Enzo Bustos Da Silva

Irregular Multivariate Time Series e o *Paranaguá Port Meteorological and Oceanographic Dataset*

São Paulo, SP

2024

Enzo Bustos Da Silva

Irregular Multivariate Time Series e o *Paranaguá Port Meteorological and Oceanographic Dataset*

Trabalho de Conclusão de Curso com objetivo de desenvolver e apresentar um conjunto de dados inédito, o *Paranaguá Port Meteorological and Oceanographic Dataset* (P^{2}MOD), explorando as abordagens modernas em problemas de Séries Temporais Multivariadas Irregulares, além de propor e validar soluções autorais para previsão de grandezas meteorológicas e oceanográficas. Apresentado para o Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo para obtenção do título de Engenheiro Elétrico - ênfase em Computação.

Área de Concentração:
Inteligência Artificial
Aprendizado de Máquina
Banco de Dados

Universidade de São Paulo – USP

Escola Politécnica

Departamento de Engenharia de Computação e Sistemas Digitais (PCS)

Orientador: Professora Doutora Anna Helena Reali Costa

Coorientador: Doutor Marcel Rodrigues de Barros

São Paulo, SP

2024

Catalogação-na-publicação

Bustos, Enzo
Séries Temporais Multivariadas Irregulares Aplicadas à Previsão de Grandezas Oceânicas no Porto de Paranaguá / E. Bustos -- São Paulo, 2024.
64 p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Séries Temporais Multivariadas Irregulares 2.Inteligência Artificial 3.Aprendizado de Máquina 4.Tarefa de Previsão 5.Grandezas Meteorológicas e Oceanográficas I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Dedico este meu Trabalho de Conclusão aos incansáveis sonhadores,
comumente chamados de Pesquisadores.

Que com seu espírito ardente pelo saber;
Com o coração apaixonado no desconhecido;
Com sua alma inquieta em romper os limites;
Com seu brilho no olhar pelo progresso e avanço;
Com sua curiosidade incessante em busca da verdade;
E com seu compromisso inabalável pelo Método Científico:

Abrem as portas das possibilidades,
por meio de toda sua capacidade e ideias,
para construir o futuro para todas as pessoas.

Agradecimentos

Agradeço imensamente à Professora Doutora Anna Helena Reali Costa, cuja orientação foi indispensável para meu desenvolvimento como pesquisador. Suas contribuições acadêmicas, cuidado e atenção foram determinantes para o sucesso deste trabalho. Estendo minha gratidão ao Doutor Marcel Rodrigues de Barros, co-orientador deste TCC, por sua dedicação, apoio e vasto conhecimento técnico, essenciais em cada etapa deste projeto. Ao Mestre Thomas Palmeira Ferraz, meu amigo e co-orientador nas minhas ICs, agradeço por ser uma meta profissional, um modelo de sucesso que admiro e me inspiro.

Aos funcionários que tornaram esse trabalho possível, meu sincero agradecimento. Destaco a Secretária Viviane Leite, sua eficiência foi essencial e sua ajuda para sanar os desafios ao longo da graduação foi crucial, uma pessoa extremamente querida, de bom coração e paciência sem iguais. Além disso, agradeço também à Administração e Autoridade Portuária de Paranaguá pelo fornecimento dos dados essenciais utilizados neste estudo.

À minha família, minha base, expresso toda minha gratidão pelo apoio incondicional, pelos valores transmitidos e pelos incentivos que me conduziram até aqui. Um carinho especial às minhas madrinhas, Flávia Bustos e Adriana Bustos, pelo amor e suporte constantes, e ao meu pai, Rogério Salviano da Silva, meu primeiro e maior exemplo na docência, cuja paixão pela educação sempre me inspirou profundamente.

Aos amigos próximos, deixo meu forte abraço e minha gratidão por estarem sempre comigo, especialmente nos momentos difíceis da graduação, ajudando a transformar dúvidas e desmotivação em boas risadas, leveza e alegria. Aos amigos: Ana Gabriela Lemes De Araújo, Marcos Vinicius Ramos Mulari, Lucas Yudi Leonardi e Zhao Yi Kuan, vocês são pessoas verdadeiramente iluminadas, que tornam meus dias e minha vida, melhores.

Por fim, um agradecimento imenso à minha parceira na vida, Maria Clara Cyrino De Souza. Você esteve ao meu lado em todas as minhas escolhas, sempre com um apoio incondicional, sendo uma fonte de motivação e inspiração que é um verdadeiro afago na minha alma. Seu amor, paciência e encorajamento constantes foram a força motriz, me impulsionando a nunca desistir e seguir de cabeça erguida por cada desafio que apareceu, e ainda aparecerá.

A cada um de vocês, meu mais sincero obrigado e minha gratidão profunda. Este trabalho é mais do que um reflexo do meu esforço; ele carrega o apoio, o carinho e a força que recebi de cada um de vocês ao longo desta jornada.

"The important thing is not to stop questioning. Curiosity has its own reason for existing."

— Albert Einstein

Resumo

Séries Temporais Multivariadas Irregulares (IMTS) representam um desafio significativo para tarefas de previsão em cenários reais, devido às irregularidades presentes nos dados coletados por sensores ambientais. Este trabalho propõe o *Paranaguá Port Meteorological and Oceanographic Dataset* (P^2MOD), um conjunto de dados inédito que captura características de séries temporais irregulares obtidas no Porto de Paranaguá, o maior porto graneleiro da América Latina. Diferente de abordagens tradicionais que dependem de regularizações complexas para tratar irregularidades, este estudo explora diretamente essas características, argumentando que modelos capazes de lidar diretamente com as irregularidades mostram-se superiores aos que regularizam essas imperfeições para ajustar os modelos às técnicas tradicionais de Processamento de Sinais. Foram implementados dois modelos para servir como caso de uso baseados em Redes Neurais, o primeiro é uma Gated Recurrent Units (GRU) implementada de forma *standard*, sendo um modelo mais simples e sem grandes capacidades de generalizações em Séries Temporais e o segundo é um modelo de Gap-Ahead integrado com a técnica de *time encoding*. A análise experimental demonstra que modelos que lidam diretamente com irregularidades superam aqueles que utilizam regularizações, tanto em precisão preditiva quanto em robustez, com a métrica de *Index of Agreement*. O P^2MOD é apresentado como um recurso público para fomentar pesquisas futuras, com potencial de promover avanços significativos na modelagem de dados meteorológicos e oceanográficos em contextos reais.

Palavras-chave: Séries Temporais Multivariadas Irregulares, Aprendizado de Máquina, Redes Neurais, Porto de Paranaguá, Sensoriamento Ambiental, *Dataset*.

Abstract

Irregular Multivariate Time Series (IMTS) pose a significant challenge for forecasting tasks in real-world scenarios due to the irregularities present in data collected by environmental sensors. This work proposes the *Paranaguá Port Meteorological and Oceanographic Dataset* (P^2MOD), an unprecedented dataset capturing the characteristics of irregular time series obtained from the Port of Paranaguá, the largest grain port in Latin America. Unlike traditional approaches that rely on complex regularizations to address irregularities, this study directly explores these characteristics, arguing that models capable of directly handling irregularities outperform those that regularize such imperfections to fit traditional Signal Processing techniques. Two models were implemented as use cases based on Neural Networks: the first is a *standard* Gated Recurrent Unit (GRU), a simpler model with limited generalization capabilities for Time Series, and the second is a Gap-Ahead model integrated with the *time encoding* technique. Experimental analysis demonstrates that models addressing irregularities directly outperform those employing regularization, both in predictive accuracy and robustness, as evaluated using the *Index of Agreement* metric. P^2MOD is presented as a public resource to foster future research, with the potential to significantly advance the modeling of meteorological and oceanographic data in real-world contexts.

Keywords: Irregular Multivariate Time Series, Machine Learning, Neural Networks, Port of Paranaguá, Environmental Sensing, *Dataset*.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

IMTS	*Irregular Multivariated Time Series*
TCC	Trabalho de Conclusão de Curso
ONU	Organização das Nações Unidas
P^2MOD	*Paranaguá Port Meteorological and Oceanographic Dataset*
AI/ML	*Artificial Intelligence and Machine Learning*
ANN	*Artificial Neural Networks*
DFT	Transformada de Fourier Discreta
iid	Independentes e Identicamente Distribuídas

Lista de símbolos

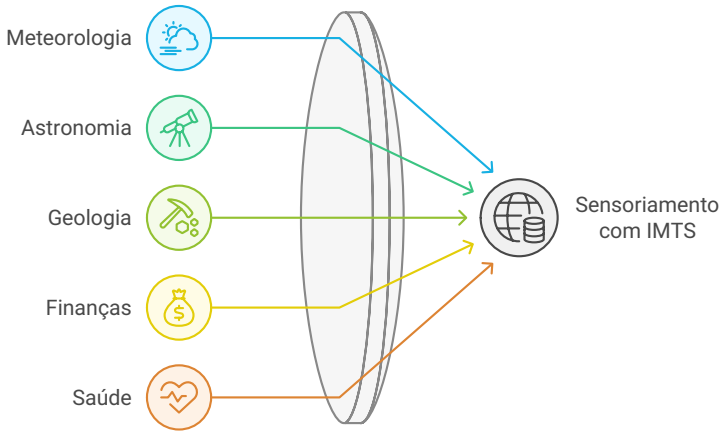
$\mathcal{D}$	Conjunto de Dados, *Dataset*
$\mathbf{X}_i$	Vetor de observações
y_i	Rótulo associado a uma observação
$\mathcal{L}$	*Loss Functions*, Função de Perdas
$x[n]$	Sinal amostrado no instante n

Sumário

1 Introdução

Previsões baseadas em *Irregular Multivariate Time Series* (IMTS, do inglês, Séries Temporais Multivariadas Irregulares) possuem ampla aplicação em diversas áreas de pesquisa e prática, abrangendo tópicos como Meteorologia, Astronomia, Geologia, Finanças e Saúde, entre outros (SCHWARZACHER, 1964; MUDELSEE, 2002; VIO et al., 2013; SEZER et al., 2020; SCOTT et al., 2022), sob a ótica de sensoriamento, ao ter múltiplas instâncias que não estão necessariamente sincronizadas, muitas vezes caímos em um problema que envolve as IMTS (ilustrado na Figura 1). No entanto, a literatura carece de conjuntos de dados públicos e metodologias específicas que capturem e tratem adequadamente as complexidades inerentes a essas irregularidades. Frequentemente, os trabalhos existentes baseiam-se em hipóteses simplificadoras e técnicas de regularização, deixando lacunas na representação dos problemas reais enfrentados por esses dados.

Figura 1 – Perspectiva Multidisciplinar das IMTS



A ilustração acima mostra como diversos tipos de problemas podem ser modelados como uma IMTS, para casos em que há múltiplos sensores ou irregularidades nos dados.

Imagem criada autorialmente.

Problemas que podem ser formalizados pelas IMTS são comumente encontrados em diversos sistemas de engenharia, especialmente no contexto de sensoriamento ambiental (ZHANG et al., 2013; MARTIN, 2014; VOOGT et al., 2003; SMITH et al., 2019). Nesse cenário, o monitoramento oceânico ganha especial relevância para questões climáticas e meteorológicas, uma importância reforçada pela Década dos Oceanos, proclamada pela ONU [1], dessa forma, o *dataset* apresentado neste trabalho assume uma relevância adicional por fornecer um conjunto de dados inédito sobre o Porto de Paranaguá, o maior porto graneleiro da América Latina (IBGE. . . , 2017), que reflete tanto as irregularidades inerentes

[1] Disponível em: ⟨https://oceandecade.org/⟩ (Promulgada em 07/12/2017)

ao sensoriamento em ambientes reais quanto sua relevância para estudos climáticos, oceanográficos e meteorológicos.

No campo de Aprendizado de Máquina e Inteligência Artificial, a disponibilidade de dados representativos para a tarefa de *IMTS-forecasting* é extremamente escassa. A maioria dos modelos e abordagens consagradas assumem uma taxa de amostragem regular dos dados (LIU et al., 2024), o que não corresponde à realidade dos sistemas reais que contém irregularidades. O *Paranaguá Port Meteorological and Oceanographic Dataset* (P^2MOD), apresentado neste trabalho, visa preencher essa lacuna, contribuindo para o desenvolvimento de métodos e modelos que considerem a natureza irregular de séries temporais reais, contribuindo para a criação de modelos preditivos mais robustos e generalizáveis para o contexto oceanográfico e meteorológico.

Adicionalmente, o conjunto de dados proposto se destaca por abordar um problema raramente explorado na literatura: a *IMTS-forecasting*. Esse *dataset* oferece dados de sensoriamento que contém as irregularidades comuns, as quais são frequentemente ignoradas ou tratadas de forma simplificada. Grande parte dos estudos e *benchmarks* sobre séries temporais recorre a dados organizados em estruturas tabulares regulares (*Grid-Like Structures*) (HEBRAIL, 2006; VITO, 2008; REYES-ORTIZ, 2013) ou submete os dados a processos extensivos de regularização antes de publicação (REISS et al., 2012; ZHOU et al., 2021). Como ilustração dessa lacuna, apenas dois artigos publicados nos anais da 38^a *AAAI Conference on Artificial Intelligence* [2] (YALAVARTHI et al., 2024; XIAO et al., 2024) trataram explicitamente das complexidades e desafios inerentes às IMTS, evidenciando a escassez de pesquisas dedicadas ao tema.

1.1 Motivação

O interesse em explorar esta temática é impulsionado por uma motivação pessoal para abordar questões ambientais e de sensoriamento *in loco*, áreas nas quais as IMTS aparecem com frequência. A experiência prática em adequar as irregularidades desses dados aos métodos e abordagens tradicionais não apenas reforça o desejo de identificar soluções mais apropriadas, mas também visa desenvolver modelos mais abrangentes que integrem essas análises eficazmente.

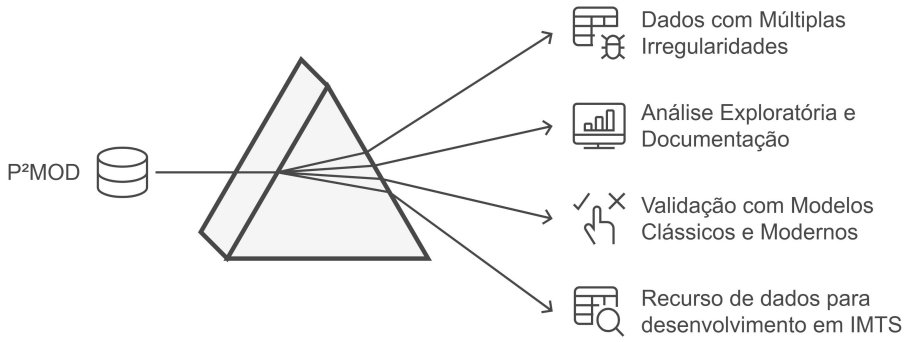
1.2 Justificativa

A presente pesquisa é motivada pela escassez de abordagens e métodos adequados para lidar com as irregularidades em séries temporais, um problema ainda sub-representado na literatura acadêmica. Embora alguns *datasets* integrem certas irregularidades, são raros

[2] Disponível em: ⟨https://ojs.aaai.org/index.php/AAAI/issue/archive⟩ (Acessado em 05/09/2024)

aqueles que representam de maneira abrangente as falhas comuns apresentadas por dados reais de engenharia, os quais reúnem diversas irregularidades em uma única base. Assim, a criação e disponibilização do *dataset* P²MOD visa preencher essa lacuna teórica (ilustrado na Figura 2), ao oferecer uma solução unificada que contempla, em um só banco de dados, as múltiplas irregularidades que podem estar presentes em dados reais.

