Universidade Federal de Minas Gerais
Departamento de Ciência da Computação
Graduação em Sistemas de Informação

# Smooth Augmentation:
# Suavização e Augmentation em Perfilagem de Poços

Romana Gallete Mota

Orientador: Prof. Renato M. Assunção, Dr.

Belo Horizonte, 2026

**Monografia**

**Smooth Augmentation: Suavização e Augmentation em Perfilagem de Poços**

Monografia submetida à banca examinadora designada pelo Colegiado Didático do Curso de Graduação em Sistemas de Informação da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Belo Horizonte, 2026

*"To those who helped me navigate the virtual insanity of these past few years..."*

# Resumo

O presente trabalho investiga novas metodologias de aumento de dados aplicadas ao treinamento de modelos de aprendizado profundo com registros de perfilagem de poços. A principal motivação desta pesquisa reside na escassez crônica e no alto custo de obtenção de dados geológicos de subsuperfície, fatores que impõem severas restrições ao volume dos conjuntos de treinamento e impactam negativamente a capacidade de generalização das redes neurais. Trata-se de um desafio crítico na análise de dados geofísicos, visto que, no cenário de escassez de dados, o uso de registros de fontes geograficamente heterogêneas pode introduzir vieses indesejados devido a particularidades regionais das bacias sedimentares.

Para mitigar essas limitações, propõe-se uma metodologia estruturada em etapas sequenciais e especificamente adaptada às restrições físicas do domínio geológico, denominada *Smooth Augmentation*. A abordagem supera os gargalos das técnicas clássicas da literatura, como *jitter*, *scaling* e *time warping* tradicionais, as quais demandam calibração manual e iterativa de parâmetros estatísticos globais e frequentemente geram anomalias que violam as leis físicas que regem os sistemas rocha-fluido. O pipeline desenvolvido atua de forma livre de ajustes empíricos, extraindo os componentes de alta frequência e de ganho local diretamente a partir da decomposição estocástica de cada perfil em relação a uma linha de base obtida via filtragem por mediana. Adicionalmente, o pipeline incorpora um algoritmo de segmentação adaptativa por *breakpoints* para garantir a invariância de escala vertical e utiliza a técnica de deslocamento circular intra-janela para salvaguardar a correlação serial e a continuidade estratigráfica local.

A validação do método foi conduzida por meio de um protocolo duplo, integrando análises quantitativas de consistência petrofísica e uma auditoria visual qualitativa em formato duplo-cego com especialistas sêniores em geociências. Os resultados analíticos revelaram que as variações sintéticas preservaram com fidelidade a distribuição multimodal e a malha de correlação multivariada original das rochas. No teste com especialistas, obteve-se uma taxa de abstenção por incerteza visual de 51,00% e uma acurácia líquida global de 46,94%. Dado que estes índices convergem estocasticamente para o comportamento de um classificador aleatório puro (*chute ao acaso*), comprova-se o realismo morfológico e o hiper-realismo das amostras geradas. Com isso, o trabalho consolida uma ferramenta automática, robusta e escalável, apta a expandir de forma consistente bases de dados geofísicos e potencializar arquiteturas preditivas complexas de subsuperfície.

**Palavras-chave:** Perfilagem de Poços. Aprendizado Profundo. Aumento de Dados. Filtro de Mediana. Deslocamento Circular. Geociências.

# Abstract

This work investigates novel data augmentation methodologies applied to the training of deep learning models using well logging records. The primary motivation for this research stems from the chronic scarcity and high cost of acquiring subsurface geological data, factors that severely restrict the volume of training sets and negatively impact the generalization capability of neural networks. This poses a critical challenge in geophysical data analysis since, in a data-scarce scenario, utilizing records from geographically heterogeneous sources can introduce undesired biases due to regional peculiarities of sedimentary basins.

To mitigate these limitations, we propose a methodology structured in sequential stages and specifically adapted to the physical constraints of the geological domain, termed *Smooth Augmentation*. This approach overcomes the bottlenecks of classical literature techniques—such as traditional *jitter*, *scaling*, and *time warping*—which demand manual and iterative calibration of global statistical parameters and frequently generate anomalies that violate the physical laws governing rock-fluid systems. The developed pipeline operates in a tuning-free manner, extracting high-frequency components and local gain directly from the stochastic decomposition of each log against a baseline obtained via median filtering. Additionally, the pipeline incorporates an adaptive segmentation algorithm based on *breakpoints* to ensure vertical scale invariance and employs an intra-window circular shifting technique to safeguard serial correlation and local stratigraphic continuity.

The validation of the method was conducted through a dual protocol, integrating quantitative analyses of petrophysical consistency and a qualitative visual audit in a double-blind format with senior geosciences specialists. Analytical results revealed that the synthetic variations faithfully preserved the original multimodal distribution and multivariant correlation mesh of the rocks. In the expert test, a visual uncertainty abstention rate of 51.00% and a global net accuracy of 46.94% were obtained. Given that these rates stochastically converge to the behavior of a pure random classifier (random guessing), the morphological realism and hyper-realism of the generated samples are formally proven. Consequently, this work consolidates an automatic, robust, and scalable tool, capable of consistently expanding geophysical databases and empowering complex subsurface predictive architectures.

**Keywords:** Well Logging. Deep Learning. Data Augmentation. Median Filter. Circular Shifting. Geosciences.

# Sumário

**1 Introdução** **6**
1.1 Motivação . . . . . 6
1.2 Objetivos . . . . . 7

**2 Referencial Teórico** **8**
2.1 Técnicas Clássicas de Aumento de Dados . . . . . 8
2.1.1 Jitter . . . . . 8
2.1.2 Scaling . . . . . 9
2.1.3 Time Warping . . . . . 9
2.2 Limitações e Desafios no Domínio Geológico . . . . . 9

**3 Pipeline Orientado por Dados** **12**
3.1 Suavização e Filtragem . . . . . 12
3.2 Segmentação Adaptativa e Invariância de Escala . . . . . 14
3.3 Derivação de Componentes e Perturbação . . . . . 15
3.4 Reconstrução das Séries Aumentadas . . . . . 17

**4 Validação e Avaliação dos Dados** **19**
4.1 Inspeção Petrofísica e Coerência Multivariada . . . . . 19
4.2 Teste Duplo-Cego com Especialistas . . . . . 21
4.3 Discussão e Diferenciais Frente aos Métodos Tradicionais . . . . . 22

**5 Conclusões** **24**

# Capítulo 1

# Introdução

O avanço do aprendizado profundo tem possibilitado importantes contribuições em diversas áreas, incluindo as geociências. No entanto, a aplicação desses métodos em dados de perfilagem de poços enfrenta desafios específicos, principalmente relacionados à escassez e heterogeneidade dos registros disponíveis. Nesse contexto, surge a necessidade de investigar estratégias que permitam ampliar a quantidade e a qualidade dos dados utilizados no treinamento de modelos, sem comprometer a fidelidade das informações originais.

## 1.1 Motivação

A motivação desta pesquisa surge da necessidade de treinar modelos de aprendizado profundo utilizando dados de perfilagem de poços. Esse processo exige grandes volumes de dados, mas a disponibilidade de tais registros é limitada, o que representa um desafio para a área. Uma alternativa seria utilizar dados de outros repositórios; no entanto, os perfis geológicos são fortemente influenciados por fatores regionais, resultando em padrões distintos entre diferentes conjuntos de dados ao redor do mundo. A incorporação de dados de fontes heterogêneas, em um contexto de escassez de dados da região alvo, poderia introduzir vieses indesejados no modelo, tornando-o altamente especializado nos padrões específicos dos dados de treinamento e comprometendo sua capacidade de generalização para poços do domínio-alvo.

