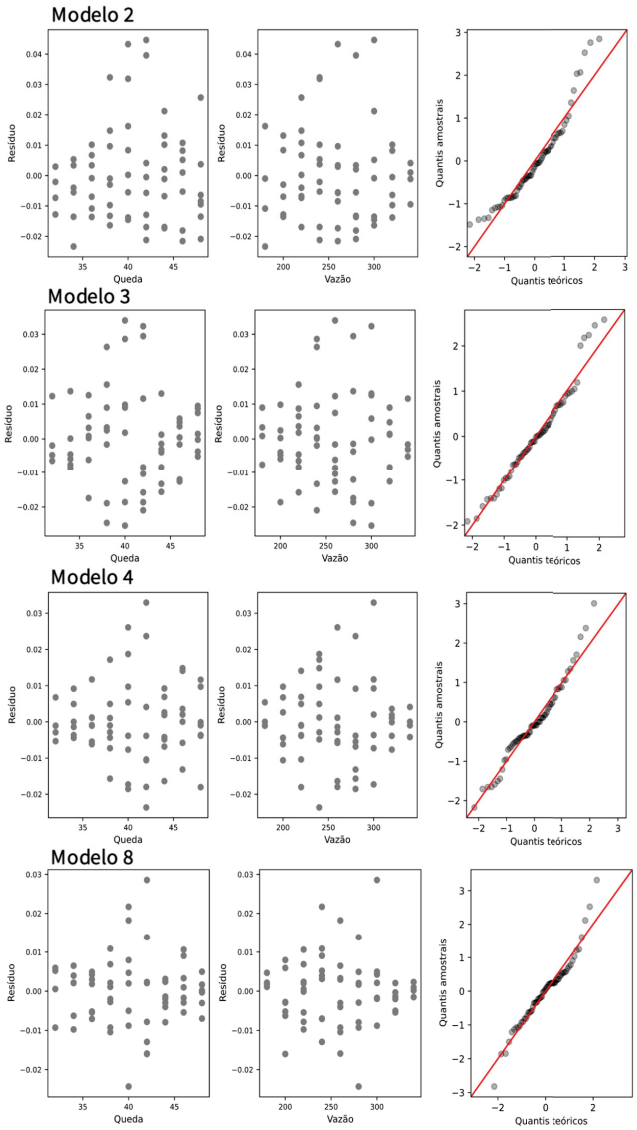

Figura 10: Gráfico de resíduos e de probabilidade normal para o subconjunto de modelos selecionados a partir da amostra da Turbina D pelo algoritmo best subsets.

válido de acordo com os resíduos.

Conforme a Tabela 2, o polinômio de segunda ordem proposto pela EPE (Modelo 2) está no subconjunto de melhores modelos de todas as amostras, mas este modelo não apresentou bons ajustes na etapa de validação resultando em resíduos sem distribuição normal. Já o Modelo 1 não apresentou bons ajustes em nenhuma das métricas.

Apesar disto, não é possível afirmar que estes modelos estão errados, mas sim que foram melhorados pela a adição de mais variáveis explicativas. Pois, a análise residual mostrou que a adição de novas variáveis (até certo ponto) aumenta o poder explicativo dos modelos polinomiais e reduz os erros de previsão.

5 Conclusão

Conforme discutido, os resultados apresentados não podem ser generalizados pois utilizam uma pequena quantidade de amostras que tem pouca representatividade sobre a população de turbinas e, principalmente, porque as amostras não foram inicialmente coletadas com a finalidade de se fazer o experimento deste estudo.

Ainda assim, este estudo lança um olhar sobre a importância de se definir métodos objetivos de coleta de dados para estimar uma forma geral para a função de rendimento de turbinas hidrogeradoras. E também indica que os modelos polinomiais tem capacidade de explicar a maior parte da variabilidade das amostras com modelos de ordem baixa.

Como continuidade, propõe-se avaliar os modelos polinomiais a partir de amostras coletadas de uma mesma curva colina com pontos dentro e fora da faixa operativa da turbina para observar o reflexo da censura desta amostra nos modelos selecionados como melhor ajuste.

Outra proposta é avaliar os efeitos da regularização L1 sobre a seleção de variáveis dos modelos e também analisar outras formas para a função de rendimento como, por exemplo, os modelos lineares generalizados que com a função de ligação *Gamma* pode modelar dados assimétricos como é o caso das amostras de rendimento das Turbinas B, C e D. Visto que os gráficos de dispersão das amostras indicam formas cônicas não lineares para a função de rendimento, outra opção também seria o estudo de modelos não lineares.

Referências

[1] Ministério de Minas e Energia. Pequenos aproveitamentos hidroelétricos - soluções energéticas para a amazônia. `https://www.mme.gov.br/luzparatodos/downloads/Solucoes_Energeticas_para_a_Amazonia_Hidroeletrico.pdf`, 2008. Acesso em: 19 de outubro de 2020.

[2] Passos I. O. Metodologia de obtenção de curvas de colina usando redes neurais para geração hidrelétrica. Master's thesis, Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará, 2011.

[3] Arce Encina A. S. *Despacho Ótimo de Unidades Geradoras em Sistemas Hidrelétricos via Heurística Baseada em Relaxação Lagrangeana e Programação Dinâmica*. PhD thesis, Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas, 2006.

[4] Diniz A. L., Esteves P. P. I., and Sagastizabal C. A. A mathematical model for the efficiency curves of hydroelectric units. *IEEE Power Engineering Society General Meeting*, pages 1–7, 2007.

[5] Empresa de Pesquisa Energética. Ministério de Minas e Energia. Estudos para a licitação da geração -

metodologia de cálculo de parâmetros energéticos médios: Rendimento e perda hidráulica, epe-dee-re-037/2011-r2. https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-497/EPE-DEE-RE-037-2011-r2.pdf, 2008. Acesso em: 7 de julho de 2020.

[6] Hidalgo I. G., Fontane D. G., Lopes J. E. G., Andrade J. G. P., and Angelis A. F. Efficiency curves for hydroelectric generating unit. *2007 IEEE Power Engineering Society General Meeting*.

[7] Phillip B. Palmer and Dennis G. O'Connell. Regression analysis for prediction: understanding the process. cardiopulm phys ther j. *Cardiopulmonary Physical Therapy Journal*, 2009.

[8] Pindyck R. S. and Rubinfeld D. L. *Econometria - Modelos e Previsões*. Editora Campus, 2004.

[9] Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. *Applied Linear Statitical Models*. McGraw-Hill Irwin, 5 edition, 2004.

[10] G. James; D. Witten; T. Hastie; R. Tibshirani. *An Introduction to Statistical Learning - with Applications in R*. Springer, 2013.

[11] Iain Pardoe. Stat 501 online course materials website / 10.5 - information criteria and press. https://online.stat.psu.edu/stat501/lesson/10/10.5, 2020. Acesso em: 18 de outubro de 2020.