Figura 2 – Lacunas Melhoradas pelo P²MOD



A ilustração acima mostra como diversas lacunas na literatura podem ser melhoradas pelo uso e divulgação de um dataset com irregularidades.

Imagem criada autorialmente.

O P²MOD se distingue por consolidar múltiplas formas de irregularidades temporais em um único banco de dados, uma abordagem rara entre as bases de dados existentes. Enquanto *datasets* convencionais abordam irregularidades de forma isolada ou limitada, o P²MOD oferece uma visão integrada, permitindo análises mais abrangentes e robustas do desempenho de algoritmos de *forecasting*. Ademais, este *dataset* reflete cenários reais de engenharia, pois se origina de um caso concreto, proporcionando uma representação fiel dos problemas enfrentados em aplicações como sensoriamento ambiental, climático e meteorológico.

Essas características tornam o conjunto de dados apresentado neste TCC um recurso valioso tanto para a comunidade acadêmica quanto para profissionais, ao viabilizar a experimentação e validação de técnicas de AI/ML voltadas diretamente ao tratamento das IMTS, sem a necessidade de pré-processamento por métodos convencionais de Processamento de Sinais ou de estruturação em *grid-like structures*. Neste contexto, o trabalho não só preenche uma lacuna significativa na literatura, como também disponibiliza uma base para a evolução das técnicas de análise de dados temporais, essencial para o desenvolvimento de abordagens generalistas e robustas que usam diretamente as técnicas mais avançadas de AI/ML disponíveis.

1.3 Objetivos

Considerando a complexidade inerente ao tratamento de irregularidades em IMTS e a limitada disponibilidade de *datasets* públicos que abordem essa questão sem simplificações, este Trabalho de Conclusão de Curso visa principalmente criar e disponibilizar um novo conjunto de dados que capture as especificidades das IMTS, com foco nas operações do Porto de Paranaguá, no Paraná. Esta pesquisa visa contribuir para a comunidade acadêmica e profissional em IMTS, atendendo aos seguintes objetivos específicos (ilustrados na Figura 3):

i. **Disponibilização do P²MOD**: Preparar, documentar e disponibilizar um *dataset* detalhado do Porto de Paranaguá, que reflita as irregularidades típicas de séries temporais reais, proporcionando um recurso valioso para pesquisas futuras.

ii. **Exploração e Análise do *Dataset***: Conduzir uma análise exploratória abrangente para identificar e caracterizar as principais irregularidades presentes no conjunto de dados oceanográficos e marítimos.

iii. **Validação e Caso de Uso do *Dataset***: Demonstrar a aplicabilidade do *dataset* em cenários práticos, empregando modelos clássicos e contemporâneos de previsão (*forecasting*) para ilustrar os desafios e o potencial do conjunto de dados como uma ferramenta preliminar de validação para pesquisas futuras.

iv. **Contribuição para a Pesquisa em IMTS**: Facilitar o avanço da pesquisa em IMTS ao fornecer um recurso de dados que permite testar abordagens diretamente voltadas ao tratamento de irregularidades, promovendo o desenvolvimento de métodos mais robustos e eficazes.

Figura 3 – Objetivos da Monografia

A imagem ilustra os objetivos da Monografia de disponibilizar um *dataset*, fornecendo métodos de análise e casos de uso para contribuir para pesquisa em IMTS.

Imagem criada autorialmente.

1.4 Organização do Trabalho

Este trabalho foi desenvolvido para oferecer uma compreensão clara e progressiva dos temas tratados. Cada capítulo foi cuidadosamente estruturado para apresentar, de maneira lógica e sequencial, uma visão abrangente e coesa do estudo (ilustrado na Figura 4). A seguir, é apresentada a organização dos capítulos:

→ **Capítulo 2: Fundamentação Teórica**
Este capítulo apresenta os conceitos e teorias que fundamentam a pesquisa, abordando os principais paradigmas e estruturas teóricas essenciais para a compreensão do problema em análise.

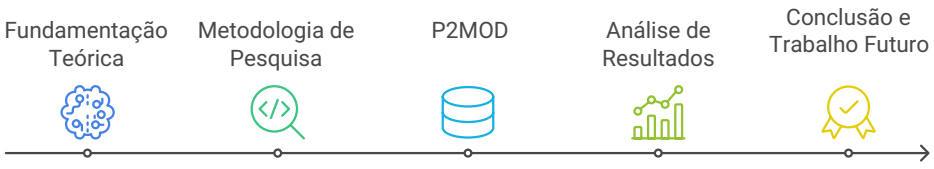
→ **Capítulo 3: Metodologia de Pesquisa**
Descreve a metodologia adotada para a execução do estudo, detalhando os métodos e procedimentos seguidos, incluindo a abordagem metodológica e as técnicas empregadas para alcançar os objetivos propostos.

→ **Capítulo 4: *Paranaguá Port Meteorological and Oceanographic Dataset***
Apresenta o conjunto de dados meteorológicos e oceanográficos do Porto de Paranaguá. São descritas a origem, natureza e características dos dados que influenciam diretamente a análise realizada.

→ **Capítulo 5: Análise dos Resultados**
Analisa os resultados obtidos a partir dos dados e dos métodos empregados, os interpretando à luz dos objetivos da pesquisa e das teorias discutidas.

→ **Capítulo ??: Conclusão e Trabalhos Futuros**
Sintetiza as principais contribuições e limitações do estudo, além de sugerir direções para pesquisas futuras.

Figura 4 – Organização do Trabalho



A imagem ilustra, de maneira sequencial, a organização da Monografia, apresentando a estrutura dos capítulos e a progressão do estudo.

Imagem criada autorialmente.

2 Fundamentação Teórica

Em problemas de sensoriamento ambiental, é frequente a presença de sistemas compostos por múltiplos sensores que operam sem uma garantia de sincronia na aquisição de dados. Essa falta de padronização na aquisição desses sinais resulta numa coleta que contém intervalos irregulares e eventos parcialmente observados, configurando uma IMTS. No entanto, a literatura atual ainda carece de uma uniformidade nas definições e abordagens aplicadas a esse tipo de problema, evidenciando uma lacuna teórica e metodológica. Este capítulo visa fornecer a fundamentação teórica necessária para a análise do P^{2}MOD, abordando os principais conceitos que serão utilizados no Capítulo 3. Entre esses conceitos, destacam-se as metodologias de análise de dados, técnicas de *forecasting*, e os avanços recentes em IA/ML, com ênfase no uso de ANN (do inglês, *Artificial Neural Networks*), ferramentas essenciais para lidar com as complexidades dos dados presentes no P^{2}MOD.

A Seção 2.1 introduz os conceitos fundamentais, abordando técnicas de aprendizado em Inteligência Artificial, o algoritmo de *Gradient Descent*, as Séries Temporais regulares e as abordagens clássicas de Processamento de Sinais. Seguida, pela apresentação das principais formas de irregularidades nas Séries Temporais que serão tratadas neste trabalho e proposição de uma definição formal para as *IMTS*, que servirá de alicerce para o desenvolvimento do restante da Monografia. A Seção 2.2 discute modelagens clássicas para Séries Temporais, incluindo métodos de *Exponential Smoothing* e a tradicional família de modelos *ARIMA* entre outros modelos consolidados. Em contraponto, a Seção 2.3 examina abordagens mais modernas, como técnicas de *Time Encoding* e modelos recentes sendo voltados para o problema de *IMTS-forecasting*. Finalmente, a Seção 2.4 analisa outros conjuntos de dados que contêm irregularidades e destaca a inovação proposta pelo P^{2}MOD na abrangência dessas irregularidades em um único banco de dados.

2.1 Conceitos Fundamentais

Para a compreensão dos conceitos de IMTS e da tarefa de *forecasting*, é essencial estabelecer as ferramentas e definições matemáticas que servirão de base para este TCC. Esta seção cogita contextualizar os principais conceitos que sustentam a análise do P^{2}MOD, fundamentando-se em estudos que oferecem diversas abordagens e perspectivas relevantes para a compreensão dos temas tratados neste trabalho.

A seção está estruturada para oferecer ao leitor uma base sólida para o entendimento do assunto. A Subseção 2.1.1 introduz conceitos essenciais de Séries Temporais no contexto de IA e ML, discutindo o algoritmo de *Backtracking Gradient Descent*, seguido das tarefas de aprendizado e métricas para avaliação de desempenho. Na Subseção 2.1.3, são definidos

os conceitos de Séries Temporais Regulares, incluindo estruturas *grid-like* e a tarefa de *forecasting*. Em seguida, a Subseção 2.1.2 aborda métodos historicamente enraizados no processamento de séries temporais como sinais, que assumem amostragem constante. Por fim, a Subseção 2.1.4 apresenta as principais formas de irregularidades nas séries temporais, seguida de uma definição formal de IMTS, que será empregada ao longo desta Monografia.

2.1.1 Aprendizado em Inteligência Artificial

O aprendizado supervisionado é um dos paradigmas fundamentais em Inteligência Artificial e Aprendizado de Máquina, caracterizado pelo treinamento de modelos a partir de pares de dados rotulados, onde cada entrada possui uma saída correspondente esperada. Utilizando esse conjunto de dados rotulados, o modelo aprenderá o mapeamento de entradas para saídas, para generalizar e realizar previsões precisas sobre novos dados. Esse tipo de aprendizado é amplamente aplicado em tarefas de classificação, regressão e previsão, sendo essencial para diversas aplicações práticas (BARBER, 2012; MOHRI et al., 2012; BISHOP et al., 2006).

Neste contexto, considera-se um conjunto de dados $\mathcal{D}$ composto por pares de observações independentes e identicamente distribuídas (iid). Formalmente, este conjunto é representado como $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, onde N é o tamanho do conjunto de dados. Cada vetor $\mathbf{x}_i$ representa uma observação descrita por um conjunto de atributos, enquanto y_i é o rótulo ou saída esperada associada a essa observação.

Segundo os autores que serviram de base para esta Subseção ((BARBER, 2012; MOHRI et al., 2012; BISHOP et al., 2006)), objetivo do aprendizado supervisionado é encontrar uma função $g : x \to y$ que mapeie as entradas $\mathbf{x}_i$ para os rótulos y_i, minimizando o erro de predição para novas observações. Para medir o *score* do modelo, define-se uma *Loss function* (função de perdas) $\mathcal{L} : \hat{y} \times y^* \to \mathbb{R}$, onde $\hat{y}$ é o valor previsto pelo modelo e y^* é o valor verdadeiro. A função $\mathcal{L}(\hat{y},\ y^*)$ quantifica o erro da previsão do modelo, sendo usada para ajustar o modelo durante o treinamento. Desse modo, dado um novo par $(\mathbf{x}^*, y^*) \notin \mathcal{D}$, gerado pelo mesmo processo subjacente de $\mathcal{D}$, o modelo pode estimar uma saída $\hat{y}$ que se aproxime de y^* e, idealmente, este modelo produz uma saída que se equipara à saída esperada, de modo que $g(\mathbf{x}^*) = \arg\max_{\hat{y}} \mathcal{L}(\hat{y}, y^*) \implies g(\mathbf{x}^*) = \hat{y} \approx y^*$ após uma quantidade suficiente de etapas de treinamento.

Para a compreensão de $g(\mathbf{x})$, adotamos uma definição amplamente utilizada (BARBER, 2012; MOHRI et al., 2012; BISHOP et al., 2006) em abordagens iniciais fundamentadas em modelos lineares. Neste contexto, o modelo é representado por uma combinação linear dos pesos $\mathbf{w}$ aplicados a cada entrada $\mathbf{x}$, de modo que:

$$g(\mathbf{x}; \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_N x_N = \hat{y} \tag{2.1}$$

Onde $\mathbf{w}$ denota o vetor de pesos e $\hat{y}$ representa a saída prevista pelo modelo. Este modelo linear estabelece uma base fundamental para a representação de relações entre variáveis e serve como uma primeira aproximação para problemas de predição em aprendizado supervisionado. O aprendizado supervisionado tem se tornado cada vez mais eficiente e poderoso, especialmente com a disponibilidade crescente de grandes volumes de dados. Esse grande volume de informações permite que os modelos de aprendizado de máquina sejam treinados de maneira mais robusta, melhorando sua capacidade de generalização e precisão. Como afirmado por (HALEVY et al., 2009), a abundância de dados muitas vezes contribui mais para o desempenho de modelos do que algoritmos sofisticados.

Já no contexto de otimização de modelos, a escolha dos pesos $\mathbf{w}$ visa minimizar o erro obtido pela **Loss Functions** $\mathcal{L}$, essa tarefa de minimização pode ser realizada por diversos algoritmos de otimização, entre os quais o *Gradient Descent* (Descida do Gradiente) é um dos mais populares. O *Gradient Descent* (Descida do Gradiente) é um algoritmo de otimização iterativo amplamente utilizado em aprendizado de máquina para ajustar os pesos $\mathbf{w}$ visando minimizar a função de perda $\mathcal{L}$ (BARBER, 2012; BISHOP et al., 2006; MOHRI et al., 2012; SUVRIT et al., 2012). A ideia fundamental desse algoritmo consiste em calcular o gradiente da função de perda $\nabla\mathcal{L}$ em relação aos pesos $\mathbf{w}$ e, em seguida, atualizar esses pesos na direção oposta ao gradiente, resultando em uma redução do erro de previsão.