Diante dessas limitações, estratégias de aumento de dados passaram a ser investigadas. Na ausência de técnicas de aumento projetadas especificamente para dados de perfilagem de poços, tais registros são tratados como séries temporais. Essa abordagem permite a aplicação de métodos tradicionais de aumento amplamente utilizados nesse domínio, como jitter, scaling, magnitude warping e time warping. Contudo, essas técnicas apresentam desafios de implementação, como o alto custo de refinamento do jitter e a necessidade de definir escalas relativas no scaling. Assim, identifica-se uma oportunidade para desenvolver abordagens inovadoras de aumento de dados que mantenham a simplicidade de implementação, ao mesmo tempo em que assegurem eficiência e preservação das características originais dos dados.

## 1.2 Objetivos

O objetivo central deste projeto é propor e avaliar novos métodos de aumento de dados especificamente adaptados a registros de perfilagem de poços. Busca-se superar as limitações das técnicas tradicionais, desenvolvendo soluções que conciliem simplicidade de implementação, eficiência computacional e preservação das propriedades dos dados, de modo a aprimorar a capacidade de generalização de modelos de aprendizado profundo em cenários de dados restritos.

Para alcançar esse objetivo geral, o projeto também se propõe a investigar o potencial de técnicas de *data augmentation* inspiradas em abordagens já consolidadas no tratamento de séries temporais, como *jitter*, *scaling* e *time warping*. A adaptação dessas técnicas ao contexto da perfilagem de poços permitirá compreender quais ajustes são necessários para que elas sejam eficazes nesse domínio, garantindo ao mesmo tempo a preservação das tendências e variações geológicas presentes nos registros originais.

Além de avaliar o desempenho dessas técnicas tradicionais adaptadas, pretende-se propor novas variações e extensões que combinem simplicidade e capacidade de geração de dados sintéticos fiéis. Dessa forma, o projeto visa não apenas reproduzir soluções conhecidas, mas também avançar no desenvolvimento de abordagens inovadoras, capazes de aumentar a diversidade dos dados de forma consistente, auxiliando na construção de modelos mais robustos e generalizáveis.

Para superar as limitações observadas nas técnicas clássicas de aumento de dados, propõe-se uma abordagem inédita, desenvolvida especificamente para registros de perfilagem de poços. Métodos tradicionais, como *jitter* e *scaling*, frequentemente falham em considerar as características particulares dos perfis geológicos, que apresentam padrões variáveis entre diferentes poços. A necessidade de refinamentos manuais, aliada a outros desafios, pode introduzir inconsistências e tornar essas técnicas menos eficientes quando aplicadas à *data augmentation* de perfis de poços.

Como alternativa, propõe-se uma abordagem mais eficaz baseada em um processo de aumento de dados estruturado em etapas bem definidas, projetado para introduzir variações controladas ao mesmo tempo em que preserva a estrutura essencial dos registros. Essa metodologia busca garantir simplicidade de implementação, consistência nos resultados e adaptabilidade às diferentes distribuições de dados encontradas em contextos reais.

# Capítulo 2

# Referencial Teórico

A aplicação de técnicas de aprendizado profundo em séries temporais tem crescido exponencialmente, impulsionando a necessidade de grandes volumes de dados para o treinamento de modelos robustos. No entanto, a coleta de dados em domínios especializados, como a geologia e a engenharia de petróleo, são processos de alto custo, dependentes de especialistas e propensos a escassez (1). Neste contexto, a técnica de *data augmentation* (aumento de dados) consolidou-se como uma estratégia fundamental para expandir artificialmente o conjunto de treinamento, melhorar a capacidade de generalização dos modelos e mitigar o *overfitting* (2, 3).

O princípio fundamental do aumento de dados consiste em aplicar transformações matemáticas ou estocásticas às amostras existentes para gerar novas variações que mantenham a consistência conceitual da classe original (invariância de rótulo). Em cenários de dados escassos, essas técnicas atuam como um regularizador do modelo, forçando a rede neural a aprender características invariantes e robustas em vez de memorizar ruídos ou padrões específicos do conjunto de treino limitado (2, 4). No entanto, a aplicação direta de métodos genéricos de aumento em dados geológicos, especificamente em perfis de poços, apresenta desafios particulares que motivam a abordagem desenvolvida neste trabalho.

## 2.1 Técnicas Clássicas de Aumento de Dados

Tradicionalmente, os métodos de aumento de dados em séries temporais buscam criar variações sintéticas preservando as características originais dos dados. De acordo com a taxonomia proposta por Iglesias et al. (4) e revisada por Iwana e Uchida (3), as técnicas clássicas operam predominantemente no domínio do tempo ou da magnitude. Entre as mais frequentemente adotadas, destacam-se o *jitter* (adição de ruído), o *scaling* (redimensionamento de magnitude) e o *time warping* (deformação temporal) (1, 5).

### 2.1.1 Jitter

O método de *jittering* consiste na introdução de ruído aditivo, tipicamente Gaussiano, a cada ponto da série temporal original. Matematicamente, dada uma série temporal original $X = \{x_1, x_2, \ldots, x_T\}$, a nova série sintetizada $X'$ é definida por:

$$x'_t = x_t + \epsilon_t, \quad \text{onde } \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

O objetivo central desta técnica é simular erros intrínsecos de medição, imperfeições de sensores ou variações ambientais aleatórias (3). Ao injetar esse ruído de baixa amplitude, impede-se que o modelo de aprendizado profundo dependa excessivamente de valores exatos ou de variações de altíssima frequência, forçando-o a focar na tendência estrutural subjacente do sinal.

### 2.1.2 Scaling

O *scaling* altera a magnitude global da série temporal multiplicando todos os seus pontos por um fator escalar constante ou ligeiramente variável. A transformação pode ser expressa como:

$$x'_t = \alpha \cdot x_t, \quad \text{onde } \alpha \sim \mathcal{N}(1, \sigma^2)$$

Nessa transformação, $\alpha$ é um escalar determinado de forma determinística ou amostrado a partir de uma distribuição de probabilidade (por exemplo, $\alpha \sim \mathcal{N}(1, \sigma^2)$ ou $\alpha \sim \mathcal{U}(1-\beta, 1+\beta)$). O propósito do *scaling* é treinar o modelo para ser invariante a mudanças de ganho ou amplitude no sinal, simulando sensores com diferentes calibrações ou fenômenos que ocorrem com intensidades distintas, sem alterar o padrão morfológico ou a frequência dos eventos (4).

### 2.1.3 Time Warping

Diferente das técnicas anteriores que alteram a magnitude, o *time warping* atua no domínio temporal, modificando a localização cronológica e a taxa de ocorrência dos eventos. Esta técnica comprime ou expande segmentos da série de forma não linear (5). O mapeamento pode ser entendido através de uma função de deformação temporal $\tau(t)$, tal que:

$$x'_t = x_{\tau(t)}$$

Na prática, isso é frequentemente realizado selecionando janelas temporais aleatórias e alterando a sua taxa de amostragem, o que faz com que certos trechos da série pareçam "estendidos"(baixa frequência local) e outros "comprimidos"(alta frequência local) (3). Essa abordagem simula variações na velocidade de ocorrência dos padrões, o que é crucial em cenários onde a dinâmica temporal do processo físico pode acelerar ou desacelerar.

## 2.2 Limitações e Desafios no Domínio Geológico

Embora eficazes em domínios gerais, como reconhecimento de atividades humanas ou áudio, a aplicação automática destes métodos em perfis de poços enfrenta limitações críticas que podem distorcer o sinal geológico e violar leis físicas subjacentes. A principal barreira reside na necessidade de preservar as características intrínsecas dos perfis,

que refletem propriedades físicas reais das rochas e fluidos (como porosidade, densidade e radioatividade natural), e não apenas variações estatísticas arbitrárias.