Matematicamente, partindo de um ponto inicial de pesos $\mathbf{w}_0$, a atualização dos pesos a cada iteração é dada pela expressão:

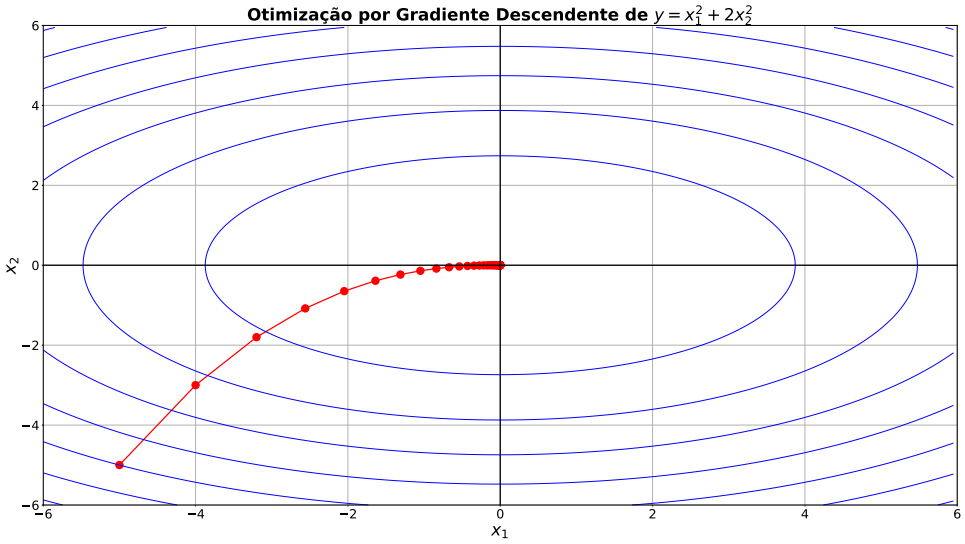
$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \nabla\mathcal{L}(\mathbf{w}_t) \tag{2.2}$$

Na Equação 2.2, η representa a taxa de aprendizado, um hiperparâmetro que determina o tamanho do passo a ser dado em cada iteração, e $\nabla\mathcal{L}(\mathbf{w}_t)$ é o gradiente da função de perda avaliado no ponto $\mathbf{w}_t$. Esse gradiente indica a direção de maior aumento da função de perda; portanto, ao nos movermos na direção oposta ao gradiente, buscamos efetivamente minimizar a função, reduzindo o erro associado à previsão do modelo. Na Figura 5, abaixo podemos ver um exemplo do algoritmo em funcionamento:

O algoritmo de *Gradient Descent* tem em vista localizar o ponto de mínimo da função de perda $\mathcal{L}(\mathbf{w})$, ou seja, o ponto de menor erro segundo nossa métrica e, ao escolhermos uma *Loss Function* que apresenta características convexas, garantimos a existência ao menos um ponto de mínimo local, que pode ser identificado quando a condição expressa na Equação 2.3 é satisfeita:

$$\nabla\mathcal{L}(\mathbf{w}) = 0 \tag{2.3}$$

Figura 5 – Algoritmo de Descida do Gradiente



O gráfico apresentado ilustra a trajetória do algoritmo de Descida do Gradiente em um espaço bidimensional para a função objetivo $y = x_1^2 + 2x_2^2$. O algoritmo começa em um ponto inicial $P_0 = (-5, -5)$ e, mediante iterações sucessivas, ajusta o ponto P em direção ao mínimo da função.

Imagem criada autorialmente.

Neste trabalho, o foco será na aplicação de redes neurais, que se destacam como uma das abordagens mais avançadas em aprendizado supervisionado, especialmente para tarefas complexas. Ao contrário de modelos lineares, as redes neurais possuem uma estrutura mais flexível, permitindo modelar relações não-lineares entre as variáveis de entrada e saída (BARBER, 2012; BISHOP et al., 2006; MOHRI et al., 2012). Elas são compostas por camadas de neurônios, onde cada neurônio realiza uma operação matemática baseada em um conjunto de entradas ponderadas. Esse cálculo pode ser descrito pela Equação 2.4:

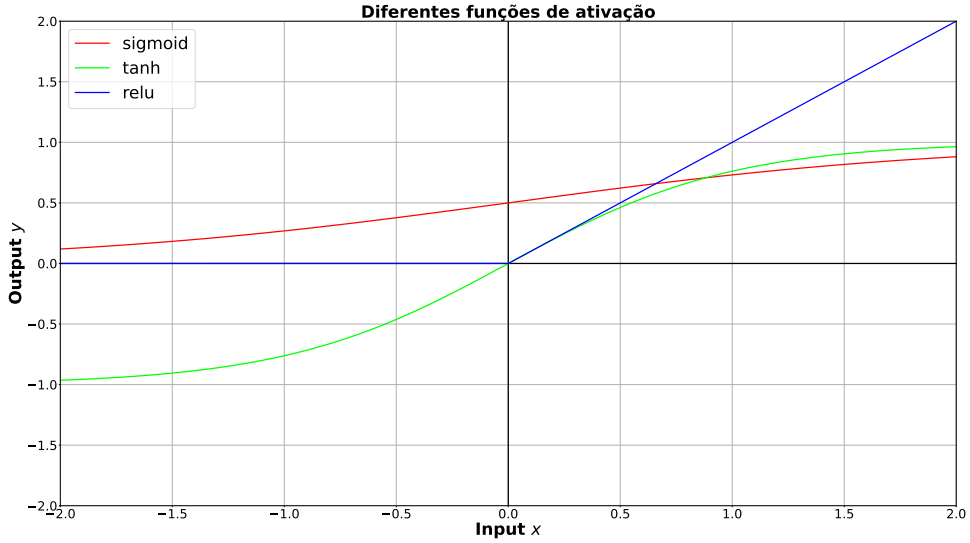
$$y = f_{act}\left(\sum_{i=1}^{N} \mathbf{w_i} \cdot \mathbf{x_i} + \mathbf{b}\right) \tag{2.4}$$

Em que f_{act} denota a função de ativação aplicada ao somatório ponderado das entradas x_i, com pesos associados w_i, e b representa o termo de viés. Alternativamente, o viés pode ser interpretado como um termo adicional $w_0 x_0$ com $x_0 = 1$. Essa estrutura é essencial para as redes neurais capturarem e aprendam representações complexas dos dados, tornando-as eficazes em problemas onde as relações entre variáveis não são triviais.

A função de ativação f_{act} introduz uma não linearidade essencial para a rede neural poder modelar relações complexas. Funções de ativação comuns incluem a sigmoide, a tangente hiperbólica e a ReLU (Todas podem ser vista nas Equações 2.5, respectivamente), as quais permitem a modelagem de comportamentos mais sofisticados em comparação a modelos puramente lineares, essas funções estão ilustradas na Figura 6.

$$\begin{aligned} \sigma(x) &= \frac{1}{1+e^{-x}} \\ \tanh(x) &= \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \\ ReLU(x) &= \max(0,\ x) \end{aligned} \tag{2.5}$$

Figura 6 – Diferentes Funções de Ativação



A imagem mostra as diferentes formas gráficas de algumas das funções de ativação mais populares no universo de AI/ML.

Imagem criada autorialmente.

O treinamento de uma rede neural envolve a otimização iterativa dos pesos w_i e dos vieses b de cada neurônio, visando minimizar a função de perda $\mathcal{L}$, neste processo, o algoritmo de retropropagação de erro, conhecido como *backpropagation*, desempenha um papel fundamental ao calcular o gradiente da função de perda em relação a cada um dos pesos da *Loss Function*. Esse algoritmo opera em duas fases principais: a primeira é a de *Forward Propagation*, onde a saída da rede é calculada para uma dada entrada; a segunda é de *Error Backpropagation*, na qual o Gradiente é calculado e atualizam-se os pesos da rede de uma maneira como no *Gradient Descent*, conforme descrito abaixo pela Equação 2.6:

$$w_i^{(t+1)} = w_i^{(t)} - \eta \cdot \frac{\partial \mathcal{L}(\mathbf{w})}{\partial w_i} \tag{2.6}$$

Em que η é a taxa de aprendizado, e $\frac{\partial \mathcal{L}}{\partial w_i}$ representa o gradiente da função de perda em relação ao peso w_i. Este processo é repetido em múltiplas iterações para cada neurônio da rede, permitindo que a rede aprenda a mapear adequadamente as entradas para as saídas.

Para otimizar o treinamento de redes neurais, uma técnica amplamente empregada é o *Stochastic Gradient Descent* (SGD) (BARBER, 2012; BISHOP et al., 2006; MOHRI et al., 2012; SUVRIT et al., 2012), uma variação da Descida do Gradiente clássica, que em contraste, ao método tradicional que calcula o gradiente da função de perda em relação a todo o conjunto de dados antes de atualizar os pesos, o SGD ajusta os pesos após cada amostra individual ou em pequenos lotes de dados (chamados de *mini-batches*). Essa característica não só reduz o custo computacional por iteração, mas também introduz variações estocásticas nas atualizações, as quais podem auxiliar o modelo a escapar de mínimos locais e favorecer a generalização, especialmente em domínios com dados extensos e complexos.

2.1.2 Abordagem Clássica com Processamento de Sinais

O processamento de sinais é uma área da engenharia que se dedica à análise, modificação e síntese de sinais, os quais podem ser qualquer forma de dados, incluindo sinais de áudio, imagens ou medições biomédicas. As abordagens tradicionais para amostragem e análise de sinais ao longo do tempo, geralmente, adotam um tratamento clássico que assume uma taxa de amostragem constante e invariante (**??**ULABY et al., 2018; WEI, 2018). Este paradigma clássico tem sido amplamente utilizado devido à sua simplicidade e eficácia, especialmente em cenários no qual os sinais são representados e processados de forma digital ou analógica, com base na tecnologia disponível e nas metodologias historicamente desenvolvidas para filtragem, compressão e transmissão.

A hipótese de uma taxa de amostragem constante pressupõe que o sinal seja amostrado de maneira uniforme ao longo do tempo, ou seja, os intervalos de tempo entre amostras consecutivas são fixos e regulares. Esta abordagem é central nas metodologias tradicionais, ao simplificar a análise e processamento dos sinais, possibilitando o uso de ferramentas matemáticas eficazes que se baseiam nesta regularidade. A amostragem uniforme, descrita matematicamente, é dada pela Equação 2.7:

$$x[n] = x(n \cdot T_s) \tag{2.7}$$

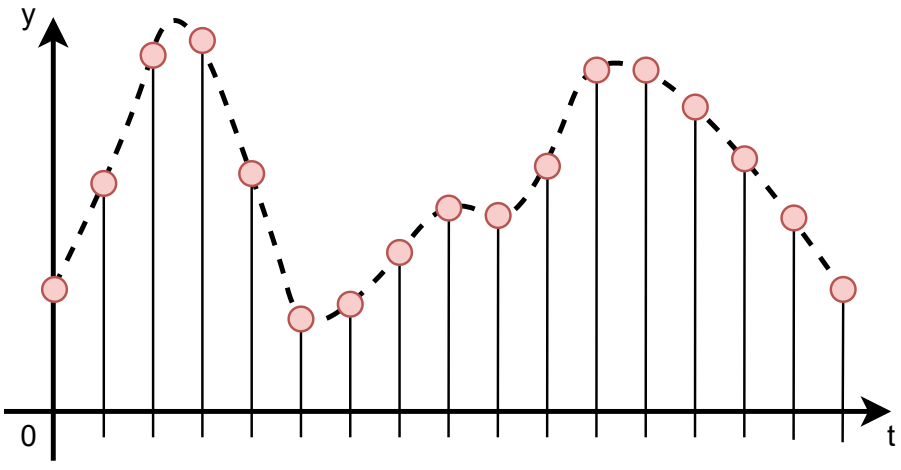
A suposição de uma frequência de amostragem constante é especialmente crítica nas técnicas de transformação e análise espectral. A transformação de sinais no domínio do tempo para o domínio da frequência, através da Transformada de Fourier Discreta (DFT), baseia-se na premissa de que as amostras são uniformemente espaçadas no tempo. Essa uniformidade garante a correta identificação das componentes de frequência presentes no sinal. A DFT, amplamente utilizada para análise espectral (**??**ULABY et al., 2018), é definida pela Equação 2.8:

$$x[e^{j\omega}] = \sum_{n=-\infty}^{+\infty} x[n] \cdot e^{-j\omega N} \tag{2.8}$$

Esta equação assume que o sinal foi amostrado em uma taxa constante, de forma que as amostras se distribuam regularmente ao longo do tempo. Essa regularidade é fundamental, ao permitir que a análise espectral produza resultados precisos e sem distorções, assegurando a integridade dos dados analisados. Na abordagem clássica para séries temporais univariadas, cada canal é tratado como uma sequência de amostras regularmente espaçadas (**??**ULABY et al., 2018), conforme o intervalo de amostragem $T_s = \frac{1}{f_s}$. Assim, com uma frequência de amostragem bem definida, técnicas de processamento de sinais, como a Transformada de Fourier, podem ser aplicadas para realizar previsões e análises espectrais.

Essas técnicas são amplamente empregadas em tarefas de previsão (ou *forecasting*) de séries temporais, permitindo a decomposição dos dados em componentes frequenciais e facilitando a identificação de padrões recorrentes, como sazonalidade e tendências (WEI, 2018). Através dessa decomposição, possibilita-se modelar e prever o comportamento futuro da série, uma vez que a estrutura regular das amostras possibilita a aplicação de métodos robustos de filtragem e análise de espectro. Essa abordagem clássica, portanto, estabelece uma base para a previsão de séries temporais, especialmente em contextos onde os dados seguem uma estrutura de amostragem regular e uniforme, como pode ser ilustrado pela Figura 7.

Figura 7 – Série Temporal Univariada como Sinal Amostrado



A imagem mostra como o sinal de um sensor que consiste em uma série temporal univariada pode ser representado como um sinal coletado mediante uma frequência de amostragem f_s.

Imagem criada autorialmente.

Historicamente, o processamento de sinais desempenhou um papel fundamental

na engenharia, fornecendo uma estrutura matemática sólida para a análise de dados sequenciais, com aplicações em áreas como telecomunicações, controle de sistemas e processamento de áudio. Com base nessa fundação rigorosa, buscou-se aplicar métodos de AI e ML sobre estruturas matemáticas semelhantes, adaptando problemas de predição e análise complexa a abordagens originalmente concebidas para sinais regulares e univariados.

Contudo, a inclusão de múltiplos sensores não sincronizados trouxe desafios à modelagem (WEI, 2018), tornando a regularização dos dados uma tarefa onerosa (BARROS et al., 2024). A aplicação direta de técnicas de AI/ML, embora efetiva em diversos contextos, esbarra em limitações quando se lida com irregularidades comuns em dados reais. Nesse cenário, os métodos tradicionais de processamento de sinais revelam-se insuficientes para capturar adequadamente as interações e as irregularidades observadas em séries temporais complexas e multivariadas, exigindo novas abordagens para explorar essas particularidades.

2.1.3 Séries Temporais Regulares

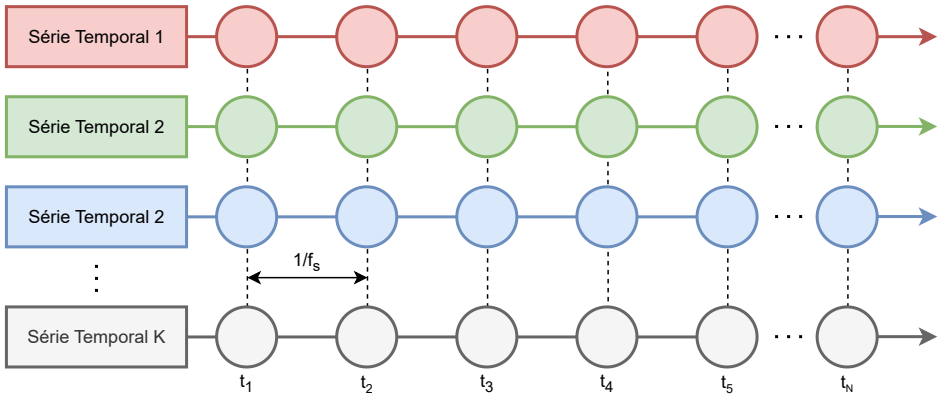
Esta Subseção apresenta os conceitos fundamentais em séries temporais regulares, essenciais para o entendimento dos métodos e técnicas básicos aplicados ao longo deste trabalho. Os principais tópicos abordados incluem a estrutura e as características das RMTS. Para construir uma base sólida sobre o tema, foram consultadas referências essenciais, incluindo estudos recentes e obras clássicas, que oferecem diferentes abordagens e percepções relevantes.