A aplicação de *jitter* exige a seleção extremamente cuidadosa do desvio padrão $\sigma$ do ruído. Se o ruído for excessivo, o sinal geológico e feições diagnósticas sutis (como transições abruptas de fácies ou pequenas camadas de folhelho) são ofuscados; se for muito pequeno, a variação produzida é negligenciável para o aprendizado do modelo. Em perfis de poços, diferentes intervalos geológicos e diferentes bacias exibem padrões estatísticos e níveis de ruído instrumentais distintos. Isto exigiria um ajuste manual, iterativo e empírico dos parâmetros de ruído para cada caso. Essa calibração manual é custosa, reduz drasticamente a escalabilidade do pipeline e limita a utilidade prática do método em grandes bases de dados multi-poço.

O método de *scaling* apresenta inconsistências matemáticas severas ao afetar diferentes regiões da série de forma desigual devido à sua natureza multiplicativa. Considere, por exemplo, um sinal físico com duas faixas características de valores associadas a diferentes litologias: uma faixa inferior próxima a $0,1$ e outra superior próxima a $0,8$. Ao aplicar um fator de escala aleatório $\alpha = 1,1$, a transformação linear gera as seguintes variações:

- **Região Inferior:** $x'_t = 1,1 \times 0,1 = 0,11 \implies \Delta = 0,01$
- **Região Superior:** $x'_t = 1,1 \times 0,8 = 0,88 \implies \Delta = 0,08$

Como resultado, os valores na faixa inferior permanecem artificialmente próximos à distribuição original, enquanto os valores na faixa superior sofrem um deslocamento oito vezes maior, desviando-se consideravelmente do comportamento físico e geológico realista.

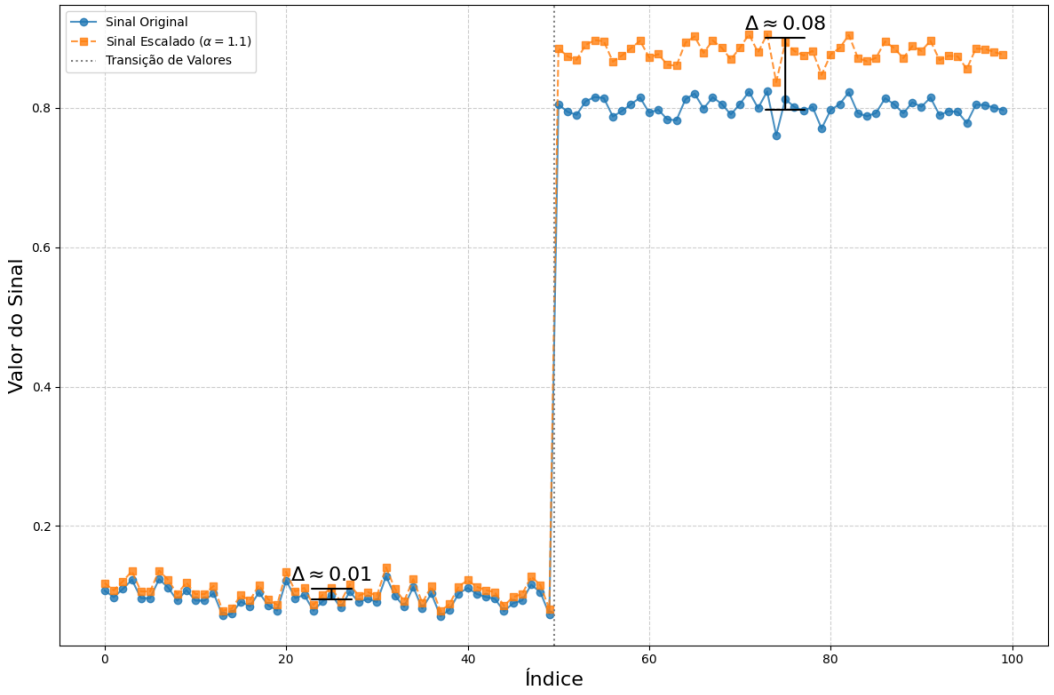



Figura 2.1: Efeito inconsistente do método de *scaling* em diferentes faixas de valores.

Essa inconsistência distorce as propriedades estatísticas da série, altera as distribuições multimodais típicas de perfis geológicos e reduz a fidelidade das amostras sintéticas, gerando dados que violam os limites físicos das rochas e, consequentemente, degradando o desempenho dos modelos de Deep Learning.

Por fim, o *time warping* apresenta severas restrições no contexto de perfis de poço, onde o eixo temporal é, na verdade, o eixo da profundidade. Alterar a taxa de amostragem vertical de forma arbitrária equivale a modificar a espessura aparente das camadas geológicas sem critério físico. Na geologia de subsuperfície, a espessura e o gradiente de transição de uma camada respondem a taxas de sedimentação e processos diagenéticos específicos; deformações temporais puramente estocásticas desconfiguram essas assinaturas depositionais.

Essas limitações demonstram de forma clara que os métodos clássicos de aumento de dados não se transferem naturalmente para o domínio de perfis de poço, evidenciando a necessidade premente de uma estratégia guiada por dados ou baseada em física que introduza variabilidade realista, mantendo estritamente a consistência geológica.

# Capítulo 3

# Pipeline Orientado por Dados

Para superar as limitações das abordagens clássicas, desenvolveu-se um pipeline de aumento de dados especificamente projetado para séries temporais de perfis de poço. Em vez de depender de regras de transformação estáticas com parâmetros globais ajustados manualmente, a abordagem proposta extrai a estrutura estatística e o ruído intrínseco de cada perfil para guiar a geração de novas amostras. O objetivo é introduzir variabilidade controlada e realista, respeitando rigorosamente os limites e padrões geológicos presentes nos dados originais.

O pipeline proposto divide-se em três macroetapas sequenciais: (i) **Decomposição e Segmentação**, onde o sinal é separado em componentes de diferentes frequências e dividido em intervalos de interesse; (ii) **Perturbação Estocástica Coerente**, na qual os componentes extraídos sofrem modificações locais preservando a vizinhança; e (iii) **Reconstrução Condicional**, onde o perfil sintético final é gerado. A seguir, detalha-se cada etapa do processo.

## 3.1 Suavização e Filtragem

A primeira etapa do pipeline visa produzir uma representação de baixa frequência (linha de base) do perfil de poço, atenuando flutuações de alta frequência associadas a ruídos de medição, sem distorcer as feições geológicas de larga escala.

Inicialmente, considerou-se o uso de filtros de média móvel. No entanto, a média tende a suavizar transições abruptas, borrando contatos estratigráficos e limites entre unidades sedimentares. Para mitigar esse efeito, adotou-se o filtro de mediana, que é inerentemente robusto a *outliers* e preserva de forma eficaz padrões constantes por partes, mantendo as descontinuidades essenciais para a interpretação geológica.

O operador de mediana foi adaptado para mitigar duas condições críticas do domínio prático:

- **Efeito de Borda:** Nas regiões limítrofes do perfil, o sinal é refletido para garantir que a janela de filtragem permaneça simétrica e bem definida.

- **Dados Ausentes:** Valores nulos ($NaN$), comuns devido a falhas de aquisição ou amostragem, são dinamicamente desconsiderados do cálculo da mediana dentro da janela móvel, impedindo a propagação de lacunas ao longo do sinal filtrado.