Primeiramente, define-se (DU et al., 2020; MAKRIDAKIS et al., 2018; SHIH et al., 2019) uma Série Temporal Multivariada Regular (RMTS) como uma sequência ordenada e finita de observações coletadas ao longo do tempo, com amostragens que ocorrem em intervalos regulares e frequência constante f_s. Formalmente, $\mathcal{S} = \{\mathbf{Z_i}\}_{i=1}^{N}$ representa um conjunto finito de séries temporais, onde cada elemento desse conjunto, $\mathbf{Z_i} = [Z_{1,t_1}, Z_{2,t_2}, \ldots, Z_{1,t_T}]$, é uma sequência de eventos temporais. Aqui, cada $Z_{i,t} \in \mathbb{R}^{K_i}$ corresponde a um vetor de K_i *features* associadas ao tempo t.

Em séries temporais regularmente amostradas, o conceito de *grid structure* assume relevância ao organizar observações temporais em intervalos constantes, ou seja, a frequência de amostragem f_s não se altera, veja a Figura 8. Nesse contexto, o *grid* visa estruturar os pontos temporais de modo a facilitar o armazenamento e a análise dos dados. Contudo, a organização de séries temporais no formato matricial $\mathbf{M} \in \mathbb{R}^{L \times K}$, onde L denota o número de observações temporais e K o número de variáveis monitoradas, não é uma tarefa trivial. Esse arranjo depende de um conjunto de pressupostos e ajustes sobre a regularidade da série, e pode ser limitado pela própria natureza dos dados coletados, o processo de regularização é algo extremamente custoso em tempo e recursos, além de introduzir vieses, geralmente só pode obter uma série regularizada quando temos um único sensor ou poucos que estão sincronizados, algo que não é comum quando temos sensores

in the wild, como no Porto de Paranaguá.

Figura 8 – Série Temporal Multivariada Regular



O esquema acima mostra uma Série Regularizada em um esquema *grid-like*, note que todas as observações estão equidistantes com intervalo de tempo: $T_s = \frac{1}{f_s}$, isso faz o problema ser simplificado.

Imagem criada autorialmente.

Para séries temporais multivariadas que apresentam regularidade na amostragem, a construção de $\mathbf{M}$ permite que algoritmos de aprendizado supervisionado apliquem métodos convencionais de modelagem e previsão (WEI, 2018), uma vez que os dados são organizados de forma que algoritmos tradicionais possam processá-los diretamente. Contudo, a organização e a padronização das séries temporais em estruturas matriciais requerem que os dados sejam consistentes e padronizados, uma condição que nem sempre é natural, sobretudo em séries que sofrem com irregularidades ou amostragens heterogêneas. A seguir, serão discutidas metodologias específicas para lidar com irregularidades e inconsistências em séries temporais multivariadas.

As principais tarefas de aprendizado para séries temporais incluem (**??**) a classificação, a detecção de anomalias e o *forecasting*. Em particular, o *forecasting* visa prever valores futuros da série com base em padrões e estruturas detectadas em observações passadas. Esse tipo de tarefa é essencial em diversos campos, como economia, engenharia e ciências ambientais, onde decisões estratégicas são baseadas em previsões acuradas. A eficácia do *forecasting* depende de fatores como a precisão do modelo, a qualidade dos dados e a capacidade do modelo de capturar tendências, padrões sazonais e irregularidades.

Antes de explorarmos detalhadamente o estudo do *IMTS-forecasting*, é necessário estabelecer o vocabulário que será utilizado ao longo deste trabalho. A análise de padrões e a previsão de valores futuros em séries temporais dependem de técnicas que estruturam a entrada de dados sequenciais em janelas de análise específicas. Essas janelas segmentam o histórico de dados em intervalos delimitados, permitindo ao modelo capturar tendências e dinâmicas locais. Esse processo de organização dos dados facilita o aprendizado e a

previsão, estruturando informações essenciais para a modelagem de séries temporais. A seguir, apresentamos um conjunto de termos que serão recorrentes neste trabalho, sendo fundamentais para a compreensão das técnicas aplicadas.

i. **Janelamento**: Consiste em segmentar uma série temporal em subsequências fixas de tamanho pré-definido, denominadas de janelas. Cada janela agrupa uma sequência de observações temporais, permitindo ao modelo identificar e aprender padrões recorrentes em intervalos limitados e estruturados.

ii. **Passo** ou ***stride***: Representa o deslocamento entre janelas consecutivas ao longo da série temporal. Um passo pequeno aumenta a sobreposição entre janelas, proporcionando uma densidade maior de dados e capturando variações sutis entre intervalos próximos. Em contrapartida, um passo maior reduz a sobreposição, diminuindo a redundância informacional e permitindo ao modelo observar padrões que se manifestam em intervalos mais distantes.

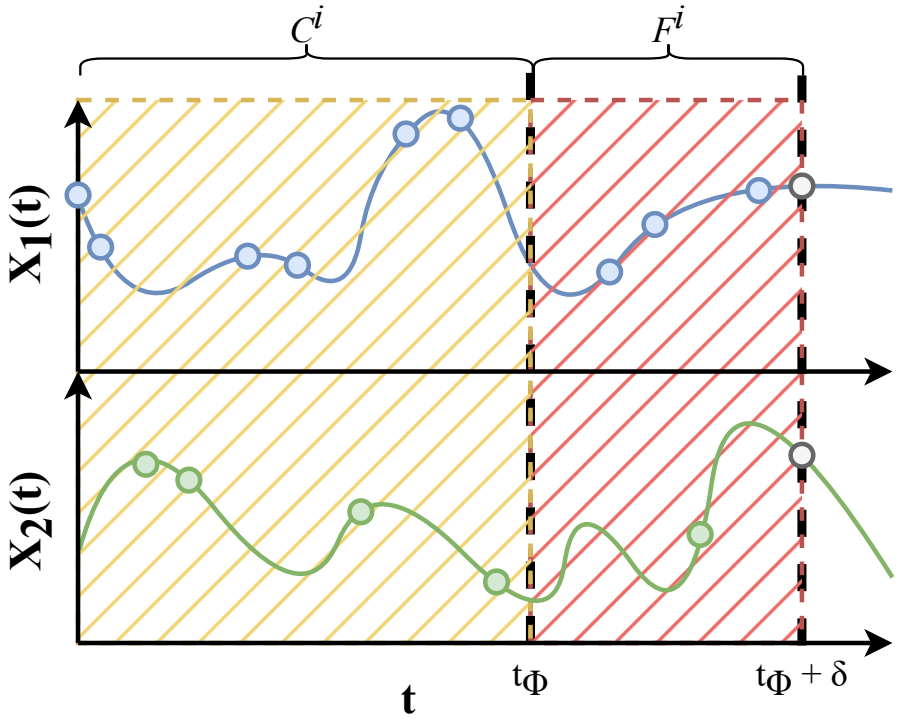
iii. **Janela de Contexto** C^i: Define o intervalo histórico de observações que o modelo utiliza para gerar previsões. Em séries temporais complexas, uma janela de contexto mais ampla pode capturar padrões sazonais ou cíclicos, enquanto uma janela menor pode ser suficiente para séries com dependências de curto prazo. A seleção apropriada do contexto é, portanto, fundamental para o aprendizado de padrões temporais relevantes para a tarefa, para a parte prática deste trabalho usa-se uma janela de contexto de cinco dias.

iv. **Janela de Forecasting**: A *janela de forecasting* ou janela de previsão específica o horizonte de valores futuros que se deseja estimar com base no contexto fornecido. A tarefa de *forecasting* utiliza o histórico contido nas janelas de contexto para prever um ou mais pontos à frente na série temporal. Abordagens para previsão variam desde modelos estatísticos clássicos até redes neurais profundas, e o horizonte de previsão pode variar de curto a longo prazo, dependendo dos objetivos específicos, para parte prática deste trabalho usa-se uma janela de *forecasting* de dois dias.

Segundo (**??**MARLIN, 2020), a tarefa de *forecasting* pode ser definida como inferir $\mathbf{x}[t_\phi + \delta]$ (para $\delta > 0$) condicionando-se nas observações $\mathcal{S}[: t_\phi]$ até o tempo t_ϕ. A tarefa de *forecasting*, tem um objetivo similar à tarefa de predição, pode-se dizer até que são análogas, porém a maior diferença à que primeiramente, estamos inferindo um ponto que está na própria série temporal, em um momento futuro. Ou seja, neste problema o instante de tempo t_ϕ refere-se ao ponto temporal no qual a previsão é feita. Deste modo, $t_\phi + \delta$ com $\delta > 0$ refere-se ao ponto temporal futuro que se deseja realizar uma previsão. Todas as observações definidas em uma janela de contexto C^i até o tempo t_ϕ podem ser utilizadas para calcular uma previsão no tempo futuro $t_\phi + \delta$. A variável δ é frequentemente referida

como o horizonte de previsão e ela engloba uma janela de *forecasting* F^i que vai de t_ϕ até $t_\phi + \delta$, representando o intervalo de tempo no futuro para os quais estamos prevendo o valor da Série Temporal. Ambas as janelas e a tarefa de *forecasting* estão ilustradas na Figura 9.

Figura 9 – Tarefa de *forecasting*



O diagrama acima ilustra como funciona o *forecasting* é mostrado como funcionam as janelas de Contexto (C^i, em amarelo) e de Previsão (F^i, em vermelho) num contexto de previsão de dados em tempo futuro de uma IMTS.

Imagem criada autorialmente, adaptada de (**??**)

2.1.4 Irregularidades nas Séries Temporais

As irregularidades em Séries Temporais referem-se a fatores que alteram ou distorcem a regularidade de uma série temporal, pensando que idealmente essas séries estariam estruturadas em uma *grid* (**??**BARROS et al., 2024; MARLIN, 2020). Essas irregularidades podem comprometer a qualidade das previsões em relação aos dados perfeitamente estruturados; no entanto, o foco deste estudo é destacar que, devido ao histórico de abordagens em Processamento de Sinais, atualmente prefere-se tratar essas irregularidades que estão presentes em sensoriamento múltiplo, em vez de desenvolver modelos específicos para lidar com essas características diretamente. Nesta análise, serão discutidas as diversas fontes de irregularidade e a importância de adotar abordagens que reconheçam e integrem essas particularidades na análise de IMTS.

As irregularidades podem ser classificadas em diferentes tipos. A seguir, são apresentados alguns dos principais tipos, com destaque para suas características e implicações na análise e modelagem de séries temporais. Dentro do Capítulo 4, discutiremos em mais detalhes as irregularidades presentes no conjunto de dados do P^{2}MOD, analisando como cada *feature* e sensor se comportam diante dessas irregularidades.

- **Dados Faltantes:** Dados ausentes surgem de forma esporádica, geralmente devido a falhas nos sensores ou problemas na transmissão das informações, resultando em séries temporais incompletas, gerando *gaps* nos dados.
- **Amostragem Irregular:** Refere-se à variação na frequência de coleta dos dados, que pode ocorrer entre diferentes séries temporais ou até mesmo numa única série. Por exemplo, algumas medições podem ser feitas a cada 5 minutos, enquanto outras são feitas a cada 20 minutos, resultando em intervalos inconsistentes entre os pontos de dados, isso pode ocorrer em uma mesma série temporal ao longo do período em que dados são coletados.
- **Deslocamentos Temporais:** Séries temporais que não estão totalmente alinhadas podem apresentar deslocamentos temporais, quando os tempos de início de diferentes séries temporais estão desalinhados. Essa desordem no tempo das observações torna a série mais esparsa em relação à grade subjacente, aumentando a complexidade na regularização dos dados para métodos convencionais e dificultando a sincronização dos dados.
- **Covariáveis Conhecidas:** São variáveis externas ao sistema de interesse, mas que influenciam o comportamento das variáveis alvo. Essas covariáveis são conhecidas para tempos futuros ($t > t_\phi$), permitindo que seu efeito seja considerado nas previsões das variáveis de interesse.
- **Eventos Parcialmente Observados:** Em dados estruturados em *grid*, assume-se que, quando um sensor coleta informações, todas as suas *features* são capturadas. No entanto, essa premissa nem sempre se aplica no contexto de IMTS.
- **Duplicatas:** Ocorre quando o mesmo instante temporal tem várias medições registradas, o que pode acontecer devido a falhas no sensor ou erros no registro dos dados. Essas duplicatas podem distorcer análises estatísticas e algoritmos que assumem carimbos de data e hora únicos.

Propõe-se baseado nos estudos de (**??**BARROS et al., 2024) uma definição formal: uma IMTS é um par de conjuntos finitos de sequências $\mathcal{I} = (\mathcal{S},\ \mathcal{T}_\mathcal{S})$. O conjunto de dados e medições é adotado como $\mathcal{S} = \{\mathbf{Z^i}\}_{i=1}^{N}$ contém $N = |S|$ sequências da forma $\mathbf{Z^i} = [z^i_{t_1}, z^i_{t_2}, \ldots, z^i_{t_T}]$, onde cada evento medido $z^i_t \in \mathcal{X}^i$ está associado a um instante

temporal $t \in \mathbb{R}$. O conjunto $\mathcal{T}_{\mathcal{S}} = \{\mathbf{T^i}\}_{i=1}^{N}$ agrupa todas as sequências de instantes de tempo $\mathbf{T^i} = [t_1^i, \ldots, t_T^i]$ associadas a $\mathcal{S}$. Essa definição abrange as diferentes formas de assincronia e as particularidades das IMTS, observe que $\mathcal{T}_{\mathcal{S}}$ é composto por diversas séries temporais que, não necessariamente, estão alinhadas ou regularizadas entre si.

2.2 Modelagem Clássica de Séries Temporais

Nesta Seção serão abordadas as técnicas fundamentais para se lidar com Séries Temporais, aquelas que servirão de base e estão mais intrinsecamente ligadas à área de Processamento de Sinais. Na Subseção 2.2.1 será abordado métodos de regularização para transformação de uma Série Irregular para uma estrutura *grid-like*, em conseguinte na Subseção 2.2.2 serão trazidos algumas fundamentações teóricas de um modelo clássico, o *Gated Recurrent Unit*, que será utilizado como um caso de estudo no P^{2}MOD.

2.2.1 Regularização

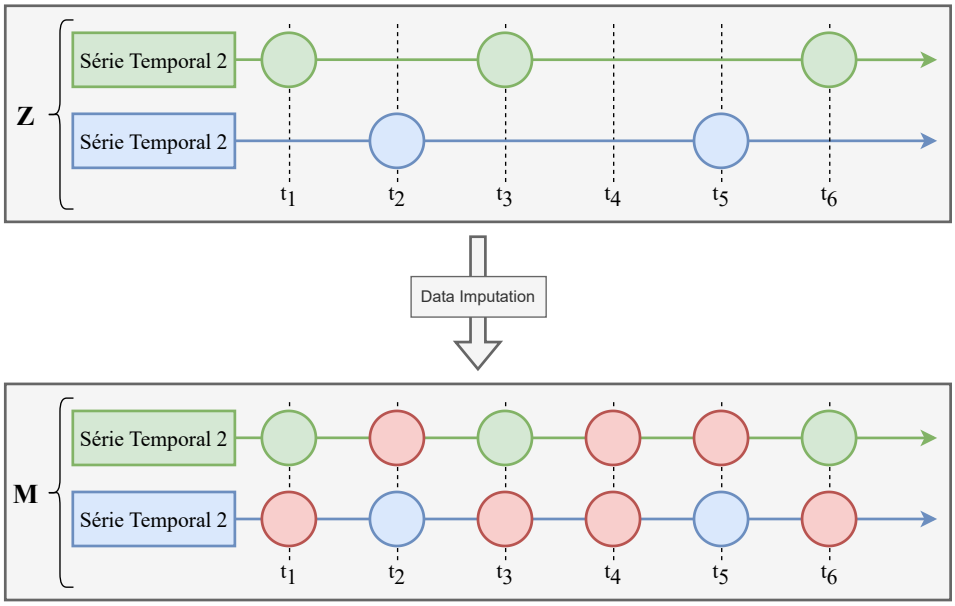
Nos contextos tradicionais, que estão mais relacionados com a área de Processamento de Sinais, técnicas de regularização são interessantes para adequar as irregularidades presentes em séries temporais ao ferramental clássico. Assim, técnicas como *masking* e de *data imputation* desempenham um papel fundamental para adequar séries que contém irregularidades em uma estrutura *grid-like*.

Primeiramente, proponho comentar sobre técnicas de *data imputation* que são comuns ao se abordar dados faltantes. Esse ferramental se refere a um conjunto de técnicas voltadas para preencher as lacunas geradas por absência de dados. A imputação tenta estimar valores plausíveis para, automaticamente, preencher esses pontos, visando obter um conjunto de dados regular que esteja completo e sem interrupções, no caso deste trabalho isso seria obter uma RMTS a partir de uma IMTS.

Existem diversas abordagens para imputação de dados, que variam em complexidade ou objetivo, sendo escolhidas conforme as características dos dados e o contexto da aplicação. Métodos mais simples incluem o preenchimento pela média e a interpolação linear. O primeiro é uma abordagem básica onde os valores ausentes são substituídos pela média dos valores observados da mesma variável, sendo um método rápido e fácil de implementar que pode, porém, introduzir viés, especialmente em séries temporais, pois não considera a variabilidade temporal dos dados. O segundo é outra técnica simples, que assumo que os pontos faltantes estão linearmente localizados entre os pontos adjacentes, dessa forma a imputação é feita através do cálculo de um plano. Para cenários mais complexos, métodos sofisticados de imputação podem ser aplicados. Um exemplo é a imputação por *K-Nearest Neighbors* (KNN), que substitui valores ausentes com base nos valores de observações similares (ou "vizinhos") em outras partes do conjunto de dados.

Observe que, independentemente do método de imputação escolhido, em IMTS, o processo de regularização para aplicação de métodos clássicos de aprendizado de máquina pode tornar-se bastante oneroso e introduzir características e vieses que não são pertinentes e interessantes para análises com modelos de AI/ML. Isso ocorre porque cada uma dos diversos tipos irregularidades precisam ser tratados e entendidos como dados faltantes frente a um *grid-like*, aumentando a complexidade e o volume de preenchimento necessário. A Figura 10 ilustra esse fenômeno em uma pequena IMTS submetida a um processo de regularização, onde a série resultante exibe mais valores imputados do que dados originais.

Figura 10 – Imputação de Dados



A figura acima mostra um processo de imputação de dados genérico para uma IMTS, observe como uma série com originalmente 5 dados, torna-se uma série completa, com 12 dados, embora 7 deles tenham sido gerados artificialmente.

Imagem criada autorialmente

Uma forma de contornar o problema gerado pelo grande número de dados imputados é utilizar o *masking*. A máscara de dados é uma técnica aplicada para lidar com dados ausentes, identificando explicitamente os pontos onde as informações estão faltantes. Ao aplicar uma máscara, o modelo consegue reconhecer as posições dos valores ausentes sem tentar ajustá-los diretamente, preservando a integridade dos dados e evitando que a ausência de informações introduza viés nos cálculos subsequentes. A máscara é representada por uma matriz binária em que cada valor booleano indica a presença ou ausência de dados, sendo que, usualmente, o valor 1 na máscara representa as posições mascaradas.

Para exemplificar, considere uma matriz de dados de entrada $\mathbf{X_{input}}$ e uma matriz de máscara $\mathbf{X_{mask}}$, ambas de dimensões 2×2, onde o valor 1 na máscara representa a presença de mascaramento:

$$\mathbf{X_{input}} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{X_{mask}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2.9}$$

Ao aplicar a máscara $\mathbf{X_{mask}}$ à matriz de dados $\mathbf{X_{input}}$, obtemos a matriz resultante $\mathbf{X_{output}} = K(\mathbf{X_{input}}, \mathbf{X_{mask}})$, no qual os valores correspondentes a 1 em $\mathbf{X_{mask}}$ são substituídos por `NaN` (*Not a Number*, uma notação para dados ausentes em `Python` e outras linguagens de programação). A matriz resultante é:

$$\mathbf{X_{output}} = K(\mathbf{X_{input}}, \mathbf{X_{mask}}) = \begin{bmatrix} \texttt{NaN} & 2 \\ 3 & \texttt{NaN} \end{bmatrix} \tag{2.10}$$

A aplicação de técnicas de regularização, como *data imputation* e *masking*, em IMTS, é particularmente desafiadora devido à complexidade multiplicada pelas várias variáveis, tornando-se uma tarefa ainda mais onerosa à medida que se aumenta a irregularidade dos dados e o número de *features*. Essa abordagem não apenas exige tratamento personalizado para cada variável, escalando o esforço computacional e humano para tratativas, mas também pode introduzir vieses devido à generalização inadequada presente nessas técnicas.

2.2.2 Gated Recurrent Unit

A *Gated Recurrent Unit* (GRU) é uma arquitetura de Rede Neural Recorrente (RNN) desenvolvida para lidar com o problema de aprendizado em séries temporais e dados sequenciais, para evitar limitações comuns em modelos recorrentes tradicionais, como o decaimento de gradiente. Proposta por (CHO, 2014), a GRU incorpora portas (*gates*) especializadas que controlam o fluxo de informações, possibilitando uma melhor modelagem de dependências de longo prazo, a GRU adota uma estrutura contendo apenas duas portas principais: a porta de atualização (*update gate*) e a porta de reinicialização (*reset gate*).

- **Porta de Atualização (Update Gate)**: A porta de atualização controla a quantidade de informação da célula anterior que será levada para a próxima etapa. Isso permite que a GRU armazene informações relevantes ao longo do tempo, enquanto descarta gradualmente informações menos relevantes. Matematicamente, a porta de atualização é definida pela função:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{2.11}$$

  onde z_t representa o valor da porta de atualização no tempo t, h_{t-1} é o estado oculto da etapa anterior, x_t é o valor da entrada atual, e W_z e b_z são, respectivamente, os pesos e o viés associados à porta de atualização.

- **Porta de Reinicialização (Reset Gate)**: A porta de reinicialização controla o quanto da informação passada deve ser esquecida ao gerar o novo estado. Ela permite que a GRU decida se deve ou não ”reiniciar”o estado oculto em certas etapas, permitindo um reset parcial ou total do estado da célula. A porta de reinicialização é definida como:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{2.12}$$

onde r_t representa o valor da porta de reinicialização no tempo t, e W_r e b_r são os pesos e o viés associados a essa porta.

Após o funcionamento das portas de atualização e reinicialização, o estado oculto da GRU é atualizado com base na combinação do estado anterior e da nova informação, ponderada pelas portas. Essa atualização do estado oculto h_t é calculada pela fórmula:

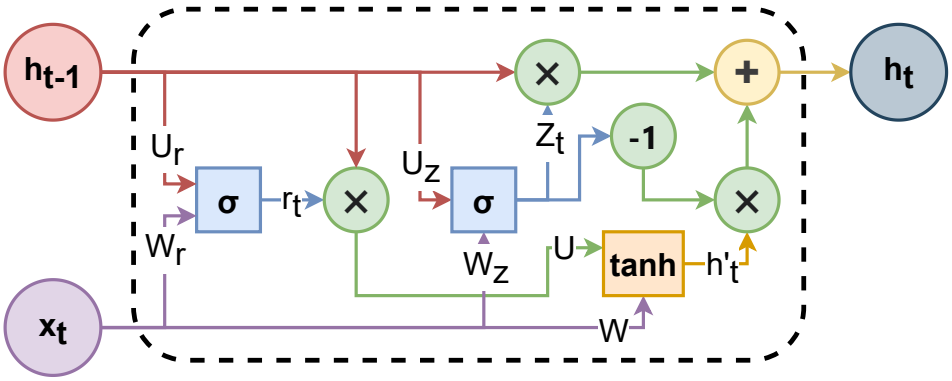
$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot h'_t \tag{2.13}$$

onde h'_t representa o novo conteúdo de memória, definido como:

$$h'_t = \tanh(W_h \cdot [r_t \cdot h_{t-1}, x_t] + b_h) \tag{2.14}$$

Essa equação permite que o modelo armazene informações de longo prazo (via z_t) enquanto mantém a flexibilidade para descartar informações irrelevantes.

Figura 11 – Diagrama Esquemático de uma GRU



A figura acima demonstra como uma unidade da GRU funciona, mostrando suas portas e operações internas.

Imagem criada autorialmente

A arquitetura simplificada da GRU (que pode ser visualizada na Figura 11), que apresenta poucos parâmetros, é uma alternativa atraente para problemas que demandam modelagem eficiente de dependências temporais. As GRUs têm sido amplamente utilizadas por sua eficiência em capturar padrões complexos em dados sequenciais, mantendo um bom equilíbrio entre desempenho e eficiência computacional.

2.3 Abordagens Modernas em Séries Temporais

Nesta seção, serão exploradas técnicas contemporâneas que aprimoram a análise de séries temporais ao abordar desafios específicos associados à dimensão temporal. Inicialmente, será apresentado o conceito de *Time Encoding* (Seção 2.3.1), uma estratégia que permite incorporar informações temporais de maneira explícita, viabilizando a aplicação de modelos que dependem de representações ordenadas do tempo. Em seguida, discutiremos a abordagem de previsão *Gap-Ahead* (Seção 2.3.2), proposta por (BARROS et al., 2024) que foca em estimativas para intervalos temporais definidos à frente, atendendo a demandas onde previsões de curto prazo ou contínuas são insuficientes. Essas técnicas refletem avanços no campo e ampliam o leque de ferramentas disponíveis para o processamento e a modelagem de séries temporais complexas.

2.3.1 Time Encoding

O *Time Encoding*, ou codificação temporal, refere-se a uma técnica fundamental em modelos que trabalham com dados sequenciais, permitindo a inclusão explícita de informações temporais ou posicionais numa série temporal. Essa codificação é crucial quando o modelo precisa considerar a ordem dos eventos ou medir intervalos entre eles, como em redes neurais aplicadas em séries temporais ou dados sequenciais de sensores.

Essa abordagem de *Temporal Encoding* é baseada diretamente no *Positional Encoding*, introduzida por (VASWANI et al., 2017), a qual é amplamente usada para codificar a posição de elementos em uma sequência, geralmente textual. Embora essa técnica seja comumente utilizada em modelos de Processamento de Linguagem Natural para adicionar informações posicionais ao mecanismo de atenção totalmente conectado nos Transformers, neste Trabalho essa técnica é adaptada para fornecer uma base matemática para o *Temporal Encoding*. Originalmente proposto por Vaswani o *Positional Encoding* é definido para uma posição pos e uma dimensão i como:

$$\begin{aligned} \mathrm{PE}_{(pos,2i)} &= \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \\ \mathrm{PE}_{(pos,2i+1)} &= \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \end{aligned} \tag{2.15}$$

onde d representa a dimensão da codificação. Essa abordagem gera representações que variam conforme a posição na sequência sendo integradas com os valores das entradas, enriquecendo o modelo com informações temporais.

Aplicado ao domínio temporal, o *Time Encoding* facilita a captura de relações temporais complexas entre eventos e melhora a interpretação e precisão dos modelos em dados sequenciais, especialmente em contextos como previsão de séries temporais e análise de comportamento temporal. Para a generalização da fórmula proposta por (VASWANI et

al., 2017) usaremos a definição de (BARROS et al., 2024) que propõem primeiramente um t' que representa t relativo ao tempo de inferência t_ϕ, em seguida o *Positional Encoding* é aplicado para criar uma representação vetorial τ_{t^i} dessa distância temporal que tem um tamanho $\mathcal{T}$, assim:

$$t' = t - t_\phi \tag{2.16}$$

$$\begin{aligned} \tau_{t',2k} &= \sin\left(\frac{t'}{1000^{2k/\mathcal{T}}}\right) \\ \tau_{t',2k+1} &= \cos\left(\frac{t'}{1000^{2k/\mathcal{T}}}\right) \end{aligned} \tag{2.17}$$

2.3.2 Gap-Ahead

O modelo *Gap-Ahead* é uma arquitetura modular desenvolvida para integrar previsões de modelos numéricos com dados observacionais disponíveis, permitindo representar eventos que abrangem tanto momentos passados quanto futuros (covariáveis conhecidas). Essa abordagem visa superar os desafios associados a séries temporais multivariadas amostradas irregularmente, garantindo robustez mesmo na presença de dados ausentes e mantendo custos computacionais reduzidos. O modelo fundamenta-se em três princípios principais:

1. **Codificação de Marcações Temporais com Funções Periódicas:** Este módulo assegura que informações temporais sejam consistentemente representadas ao longo das séries temporais, capturando padrões cíclicos através do uso de funções periódicas aplicadas às diferenças temporais.

2. **Codificação Temporal Independente das Séries Temporais:** Cada série temporal é processada por uma rede neural recorrente (RNN) independente, o que permite que o modelo continue funcional, mesmo na ausência de dados de sensores específicos.

3. **Difusão de Informação com Redes de Atenção de Grafos Heterogêneos Regularizados (RHGAT):** Este mecanismo facilita a disseminação de informações entre os nós do grafo, enriquecendo as representações de cada nó com base nas informações dos vizinhos.