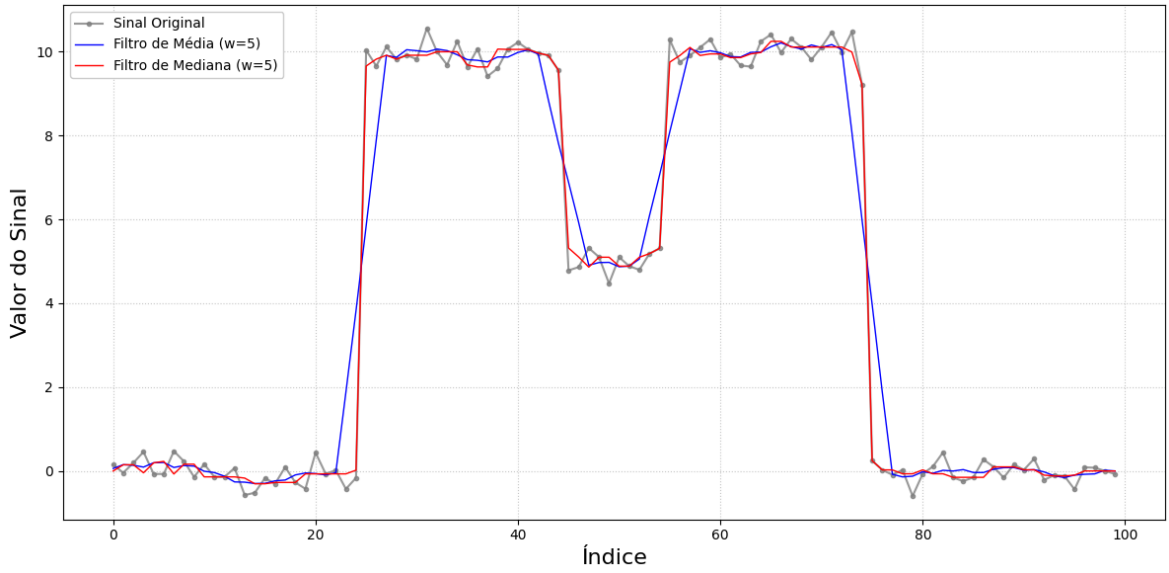



Figura 3.1: Série original e resultados da suavização através de filtros de média e mediana.

A Figura 3.1 ilustra o efeito da operação de suavização. O filtro de mediana (curva em vermelho) remove eficientemente o ruído de alta frequência e preserva os patamares litológicos, servindo como a linha de base estável ($y_{\text{smoothed}}$) necessária para as etapas subsequentes.

O comportamento desse filtro é governado pelo tamanho de sua janela de amostragem (comumente denominada *span*). A seleção desse parâmetro dita o balanço entre a preservação de feições e a eliminação do ruído:

- **Janelas reduzidas:** Promovem uma filtragem sutil. Embora preservem transições de curtíssima escala, reduzem pouco o ruído instrumental, mantendo a série sintética final muito próxima à assinatura ruidosa original.

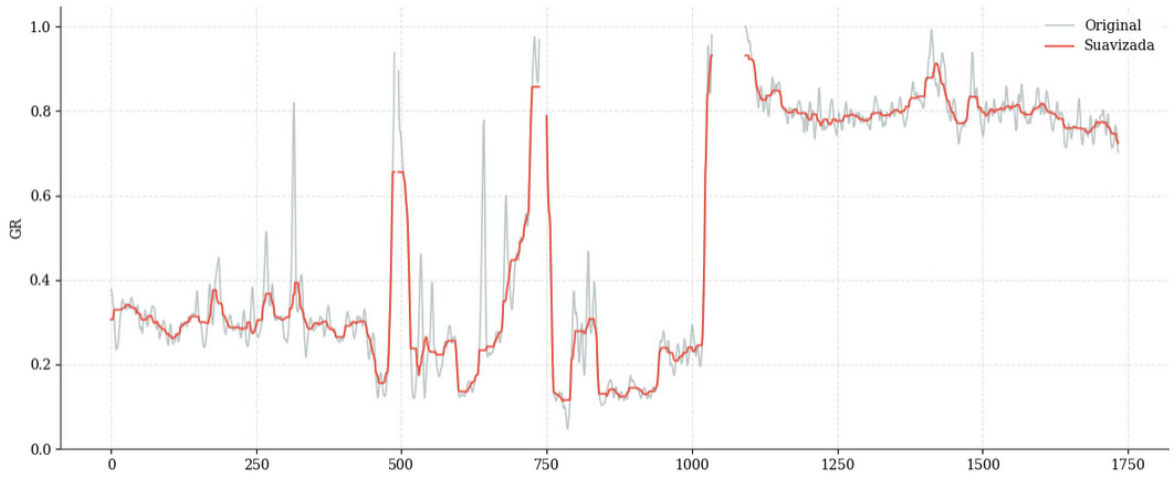



Figura 3.2: Suavização sutil através do filtro de mediana utilizando um tamanho de janela reduzido (span = 25).

- **Janelas amplas:** À medida que o tamanho da janela aumenta, o nível de suavização torna-se significativamente mais intenso, reduzindo drasticamente o ruído de alta frequência e isolando uma linha de base mais macroscópica.

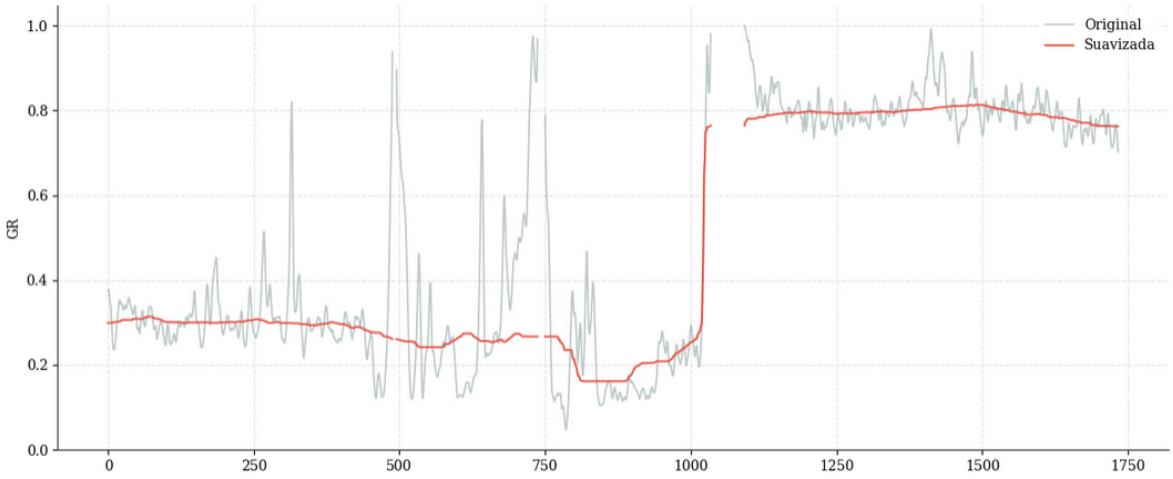



Figura 3.3: Suavização intensa através do filtro de mediana utilizando um tamanho de janela amplo (span = 250).

Para garantir que o modelo de aprendizado profundo seja exposto a um mesmo nível de suavização em todas as amostras, estabeleceu-se o uso de uma janela de tamanho fixo em pontos.

Essa decisão de projeto fundamenta-se na preservação do nível de detalhe físico e evita uma ilusão visual comum ao analisar perfis de diferentes comprimentos. Ao plotar duas séries de comprimentos distintos por inteiro - por exemplo, uma seção curta de 100 pontos e uma longa de 10.000 pontos -, uma janela proporcionalmente maior na série longa pode parecer visualmente mais adequada na tela, pois mascara o ruído em macroescala. No entanto, ao aproximar o zoom e visualizar simultaneamente uma mesma quantidade absoluta de pontos em cada uma, constata-se que a série tratada com a janela maior sofreu uma perda severa de informação geológica e detalhes estratigráficos (sub-suavização versus super-suavização). Portanto, fixar o número absoluto de pontos da janela garante que a resolução vertical e o nível de detalhe preservado sejam idênticos em todo o banco de dados, independentemente da extensão do poço.

## 3.2 Segmentação Adaptativa e Invariância de Escala

A segunda etapa identifica posições ao longo do perfil onde o sinal exibe mudanças significativas de comportamento estatístico ou de tendência, denominadas *breakpoints*. Esses pontos segmentam a série suavizada, permitindo que as transformações subsequentes ocorram de forma localizada e contextualizada com a geologia local.