O funcionamento do modelo pode ser descrito pelas etapas a seguir. Primeiramente, uma amostragem uniforme de tempos de referência t_ϕ é realizada no intervalo temporal disponível. Para cada t_ϕ, são determinados os intervalos de contexto e de previsão:

$$t'_\phi = t_\phi + O_i, \tag{2.18}$$

$$Z_i^c = \left[Z_{i,t} \in Z_i \,|\, t'_\phi - C_i \leq t < t'_\phi \right], \tag{2.19}$$

$$Z_i^f = \left[Z_{i,t} \in Z_i \,|\, t'_\phi \leq t < t'_\phi + F_i \right], \tag{2.20}$$

onde O_i representa o deslocamento temporal para uma série específica, C_i é o tamanho do intervalo de contexto e F_i é o tamanho do intervalo de previsão.

Para lidas com os intervalos temporais irregulares, é utilizada a técnica de *Time Encoding* descrita anteriormente com a Equação 2.17, permitindo que a diferença temporal entre observações seja representada de forma contínua. Cada marcação temporal t é representada em relação ao tempo de referência t_ϕ. Então, cada série temporal é processada por uma RNN que incorpora a codificação temporal:

$$h_{i,t} = \mathrm{RNN}_i \left(Z_{i,t-1} \parallel \tau_{t'}, h_{t-1} \right), \tag{2.21}$$

onde $\parallel$ denota a operação de concatenação. A difusão de informações é realizada por meio de uma Rede de Atenção de Grafos Heterogêneos Regularizada (RHGAT), que atualiza as representações dos nós alvo com base nas informações dos vizinhos:

$$h'_{i,t} = \sum_{r \in R} \sum_{b=1}^{B} \mathrm{GAT} \left(\{ a_{r,b} \cdot h_j : j \in N^{(r)}(i) \} \right), \tag{2.22}$$

onde $a_{r,b}$ é um coeficiente escalar treinável, e $N^{(r)}(i)$ representa o conjunto de vizinhos de um nó alvo.

Finalmente, as previsões são geradas utilizando as representações latentes enriquecidas:

$$\hat{Z}_{i,t} = W_i h'_{i,t} + b_i, \tag{2.23}$$

em que W_i e b_i são parâmetros aprendidos para cada nó.

O modelo *Gap-Ahead* demonstra sua eficácia ao integrar dados multimodais, incluindo imagens, medições locais e previsões numéricas, permitindo previsões robustas e flexíveis em condições adversas, como ausência de dados ou ruídos. Essa abordagem inovadora contribui significativamente para a previsão precoce de eventos extremos, como marés de tempestade, promovendo a segurança comunitária e a tomada de decisão informada (BARROS et al., 2024; WANG et al., 2019; VELIcKOVIć et al., 2017).

2.4 Datasets existentes com Irregulatidades

Diversos datasets da literatura são utilizados para diferentes tarefas em áreas variadas, apresentando características específicas que os tornam adequados a determinados objetivos. A Tabela 1 apresenta uma comparação detalhada de alguns desses datasets relevantes, descrevendo suas características, tarefas e principais aplicações:

- **GRID (COOKE et al., 2006):** Este dataset foi desenvolvido para a tarefa de *Speech Recognition* (Reconhecimento de Fala), com foco na classificação. Contudo, não considera irregularidades como dados ausentes, amostragem irregular ou deslocamento temporal (*timeshift*).
- **PhysioNet 2012 (SILVA et al., 2012):** Voltado para a área de Cuidados de Saúde, este dataset é utilizado para tarefas de classificação e regressão. Ele introduz desafios como dados ausentes e amostragem irregular, mas não contempla a presença de deslocamento temporal e covariáveis conhecidas.
- **PAMAP2 (REISS et al., 2012):** Focado no monitoramento de atividades, este dataset suporta múltiplas tarefas, incluindo classificação, previsão e regressão. Apesar de incluir amostragem irregular, dados ausentes e deslocamento temporal; ele não aborda as covariáveis conhecidas.
- **MIMIC-III (JOHNSON et al., 2016):** Este dataset, utilizado na área de Cuidados de Saúde, é projetado para tarefas de classificação, previsão e regressão. Ele oferece uma base com a presença de todas as irregularidades, exceto covariáveis conhecidas.
- **USHCN (MENNE et al., 2016):** Voltado para estudos de climatologia, este dataset é utilizado para a tarefa de previsão. Ele incorpora apenas os dados faltantes.
- **Beijing Air PM2.5 (Zhang et al., 2017):** Este dataset concentra-se na qualidade do ar tendo sido projetado para tarefas de imputação e regressão. Ele aborda apenas as irregularidades de dados faltantes.
- **PhysioNet 2019 (Reyna et al., 2020):** Focado na área de *Healthcare*, este dataset suporta tarefas de classificação e regressão. Ele inclui desafios relacionados a dados ausentes, deslocamento temporal e amostragem irregular, mas não aborda covariáveis conhecidas.
- **TGB (Huang et al., 2023):** Este dataset apresenta múltiplos campos de aplicação, sendo construído para ser um *benchmark* para as tarefas de classificação e regressão, abordando todos os tipos de irregularidades, exceto as covariáveis conhecidas.

- **P²MOD:** Desenvolvido para a área de oceanografia e meteorologia, este dataset é projetado para tarefas de previsão e apresenta todas as irregularidades propostas: dados ausentes, amostragem irregular, deslocamento temporal e covariáveis conhecidas.

Tabela 1 – Outros *datasets* com Irregularidades

Dataset	**Dados Faltantes**	**Amostragem Irregular**	**Deslocamento Temporal**	**Covariáveis Conhecidas**
GRID (COOKE et al., 2006)	✗	✓	✗	✗
PhysioNet 2012 (SILVA et al., 2012)	✓	✓	✗	✗
PAMAP2 (REISS et al., 2012)	✓	✓	✓	✗
MIMIC-III (JOHNSON et al., 2016)	✓	✓	✓	✗
USHCN (MENNE et al., 2016)	✓	✗	✗	✗
Beijing Air PM2.5 (ZHANG et al., 2013)	✓	✗	✗	✗
PhysioNet 2019 (REYNA et al., 2020)	✓	✓	✓	✗
TGB (HUANG et al., 2024)	✓	✓	✓	✗
P²MOD	✓	✓	✓	✓

A tabela acima mostra uma comparação qualitativa com outros *datasets* encontrados, mostrando que apenas o P²MOD apresenta todos os tipos de irregularidades propostas.

Conforme observado na Tabela 1, o **P²MOD**, proposto neste trabalho, destaca-se por ser o único dataset que integra todas as irregularidades consideradas: dados ausentes, amostragem irregular, deslocamento temporal e covariáveis conhecidas. Essas características tornam o **P²MOD** um recurso inédito para a comunidade científica, uma vez que os datasets existentes geralmente abordam apenas um subconjunto desses desafios. Ao reunir todas essas irregularidades em um único conjunto de dados, o **P²MOD** cria um cenário mais próximo das condições reais enfrentadas por sistemas de previsão, permitindo a avaliação mais abrangente e robusta de modelos preditivos.

Essa contribuição é particularmente relevante para a área de previsão em séries temporais multivariadas irregulares, o **P²MOD** fornece uma base única para explorar e validar abordagens que enfrentam essas dificuldades, contribuindo para avanços significativos em áreas críticas, como a oceanografia e meteorologia.

3 Metodologia do Trabalho

Nesta seção, apresentamos detalhadamente os procedimentos metodológicos adotados para a preparação, análise e experimentação do dataset P^{2}MOD. O capítulo está estruturado de maneira a fornecer uma visão clara e sequencial de todo o processo, dividindo-se em quatro etapas principais. A Seção 3.1 aborda as estratégias de preparação e pré-processamento dos dados, destacando as técnicas de normalização e transformação aplicadas. A Seção 3.2 discute a análise exploratória de dados (EDA), que foi fundamental para a compreensão das características do dataset. Na Seção 3.3, detalhamos os experimentos realizados com os modelos de aprendizado de máquina *Gated Recurrent Unit* (GRU) e *Gap-ahead* com *Time Encoding*, ambos escolhidos por sua capacidade de lidar com séries temporais irregulares. Por fim, a Seção **??** apresenta as contribuições metodológicas e a importância do trabalho no contexto de previsão em séries temporais multivariadas irregulares. A Figura 12, abaixo, demonstra como foi dividido e organizado este Trabalho.

Figura 12 – Esquema da divisão da Metodologia do Trabalho

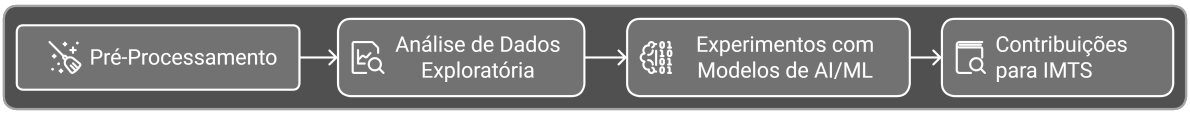


Imagem criada autorialmente

3.1 Preparação e Pré-Processamento dos Dados

A primeira etapa da metodologia envolveu a preparação e o pré-processamento do dataset P^{2}MOD, visando garantir a qualidade e a consistência dos dados utilizados nos experimentos. Inicialmente, os dados foram organizados e convertidos para o Sistema Internacional de Unidades (SI), ou seja, colunas que originalmente estavam em unidades de Nós (kts) foram convertidas para metros por segundo (m/s), assegurando a uniformidade nas escalas das variáveis e possibilitando uma análise mais rigorosa e confiável. Posteriormente, aplicou-se uma normalização à algumas colunas do dataset, com o intuito de reduzir discrepâncias entre diferentes variáveis e facilitar o treinamento dos modelos de aprendizado de máquina, que podem ser sensíveis a escalas divergentes.

Além disso, para lidar com variáveis de direção, como ângulos representados em graus, foi realizada uma transformação para as componentes trigonométricas *sin* e *cos*. Essa abordagem permitiu que as direções fossem representadas de maneira contínua e adequada, eliminando ambiguidades decorrentes da descontinuidade angular, como a transição de

360° para 0°. Essas etapas de pré-processamento foram fundamentais para assegurar que os dados estivessem devidamente preparados para a análise exploratória e a modelagem subsequentes. A ilustração da Figura 13 mostra a divisão da preparação de dados neste Trabalho.

Figura 13 – Ilustração da Preparação dos Dados

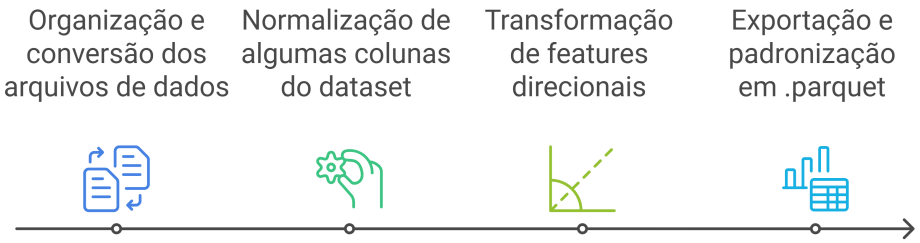


Imagem criada autorialmente

3.2 Análise Exploratória dos Dados

A segunda etapa da metodologia foi a realização de uma análise exploratória de dados, cujo objetivo principal foi investigar as características intrínsecas do dataset P^{2}MOD.

Durante essa etapa, foram analisadas as distribuições estatísticas das variáveis, permitindo identificar padrões relevantes e possíveis anomalias nos dados. Além disso, exploraram-se as correlações entre as variáveis, com o intuito de identificar relações importantes que poderiam influenciar o desempenho dos modelos preditivos, dessa forma a EDA serviu para descartar algumas colunas que seriam usadas nos modelos de caso de uso e também serviram para ter uma ideia qualitativa das distribuições e grandezas do *dataset*.

Figura 14 – Divisão de tarefas realizada no EDA

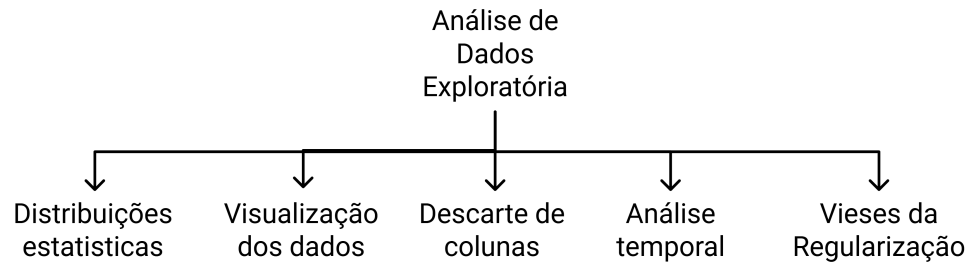


Imagem criada autorialmente

Adicionalmente, foi realizada uma análise temporal detalhada, visando compreender os comportamentos das séries ao longo do tempo e verificar a presença de irregularidades, as percepções obtidos nesta etapa foram cruciais para orientar as decisões tomadas nas etapas

subsequentes, focando em mostrar como o método de Regularização introduz diversos vieses nos dados mediante uma sobrecarga de dados imputados em relação ao número de dados reais. A Figura 14 mostra como foi realizada a divisão das tarefas da etapa de EDA de maneira ilustrativa.

3.3 Modelagem e Testes com AI/ML

Com os dados devidamente preparados e explorados, foram realizados dois experimentos principais com modelos de aprendizado de máquina. Os quais são explicitados pela Figura 15, abaixo:

Figura 15 – Modelos usados no Trabalho

Imagem criada autorialmente

O primeiro experimento empregou o modelo *Gated Recurrent Unit* (GRU), uma variante de redes neurais recorrentes amplamente utilizada para lidar com séries temporais. O GRU foi escolhido por sua capacidade de capturar dependências de longo prazo e modelar dinâmicas temporais complexas, características essenciais no contexto de séries temporais irregulares. A ideia de usar uma *GRU-standard* foi a de justamente mostrar como um modelo mais simples e, sem as capacidades intrínsecas de relacionar todas as *features* pode desempenhar em um problema complexo de *IMTS-forecasting* de uma série fundamentalmente cheia de irregularidades.