A detecção de mudanças é realizada através da biblioteca *ruptures*, conforme ilustrado na Figura 3.4. Um aspecto crucial da metodologia é que o número de segmentos não é um valor absoluto fixo, mas parametrizado como uma proporção ($\gamma$) em relação ao comprimento total do perfil ($T$).

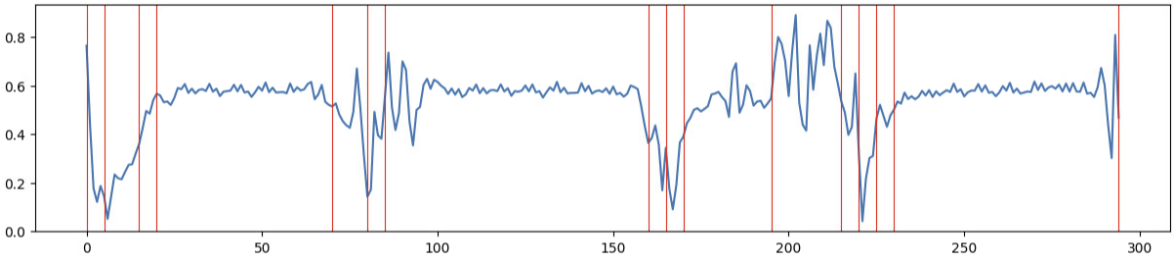


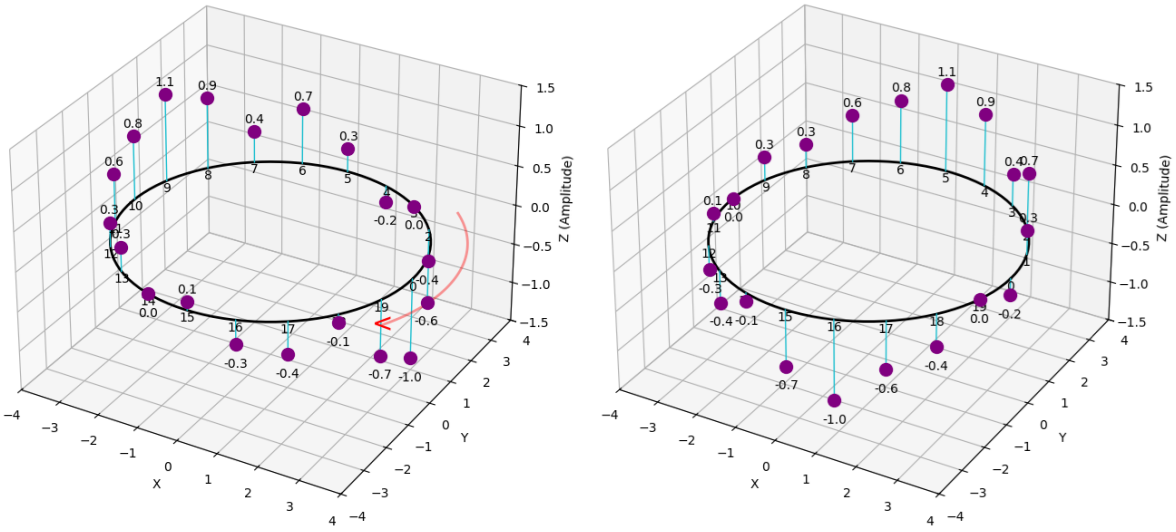

Figura 3.4: Série temporal de perfil de raios gama (GR) segmentada adaptativamente.

Essa abordagem garante a **invariância de escala vertical** do pipeline: fixando-se $\gamma = 2,5\%$, um perfil curto de 400 amostras gerará aproximadamente 10 segmentos, enquanto um perfil longo de 8.000 amostras gerará cerca de 200 segmentos. Como consequência, o comprimento médio das janelas de perturbação permanece estatisticamente similar, estabilizando o comportamento dos algoritmos de aumento independentemente da profundidade ou extensão do poço.

## 3.3 Derivação de Componentes e Perturbação

Para eliminar a necessidade de calibração manual de ruído, o pipeline extrai os componentes de variabilidade diretamente da relação matemática entre o sinal original ($y_{\text{original}}$) e a linha de base suavizada ($y_{\text{smoothed}}$). A partir dessa decomposição, o pipeline isola os operadores de *Jitter* e *Scaling*, aplicando em seguida a técnica de Deslocamento Circular (*Circular Shifting*) de forma independente dentro de cada segmento delimitado na seção anterior.

O deslocamento circular consiste em amostrar, para um segmento de comprimento $L$, um valor inteiro de translação $s \sim \mathcal{U}(0, L-1)$ e rotacionar o vetor correspondente.




(a) Componente original do segmento.

(b) Componente após rotação stocástica.

Figura 3.5: Representação esquemática do deslocamento circular ($s = 4$) aplicado individualmente sobre os componentes de um segmento.

Esta técnica foi escolhida em detrimento do embaralhamento (*shuffle*) convencional porque preserva rigorosamente a correlação serial, as relações de vizinhança e a textura do ruído real, impedindo a geração de frequências fisicamente impossíveis no domínio geológico (Figura 3.5).

A extração e a perturbação desses componentes operam sob três definições:

- **Componente Residual (Jitter):** Determinado através da diferença aditiva:

$$r(t) = y_{\text{original}}(t) - y_{\text{smoothed}}(t)$$

O sinal $r(t)$ isola o ruído instrumental e as flutuações de altíssima frequência (Figura 3.6). Uma vez extraído, o vetor de resíduos de cada segmento sofre deslocamento circular independente, gerando o componente de ruído perturbado $R_{\text{shifted}}(t)$.

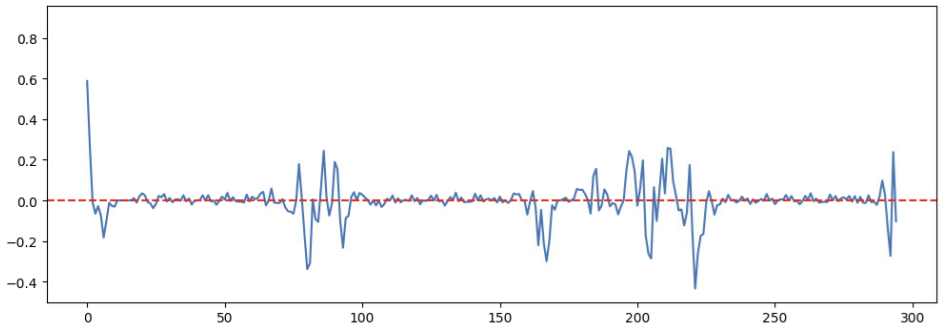



Figura 3.6: Perfil de resíduos extraídos do GR.

- **Componente de Escala (Scaling):** Mensurado por meio da razão multiplicativa ponto a ponto:

$$c(t) = \frac{y_{\text{original}}(t)}{y_{\text{smoothed}}(t)}$$

$c(t)$ mapeia os desvios proporcionais do sinal em relação à sua tendência central. Ao aplicar o deslocamento circular sobre esses coeficientes por segmento, obtém-se $C_{\text{shifted}}(t)$, permitindo alterar a assinatura de amplitude local sem violar os limites impostos pela magnitude da linha de base.

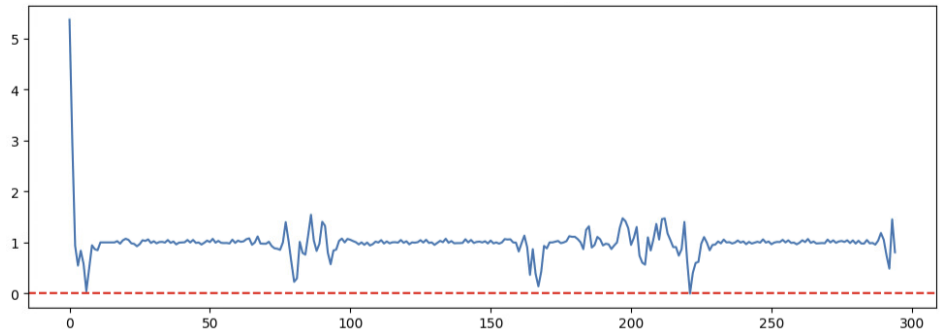



Figura 3.7: Perfil de coeficientes extraídos do GR.

- **Fatores de Deformação (Warping):** Diferente dos anteriores, o componente de deformação temporal não é extraído ponto a ponto, mas modelado diretamente por segmento. Para cada intervalo delimitado pelos *breakpoints*, amostra-se um fator escalar $\alpha$ a partir de uma distribuição Gaussiana truncada ou controlada:

$$\alpha \sim \mathcal{N}(1,0,\ 0,2^2)$$

A centralização em $1,0$ impede desvios sistemáticos na espessura total do poço, enquanto a variância controlada previne distorções morfológicas severas.

## 3.4 Reconstrução das Séries Aumentadas

A macroetapa final do pipeline consiste em recombinar a linha de base preservada ($y_{\text{smoothed}}$) com os componentes perturbados de forma a isolar ou combinar os efeitos de aumento de dados. Partindo-se do perfil original (Figura 3.8), três variantes de aumentos são formuladas de maneira independente:

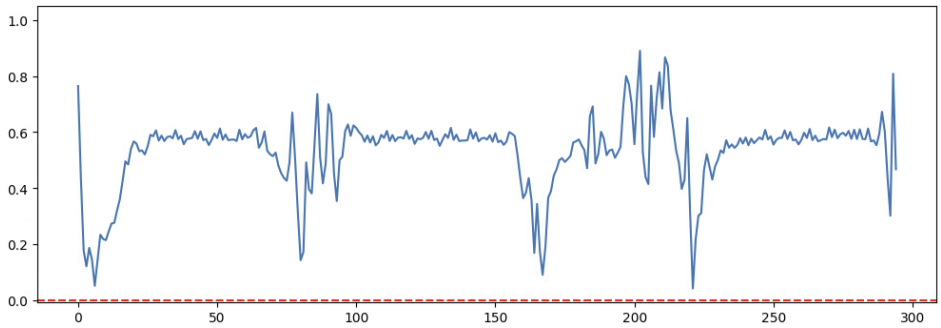



Figura 3.8: Série temporal do perfil de poço original (Referência).

**Jitter Orientado por Dados:** A série sintetizada é reconstruída somando-se os resíduos rotacionados de volta à curva suavizada. Esse método simula uma reamostragem do ruído real do sensor sobre a mesma estrutura geológica:

$$y_{\text{jitter_aug}}(t) = y_{\text{smoothed}}(t) + R_{\text{shifted}}(t)$$

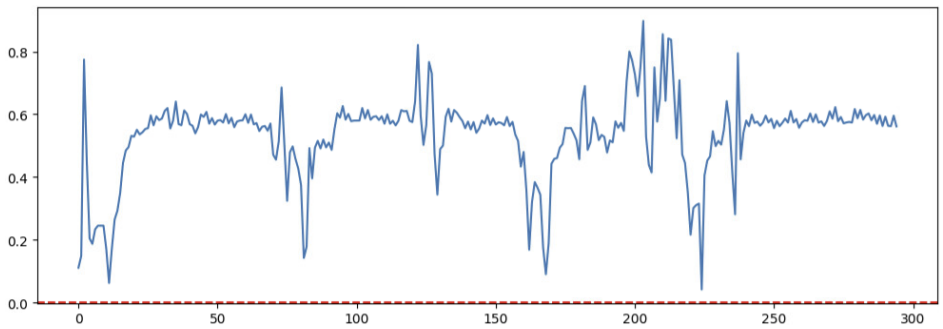



Figura 3.9: Série GR aumentada com *jitter* modificado.

**Scaling Orientado por Dados:** A série é reconstruída multiplicando-se a tendência central pelos coeficientes de escala rotacionados. Isso altera estocasticamente a dispersão local preservando a coerência das transições físicas:

$$y_{\text{scale_aug}}(t) = y_{\text{smoothed}}(t) \times C_{\text{shifted}}(t)$$

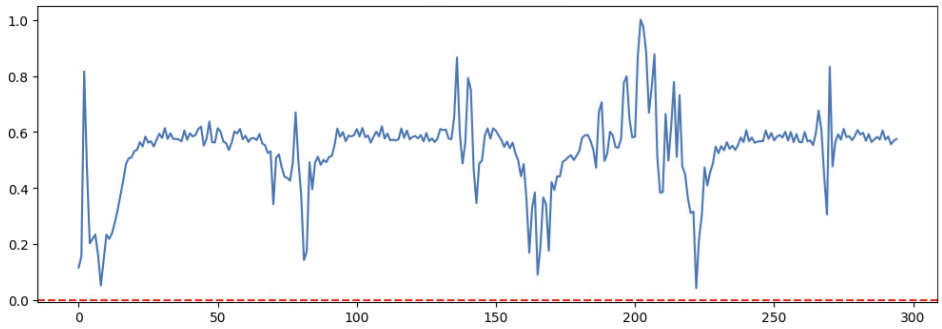



Figura 3.10: Série GR aumentada com *scale* modificado.

**Time Warping Segmentado:** A deformação temporal é aplicada diretamente sobre a geometria de cada segmento da série suavizada baseando-se no fator $\alpha$ sorteado para aquele intervalo:

- Se $\alpha > 1$: O segmento sofre uma expansão física através de interpolação linear, emulando um aumento local de espessura da camada geológica.
- Se $\alpha < 1$: O segmento é comprimido. Para manter o alinhamento de contorno sem criar descontinuidades nas bordas do segmento, realiza-se o corte e redimensionamento de uma subjanela contínua do trecho original.

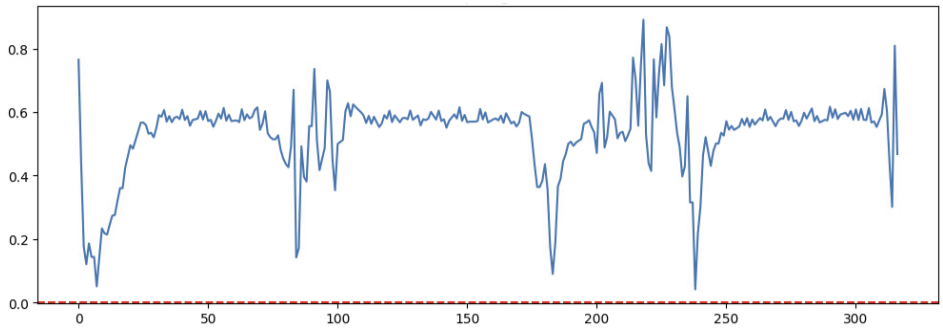



Figura 3.11: Série GR aumentada com *warp* modificado.

# Capítulo 4

# Validação e Avaliação dos Dados

A validação de técnicas de *data augmentation* em domínios altamente especializados, como a petrofísica e a geologia de subsuperfície, impõe desafios que transcendem as métricas estatísticas tradicionais de aprendizado de máquina. Enquanto em cenários computacionais distintos, como visão computacional ou reconhecimento de voz, a plausibilidade visual ou a variação de entropia podem ser suficientes, o aumento de dados de perfis de poço exige a preservação estrita de propriedades físicas e correlações multivariadas fundamentais. Para validar o pipeline proposto, implementou-se um protocolo de avaliação duplo, integrando uma abordagem analítico-petrofísica quantitativa e uma auditoria empírica qualitativa com especialistas do domínio.