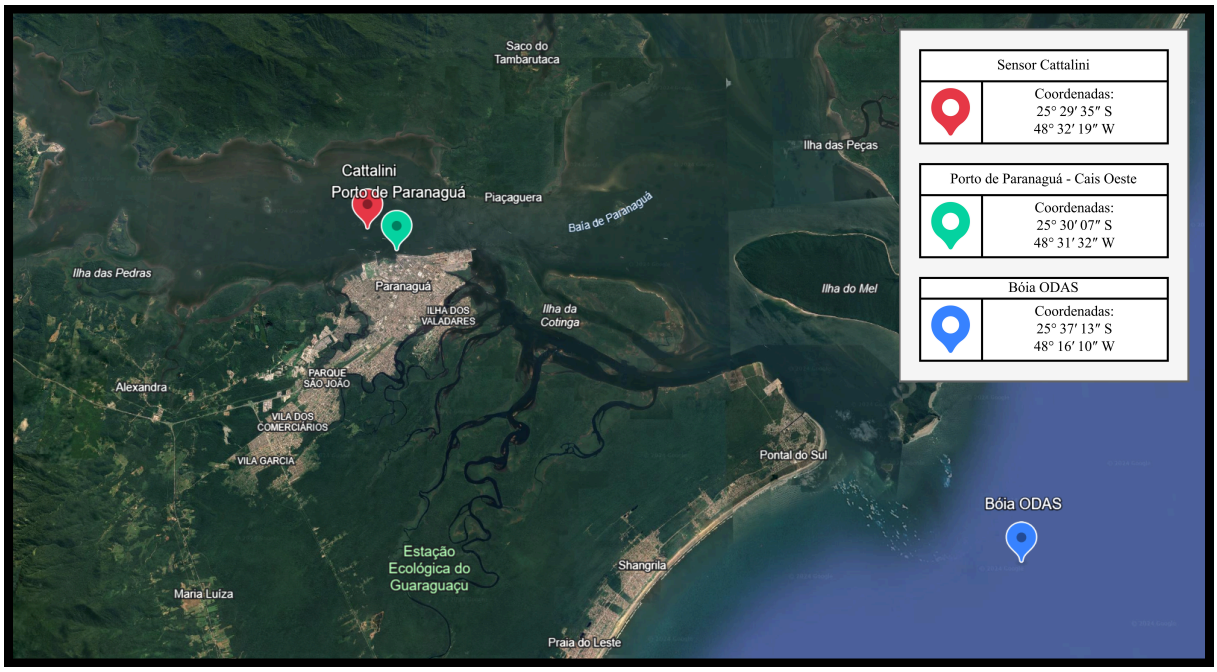
O segundo experimento utilizou o modelo bem mais sofisticado, o *Gap-Ahead* com *Time Encoding* para lidar de maneira eficaz com as irregularidades características do dataset P^2MOD, como amostragem irregular e deslocamentos temporais. Esse modelo por incorporar, além das GRUs, também um módulo de GNNs e *Time Encoding* que se mostram muito mais eficientes em representar e lidar com as irregularidades dos dados.

Ambos os modelos foram treinados e avaliados nas mesmas partições de dados utilizando a métrica de *Index of Agreement* (IoA) (WILLMOTT, 1981) (que será discutida em mais detalhes no Capítulo 5) e que permitiu a comparação entre os modelos de GRU com o Gap-Ahead + *Time Encoding* para a tarefa de *IMTS-forecasting*

4 Paranaguá Port Meteorological and Oceanographic Dataset

Neste capítulo, apresentamos o conjunto de dados utilizado neste estudo, coletado em três localidades estratégicas: a Boia ODAS, o terminal da Cattalini e o Porto de Paranaguá. Primeiramente, discutimos as características dos dados provenientes da Seção 4.1, com foco nas informações de correntometria e medições meteorológicas. Em seguida, abordamos os dados coletados no Terminal da Cattalini na Seção 4.2, que incluem variáveis meteorológicas, correntometria e informações sobre o nível do mar. Posteriormente, exploramos os dados obtidos pelos Sensores do Porto de Paranaguá na Seção 4.3, que incluem medições do nível do mar e valores teóricos de marés harmônicas e astronômicas. Por fim, discutimos as Discrepâncias e Irregularidades do Dataset na Seção 4.4, analisando os desafios enfrentados e as soluções implementadas para garantir a confiabilidade e a integridade dos dados.

Figura 16 – Localização dos Sensores de Coleta de Dados



A Figura mostra a localização dos sensores ambientais e marégrafos instalados nas proximidades da região do Porto de Paranaguá.

Imagem criada autorialmente

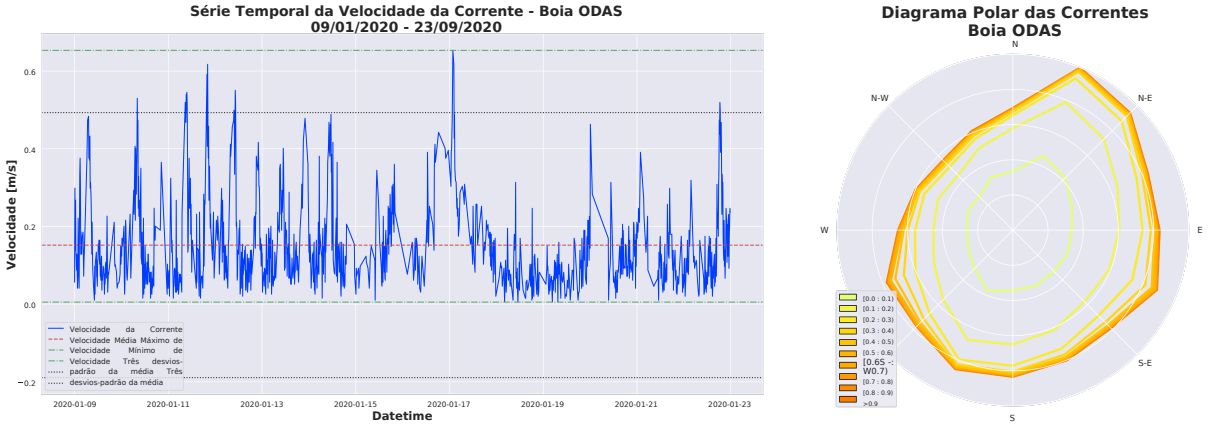
Esse conjunto de dados é de grande relevância para o problema de *IMTS-forecasting*, por oferecer abrangente e integrada das condições marítimas e atmosféricas em uma região

de alta importância econômica e a Figura 16 mostra a localização de cada um dos sensores.

4.1 Boia ODAS

A Boia ODAS é responsável pela coleta de duas bases de dados principais: uma relacionada à meteorologia e outra à dinâmica das correntes marinhas. A base de meteorologia inclui colunas referentes à velocidade dos ventos (em nós, *kts*), rajada dos ventos e direção dos ventos (em graus, $°N$). Já a base de correntes fornece informações detalhadas sobre a velocidade e direção das correntes em diferentes profundidades. Essas profundidades são representadas por seis níveis, com as variáveis *Velocidade 1* e *Direção 1* até *Velocidade 6* e *Direção 6*, correspondendo às medições feitas em diferentes camadas da coluna d'água.

Figura 17 – Análise dos Dados de Corrente — Boia ODAS



A esquerda temos uma série temporal e à direita um diagrama polar, ambas da velocidade das correntes.

Imagem criada autorialmente

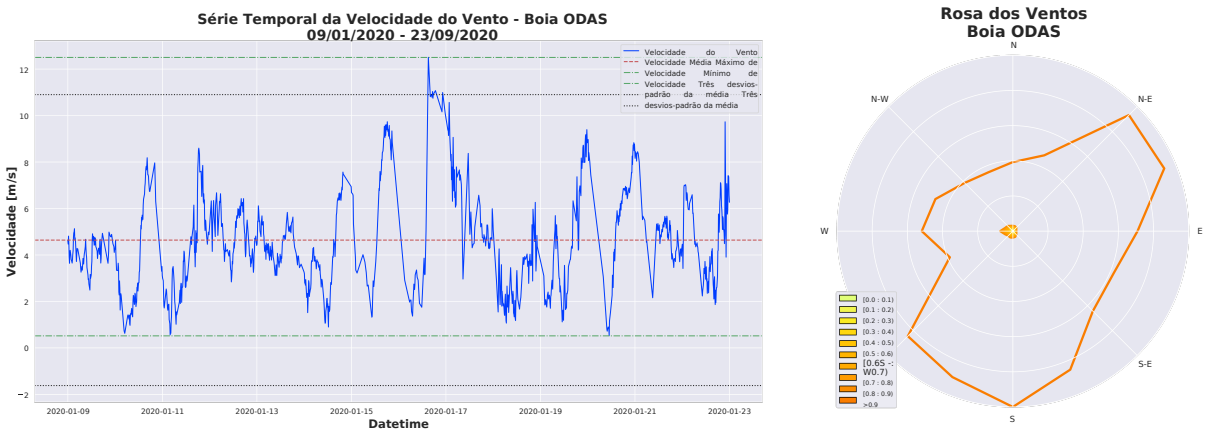
O período representado na Figura 17 mostra ilustrativamente como se comporta essa série, mostrando componentes sazonais e a predominância de uma direção.

De forma análoga, a Figura 18 traz uma análise gráfica dos dados meteorológicos. A série temporal ilustra as flutuações na velocidade dos ventos, enquanto o diagrama polar evidencia as direções predominantes dos ventos registrados ao longo do período analisado. Essa visualização auxilia na compreensão das interações atmosféricas que influenciam a região monitorada pelas Boia ODAS.

4.2 Terminal da Cattalini

O Terminal da Cattalini é uma das localidades estratégicas deste estudo, contendo três bases de dados principais: meteorologia, correntes e marégrafo. Cada uma dessas

Figura 18 – Análise dos Dados de Meteorologia — Boia ODAS

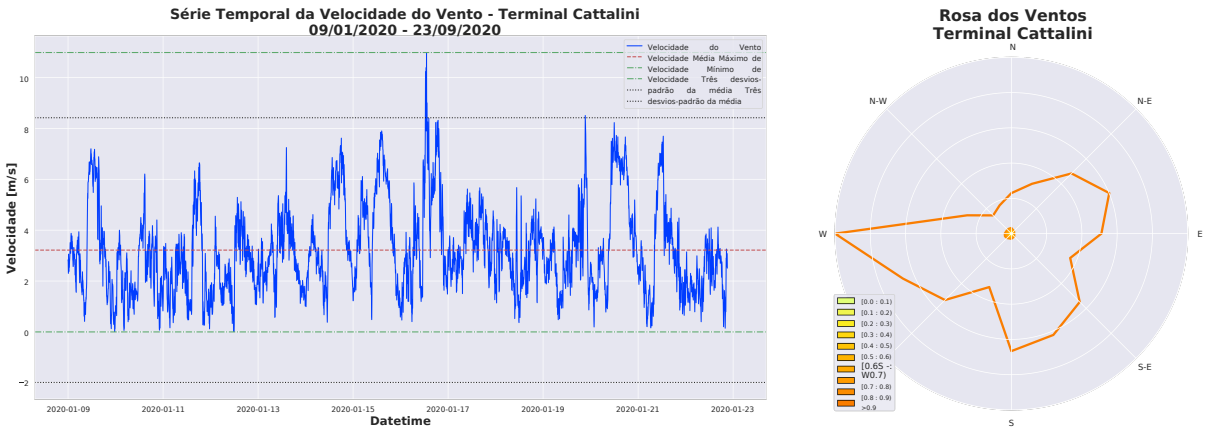


A esquerda temos uma série temporal e à direita um diagrama polar, ambas da velocidade dos ventos.

Imagem criada autorialmente

bases oferece informações detalhadas e complementares sobre as condições oceânicas e atmosféricas da região, capturadas por diferentes sensores ao longo do tempo.

Figura 19 – Análise dos Dados de Meteorologia — Terminal Cattalini



A esquerda temos uma série temporal e à direita um diagrama polar, ambas da velocidade dos ventos.

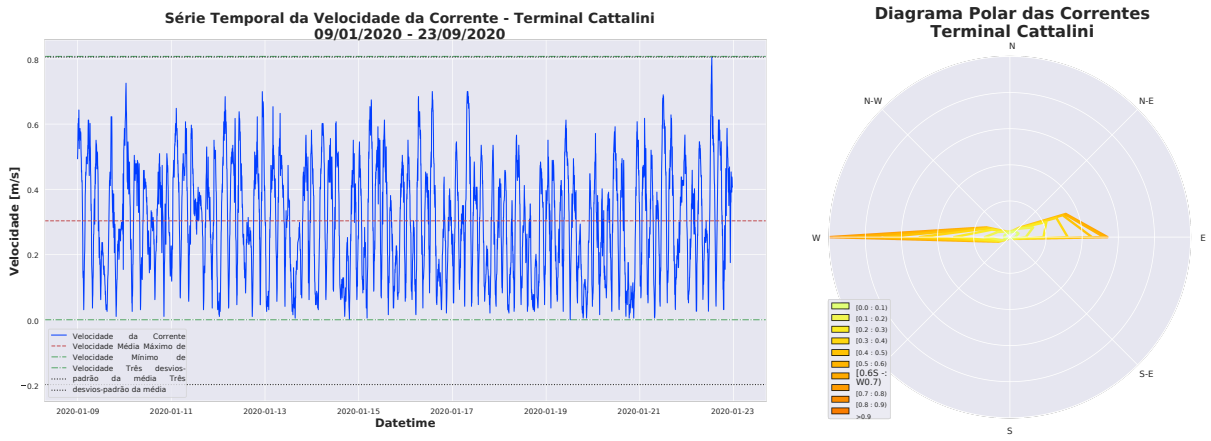
Imagem criada autorialmente

A base meteorológica inclui as variáveis de velocidade dos ventos, rajada dos ventos e direção dos ventos. Essas informações são cruciais para a análise da interação entre os sistemas atmosféricos e oceânicos, auxiliando na caracterização de padrões climáticos locais. A Figura 19 apresenta o diagrama polar que ilustra a distribuição direcional dos ventos registrados nessa localidade, destacando a intensidade e a predominância de determinadas

direções ao longo do período analisado, além disso, a imagem também mostra a série temporal da velocidade do vento para o mesmo período visto anteriormente. Vale ressaltar que, analogamente ao exemplo anterior com a Boia ODAS.

A base de correntes é composta por medições de velocidade e direção em diferentes profundidades, distribuídas da mesma forma que o correntômetro da Boia ODAS. Esses dados permitem uma análise vertical detalhada da dinâmica das correntes marinhas, a Figura 20 apresenta o diagrama polar das correntes, destacando como as direções e intensidades variam ao longo do tempo e entre as diferentes camadas de profundidade e também uma faixa de tempo da série temporal.

Figura 20 – Análise dos Dados de Correntes — Terminal Cattalini



A esquerda temos uma série temporal e à direita um diagrama polar, ambas da velocidade das correntes.