## 4.1 Inspeção Petrofísica e Coerência Multivariada

A primeira linha de validação consistiu em verificar se as transformações estocásticas locais preservaram as leis físicas intrínsecas que regem os sistemas rocha-fluido. Em perfis de poço, as diferentes curvas, tais como Raios Gama, Densidade Bulk, Porosidade Neutrônica e Resistividade, não são independentes; elas respondem de forma acoplada à mesma litologia, geometria de poros e fluidos saturantes. Métodos clássicos de aumento de dados, ao tratarem as curvas como canais isolados ou ao aplicarem ruído e fatores de escala globais e arbitrários, destroem sistematicamente essas interdependências, gerando assinaturas petrofísicas impossíveis.

Para mitigar esse risco e comprovar a robustez do pipeline orientado por dados, foram geradas matrizes de correlação cruzada e gráficos de dispersão comparando as populações de dados originais e sintéticos. Focou-se especialmente nas relações clássicas da física de rochas, como o domínio de densidade versus porosidade. A análise estatística demonstrou categoricamente que as dispersões geradas pelas três vertentes do pipeline proposto se sobrepuseram de forma análoga às tendências da rocha real.

Como as perturbações de *Jitter* e *Scaling* propostas utilizam resíduos e coeficientes extraídos de forma condicionada à própria linha de base suavizada da rocha (e aplicados via deslocamento circular por segmentos geologicamente coerentes), a textura estrutural e o acoplamento físico original foram protegidos. Os coeficientes de correlação de Pearson entre as variáveis mantiveram os mesmos sinais e magnitudes relativos (Figuras 4.1, 4.2 e 4.3), demonstrando que o aumento introduziu variabilidade sem corromper a assinatura geológica de subsuperfície.

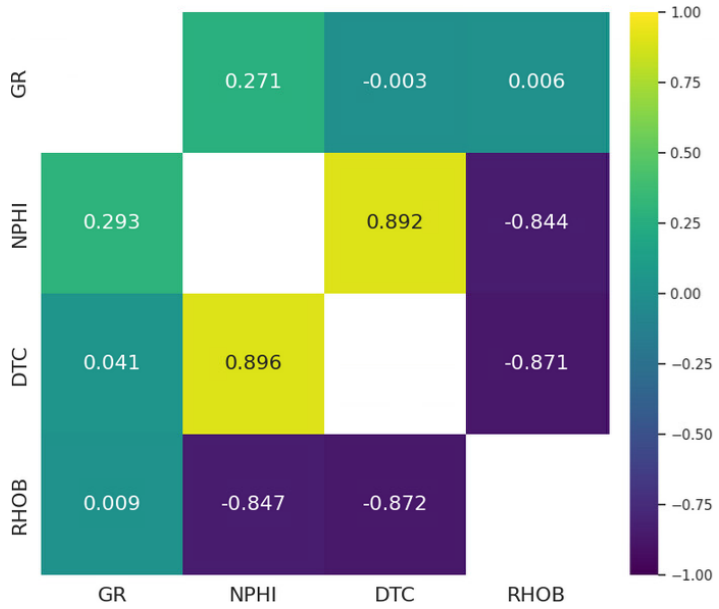



Figura 4.1: Matriz de Correlação Combinada entre perfis originais e sintéticos utilizando o método *Shift Augmentation*.

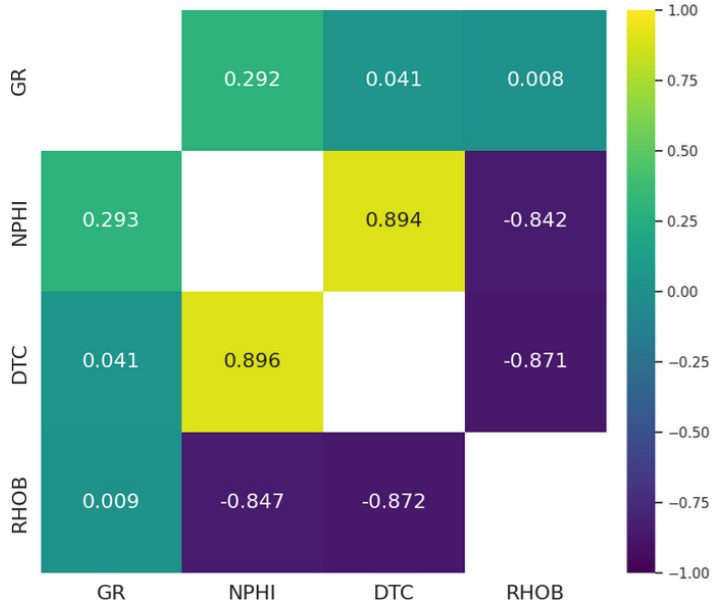



Figura 4.2: Matriz de Correlação Combinada entre perfis originais e sintéticos utilizando o método *Scale Augmentation*.

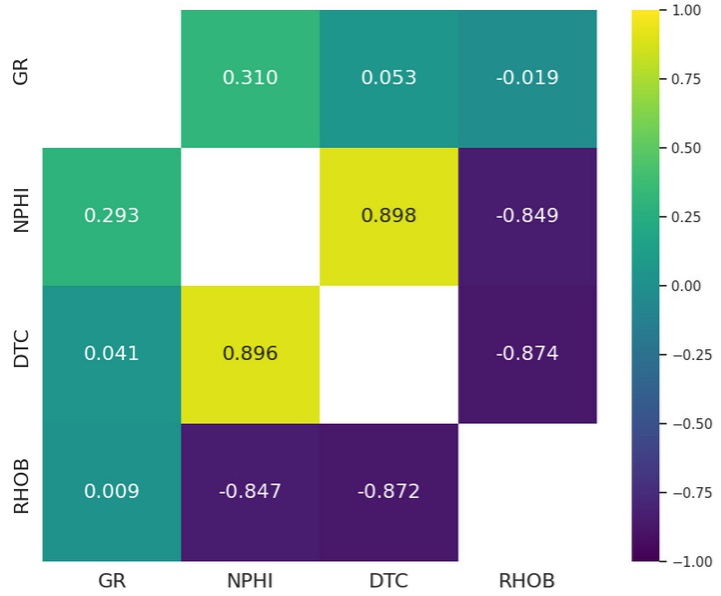



Figura 4.3: Matriz de Correlação Combinada entre perfis originais e sintéticos utilizando o método *Warp Augmentation*.

## 4.2 Teste Duplo-Cego com Especialistas

Para atestar o realismo e a fidelidade morfológica das amostras geradas face à percepção humana altamente treinada, instanciou-se um teste empírico baseado no protocolo duplo-cego. Submeteu-se um conjunto de 100 fragmentos de perfis de poços à auditoria visual de uma banca composta por geocientistas, geólogos e geofísicos sêniores.

O experimento foi conduzido por meio de uma interface computacional customizada que exibia, aleatoriamente, pares de perfis logados. Cada par continha uma assinatura real e uma sintética derivada das metodologias propostas de *Scaling*, *Shifting* (Jitter) e *Warping*. Os especialistas eram desafiados a identificar qual perfil era o verdadeiro. Crucialmente, para mitigar o viés de adivinhação forçada e isolar o ruído estatístico de escolhas puramente aleatórias, introduziu-se a opção de resposta "Não Sei / Inseguro". Os resultados consolidados dessa avaliação são apresentados na Tabela 4.1.

Tabela 4.1: Resultados da avaliação cega dos perfis de poço aumentados por especialistas do domínio. A acurácia líquida é calculada excluindo-se as abstenções (respostas "Não Sei").

| | **Nº de Amostras** | **Correto** | **Incorreto** | **Não Sei** | **% Não Sei** | **Acurácia** |
|---|---|---|---|---|---|---|
| Scale | 33 | 9 | 7 | 17 | 51,52% | 56,25% |
| Shift | 36 | 6 | 11 | 19 | 52,78% | 35,29% |
| Warp | 31 | 8 | 8 | 15 | 48,39% | 50,00% |
| **Total** | **100** | **23** | **26** | **51** | **51,00%** | **46,94%** |

A análise quantitativa dos laudos emitidos revela dois indicadores fundamentais do sucesso da metodologia em contraste com os métodos de aumento clássicos:

1. **Taxa Elevada de Incerteza (Abstenção):** Do volume total de avaliações, 51,00% das respostas concentraram-se na categoria "Não Sei / Inseguro". Esse índice demonstra que, em mais da metade dos cenários experimentais, profissionais experientes consideraram ambos os sinais igualmente verossímeis, sendo incapazes de detectar artefatos digitais, descontinuidades matemáticas ou anomalias estruturais que denunciassem a origem sintética do dado.

2. **Acurácia Equivalente ao Acaso (Convergência Estocástica):** Ao isolar os casos em que os especialistas emitiram um veredito técnico assertivo (excluindo-se as abstenções), a acurácia líquida global calculada foi de 46,94%. Do ponto de vista da teoria estatística, uma acurácia em torno de 50% em um teste de escolha binária (falso *versus* verdadeiro) é indistingüível do comportamento de um classificador randômico puro (chute aleatório). Destaca-se o comportamento do método *Shift*, no qual os especialistas erraram mais do que acertaram (acurácia líquida de 35,29%), confundindo o ruído real deslocado circularmente com a assinatura instrumental legítima do poço.

## 4.3 Discussão e Diferenciais Frente aos Métodos Tradicionais

Os resultados obtidos confirmam a hipótese de que o pipeline proposto resolve os gargalos críticos que historicamente invalidam o uso de *data augmentation* clássico em dados geológicos. Abordagens tradicionais de *Jittering* Gaussiano ou *Scaling* multiplicativo global falham em testes duplo-cegos porque alteram as frequências do sinal de forma geometricamente uniforme ou violam a escala de magnitudes das litologias. Um geólogo facilmente identificaria um perfil modificado por *scaling* clássico ao notar que valores de folhelho e arenito sofreram dispersões desproporcionais, quebrando o padrão morfológico esperado para a bacia sedimentar.

O diferencial do método proposto reside na sua natureza **orientada por dados e contextualizada por segmento**. Ao extrair os resíduos e os coeficientes ponto a ponto em relação à mediana e realizar o deslocamento circular confinado aos limites dos *breakpoints*, o pipeline garante que:

- O ruído adicionado possui exatamente o mesmo espectro de frequência e a mesma variância do instrumento real daquele poço específico, eliminando a necessidade de parametrização manual de distribuições teóricas;

- As variações de escala preservam o gradiente e a proporção entre as diferentes faixas de valores (altos e baixos), respeitando os limites físicos das rochas;

- As deformações temporais (*warping*) aplicadas por segmento preservam a integridade dos contatos litológicos abruptos identificados pelo algoritmo de detecção de mudanças, emulando perfeitamente variações geológicas na espessura de deposição das camadas sem criar assinaturas artificiais em zonas de transição estáveis.

O fato de a acurácia dos especialistas ter convergido para a probabilidade de um acerto ao acaso prova que as curvas sintéticas geradas são indistingüíveis das originais

sob critérios visuais e petrofísicos rígidos, validando o pipeline como uma ferramenta confiável para expansão de bases de dados geológicos destinadas a modelos de Deep Learning.

# Capítulo 5

# Conclusões

Este trabalho apresentou uma metodologia inovadora de aumento de dados especificamente projetada para séries temporais de perfis de poço, superando os gargalos críticos de generalização e as distorções físicas associadas aos métodos clássicos do domínio. Diante da histórica escassez e do alto custo de obtenção de dados geológicos em subsuperfície, a pesquisa consolida uma estratégia robusta para viabilizar o treinamento eficiente de arquiteturas de aprendizado profundo (*Deep Learning*) sem comprometer a fidelidade física do sinal subjacente.

A principal contribuição conceitual da abordagem proposta reside na sua natureza intrinsecamente livre de parametrização empírica (*tuning-free*). Em contraste com as técnicas tradicionais da literatura, que demandam o ajuste manual e iterativo de distribuições de ruído teóricas ou fatores de ganho globais, o pipeline desenvolvido extrai a estrutura de variabilidade estatística diretamente dos dados brutos de cada poço. Ao isolar os componentes de alta frequência (resíduos aditivos) e flutuações de amplitude (coeficientes multiplicativos) em relação a uma linha de base obtida via filtragem por mediana, o método elimina a arbitrariedade humana do processo de síntese de dados.

A arquitetura do pipeline introduz avanços significativos em três pilares fundamentais:

- **Invariância de Escala Vertical:** A segmentação adaptativa baseada em algoritmos de detecção de *breakpoints* proporcionais ao comprimento do perfil garante que a amostragem das janelas de perturbação seja invariante à profundidade do poço investigado, estabilizando a resolução matemática em grandes bases de dados multi-poço;

- **Preservação de Correlação Multivariada:** Ao condicionar os operadores de *Jitter* e *Scaling* à morfologia original das camadas, a metodologia impede a quebra de acoplamento físico entre múltiplas curvas síncronas (e.g., Raios Gama, Densidade e Porosidade), salvaguardando as assinaturas petrofísicas legítimas do sistema rocha-fluido;

- **Coerência Textural e Estratigráfica:** A substituição do embaralhamento estocástico convencional pelo Deslocamento Circular confinado aos limites de cada segmento preserva rigorosamente a correlação serial, a continuidade local e a vi-

zinhança dos pontos, emulando a textura do ruído real de aquisição e respeitando a deposição geológica moldada ao longo do tempo geológico.

A validação experimental do pipeline foi submetida a um rigoroso protocolo duplo-cego com especialistas seniores de geociências, engenharia de petróleo e geofísica. Os resultados quantitativos demonstraram que os profissionais foram incapazes de isolar estatisticamente os perfis sintéticos das assinaturas naturais. Com uma taxa de abstenção por incerteza visual de 51,00% e uma acurácia líquida global de 46,94%, valor estatisticamente equivalente ao comportamento de um classificador aleatório puro (*chute ao acaso*), comprovando o realismo morfológico e geológico das amostras geradas. Adicionalmente, as análises de gráficos cruzados petrofísicos ratificaram a manutenção das distribuições multimodais originais nas amostras expandidas.

Conclui-se que a metodologia proposta cumpre com êxito os objetivos estipulados, preenchendo uma lacuna crítica na intersecção entre a Inteligência Artificial e as Geociências. Ao automatizar a geração de amostras realistas de perfis de poço de forma robusta, escalável e fiel às leis físicas da natureza, este trabalho fornece um alicerce para expandir conjuntos de dados e potencializar a capacidade de generalização de modelos preditivos profundos aplicados à caracterização de reservatórios e modelagem de subsuperfície.

# Referências Bibliográficas

1 ANNAKI, I.; RAHMOUNE, M.; BOURHALEB, M. Overview of data augmentation techniques in time series analysis. *International Journal of Advanced Computer Science and Applications*, The Science and Information Organization, v. 15, n. 1, 2024. Disponível em: <http://dx.doi.org/10.14569/IJACSA.2024.01501118>.

2 SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data*, v. 6, p. 1–48, 2019. Disponível em: <https://api.semanticscholar.org/CorpusID:195811894>.

3 IWANA, B. K.; UCHIDA, S. An empirical survey of data augmentation for time series classification with neural networks. *PLOS ONE*, Public Library of Science, v. 16, n. 7, p. 1–32, 07 2021. Disponível em: <https://doi.org/10.1371/journal.pone.0254841>.

4 IGLESIAS, G. et al. Data augmentation techniques in time series domain: a survey and taxonomy. *Neural Computing and Applications*, v. 35, p. 10123–10145, 2023. Disponível em: <https://doi.org/10.1007/s00521-023-08459-3>.

5 LIU, B.; ZHANG, Z.; CUI, R. yi. Efficient time series augmentation methods. *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, p. 1004–1009, 2020. Disponível em: <https://api.semanticscholar.org/CorpusID:227219603>.