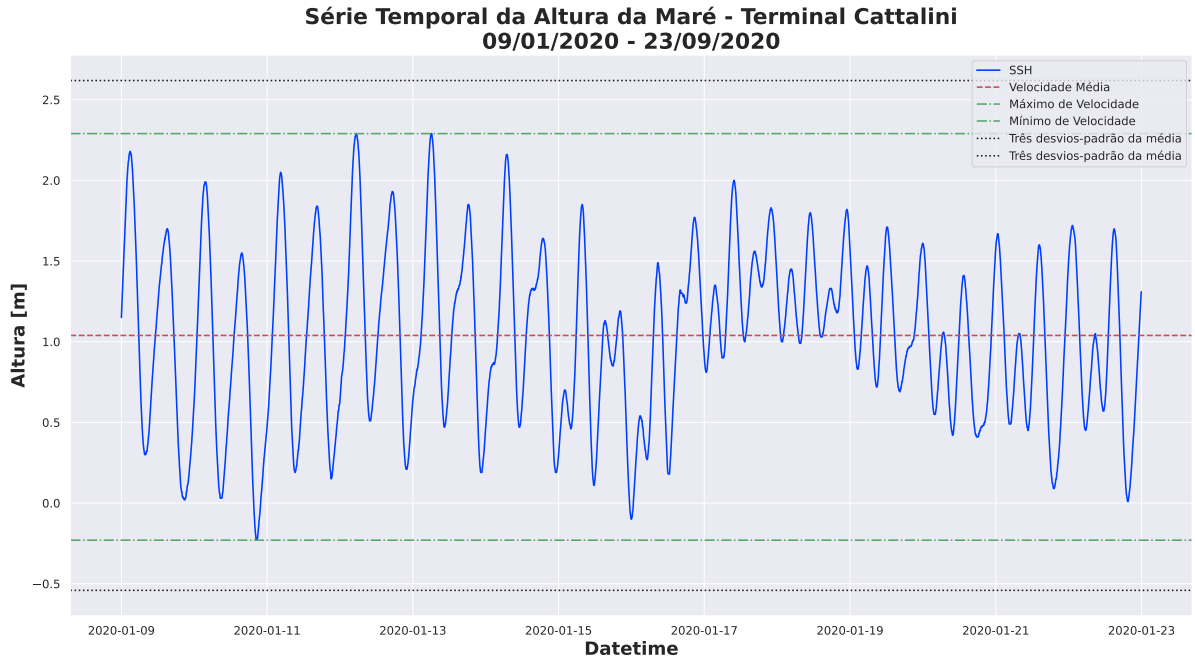
Imagem criada autorialmente

Por fim, o marégrafo fornece a altura da superfície do mar (*Sea Surface Height*, SSH) em metros, que é uma variável crítica para a análise do comportamento das marés e para a identificação de eventos extremos, como tempestades e ressacas. A série temporal da SSH coletada no terminal da Cattalini é apresentada na Figura 21, evidenciando variações características ao longo do período estudado.

4.3 Porto de Paranaguá

Os sensores instalados no Porto de Paranaguá fornecem um conjunto de dados diversificado, composto por três bases principais: medições do nível do mar capturadas por um marégrafo, além de valores calculados para a maré harmônica e a maré astronômica. Esses três tipos de dados são fundamentais para modelar as dinâmicas oceânicas da região, sendo descritos em detalhes a seguir.

Figura 21 – Análise dos Dados de Altura de Maré — Terminal Cattalini



Variação da *Sea Surface Height* (SSH) ao longo do tempo.

Imagem criada autorialmente

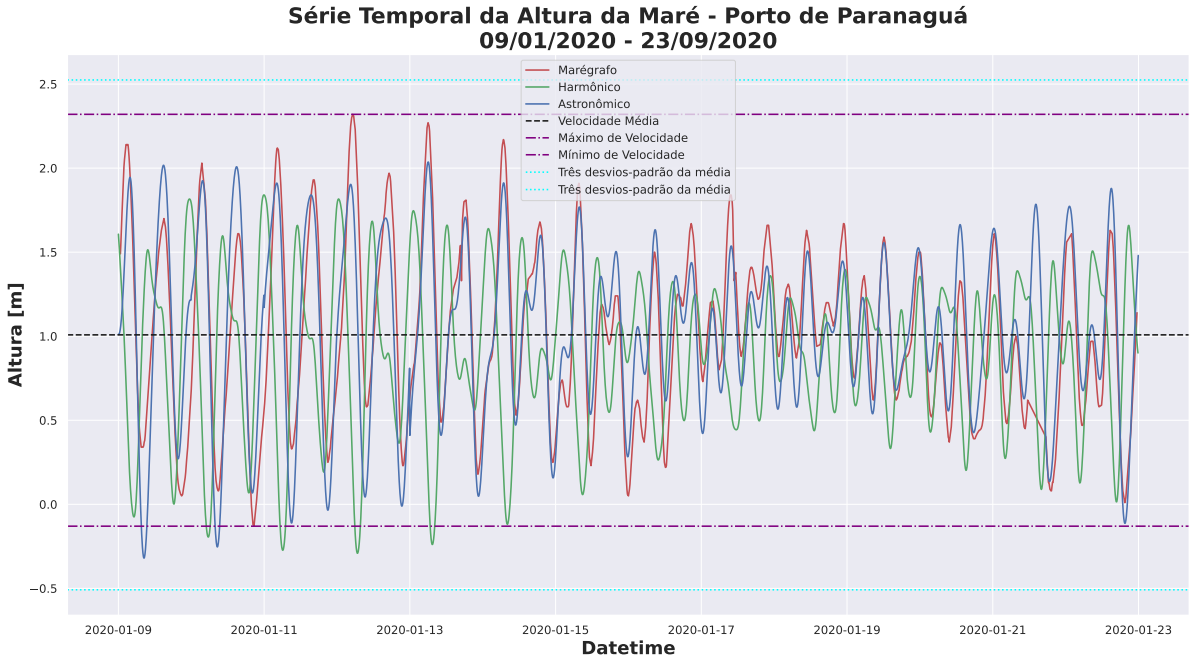
O marégrafo é responsável por medir o SSH em intervalos regulares, fornecendo dados diretamente observados sobre as variações no nível do mar ao longo do tempo. Esses dados, no entanto, estão sujeitos a lacunas e erros instrumentais, que podem ser causados por falhas no equipamento ou condições ambientais adversas.

Em complemento, os valores de maré harmônica e astronômica são calculados a partir de modelos físicos. A maré harmônica é obtida pela decomposição de séries históricas do nível do mar em componentes periódicas, enquanto a maré astronômica é estimada teoricamente com base nas forças gravitacionais exercidas por corpos celestes, como o Sol e a Lua. Diferentemente das medições do marégrafo, essas duas variáveis são consideradas covariáveis conhecidas, o que significa que o modelo de previsão tem acesso a seus valores futuros no instante de referência t_ϕ. Essa característica é particularmente útil para a tarefa de *forecasting*, ao fornecer informações adicionais que podem auxiliar na precisão das previsões.

Embora as marés harmônicas e astronômicas não apresentem falhas, como lacunas ou erros instrumentais, suas estimativas podem ter limitações em função de simplificações nos modelos ou da falta de representatividade para condições locais específicas. Já os dados medidos pelo marégrafo frequentemente enfrentam desafios relacionados à continuidade e à qualidade das medições. A Figura 22 ilustra as diferenças e a complementaridade entre

os dados calculados e os observados.

Figura 22 – Análise dos Dados de Correntes — Terminal Cattalini



Comparativo visual das três medições de SSH (marégrafo, maré harmônica e maré astronômica) no Porto de Paranaguá.

Imagem criada autorialmente

4.4 Irregularidades e Características do Dataset

O conjunto de dados utilizado neste estudo apresenta diversas irregularidades que tornam inviável a aplicação de técnicas convencionais de regularização. Dados coletados nas três localidades — Boia ODAS, Terminal da Cattalini e Porto de Paranaguá — exibem características distintas e comportamentos heterogêneos que, se regularizados, poderiam introduzir vieses significativos e comprometer a qualidade das análises.

Regularizar o dataset, isto é, alinhar os dados em intervalos de tempo fixos ou preencher lacunas com valores imputados, pode resultar em várias limitações. Primeiramente, a imputação excessiva de dados ausentes tende a mascarar padrões reais, criando informações artificiais que não refletem as dinâmicas naturais do sistema estudado. Isso adiciona viés ao modelo, especialmente em séries temporais altamente dependentes, onde os valores imputados podem distorcer tendências e sazonalidades. Além disso, ao forçar uma homogeneidade inexistente, a regularização ignora a natureza intrínseca do problema, reduzindo a capacidade dos modelos preditivos de lidar com dados reais e irregulares, como aqueles frequentemente encontrados em aplicações práticas de *IMTS-forecasting*.

Localização	Sensor	Frequência Estimada	Faltantes
Bóia ODAS	Correntômetro	10 minutos	18.004%
	Meteorológico	5 minutos	24.023%
Cattalini	Correntômetro	10 minutos	0.945%
	Meteorológico	1 minuto	56.892%
	Marégrafo	5 minutos	0.953%
Porto de Paranaguá	Marégrafo	5 minutos	25.836%
	Maré Astronômica	5 minutos	0.000%
	Harmônicos	10 minutos	0.000%

Tabela 2 – Qualidade dos dados coletados nas diferentes localizações.

A Tabela 2 apresenta a porcentagem de dados faltantes no *dataset*, considerando uma frequência estimada de amostragem definida com base em uma análise visual das linhas e defasagens de tempo presentes nas tabelas. É importante destacar que essas porcentagens refletem a comparação com um cenário ideal regularizado, que não corresponde à realidade dos dados observados. Regularizar o *dataset* neste contexto resultaria em um aumento substancial de valores ausentes (`NaNs`), introduzindo vieses significativos e comprometendo a representatividade das dinâmicas naturais das séries temporais.

Figura 23 – Faltas no P^{2}MOD

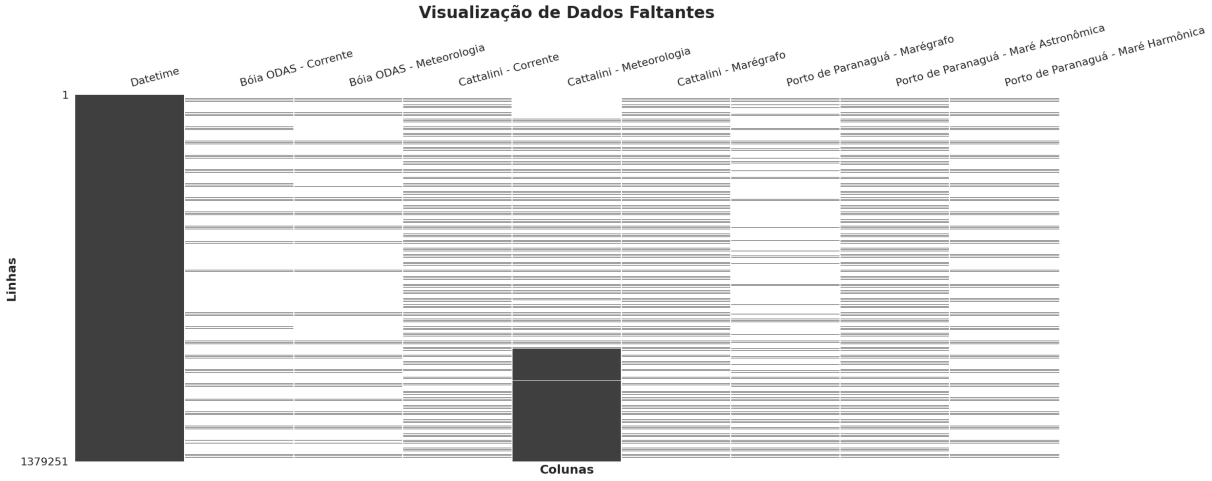


Ilustração mostrando as faltas no *dataset* regularizado, observe que os tons escuros são os dados presentes e estão referenciados à menor frequência estimada de `1 minuto`.

Imagem criada autorialmente

O período de coleta do dataset abrange medições de `2019-04-25 19:30:00` até `2021-12-08 15:00:00`. Embora o intervalo seja relativamente curto, ele captura informações valiosas para a modelagem oceânica e atmosférica. No entanto, nem todos

os sensores possuem medições completas para todo o período, isso é explicitado na Figura 23, resultando em frequentes lacunas nos dados. Essas lacunas, se preenchidas de maneira artificial, poderiam comprometer a fidelidade das análises e diminuir a capacidade do modelo de lidar com as irregularidades naturais do conjunto.

Dessa forma, este trabalho adota uma abordagem que respeita a irregularidade dos dados e evita a regularização excessiva, explorando técnicas avançadas para modelar diretamente as dinâmicas temporais irregulares. Esse método não apenas mantém a integridade dos dados originais, mas também oferece uma base mais robusta para a análise e previsão em séries temporais multivariadas irregulares.

5 Finalização

Neste capítulo, apresentamos o setup experimental e os resultados obtidos durante os testes realizados com o dataset **P²MOD**. O principal objetivo dos experimentos foi demonstrar que considerar as irregularidades inerentes ao dataset, ao invés de tratá-las por meio de regularização, pode ser significativamente mais eficiente para a tarefa de previsão em séries temporais multivariadas irregulares (*IMTS-forecasting*). Para isso, foram implementados e avaliados dois modelos distintos: uma GRU *standard* e o modelo *Gap-Ahead* com *Time Encoding*.

5.1 Setup Experimental

O experimento foi conduzido utilizando exclusivamente as primeiras componentes de correntes, velocidade do vento e altura do nível do mar das bases de dados coletadas. As demais variáveis foram descartadas visando simplificar o problema e focar nas características principais do dataset. Para evitar problemas relacionados à descontinuidade angular, as variáveis direcionais (e.g., direção do vento e direção das correntes) foram convertidas em suas representações trigonométricas, ou seja, valores de seno e cosseno. Essa transformação garante uma representação contínua e consistente para os modelos preditivos, além disso, a *feature target* foi considerada somente o SSH do marégrafo do Porto de Paranaguá. Os dois modelos utilizados no experimento foram configurados da seguinte forma:

- **GRU-standard:** Um modelo simples baseado em *Gated Recurrent Units*, utilizado como *baseline* para avaliar o desempenho de um modelo tradicional em séries temporais. Esse modelo não possui mecanismos avançados para lidar diretamente com irregularidades nos dados, dependendo da estrutura da série temporal para capturar dinâmicas temporais.
- **Gap-Ahead com Time Encoding:** Um modelo mais sofisticado que incorpora codificação temporal e redes de atenção baseadas em grafos. Esse modelo foi projetado especificamente para lidar com dados irregulares, permitindo a captura de padrões mais complexos.

Ambos os modelos foram treinados e avaliados utilizando o *Index of Agreement* (IoA), uma métrica robusta para comparar a similaridade entre as previsões dos modelos e os valores reais observados (WILLMOTT, 1981; BARROS et al., 2024; NETTO et al., 2022). Definido como:

$$\mathcal{L}(\hat{y}_i, y_i) = 1 - IoA = 1 - \frac{\sum_{n=1}^{N}(y_i - \hat{y}_i)}{\sum_{n=1}^{N}(|\hat{y}_i - \bar{y}_i)| + |y_i - \bar{y}_i|)^2} \tag{5.1}$$

5.2 Resultados

Os resultados obtidos demonstram diferenças significativas entre os dois modelos. A GRU-standard apresentou desempenho inferior. Esse comportamento pode ser atribuído à incapacidade do modelo de lidar diretamente com lacunas temporais e informações heterogêneas, o que comprometeu sua capacidade de generalização, obtendo um valor de $IoA_{GRU} = 0.4452425$.

Modelo	**IoA**
GRU	0.4452425
Gap-Ahead	0.1027102

Tabela 3 – *Index of Agreement* dos modelos treinados.

Por outro lado, o modelo *Gap-Ahead* com *Time Encoding* mostrou-se altamente eficaz em capturar as dinâmicas subjacentes do dataset, mesmo diante das irregularidades, obtendo um valor de $IoA_{Gap-Ahead} = 0.1027102$. Esse resultado reforça a hipótese de que tratar os dados em seu estado original, sem tentar regularizá-los, pode oferecer vantagens significativas em termos de precisão preditiva. A abordagem do *Gap-Ahead* demonstra que a integração de técnicas avançadas, como codificação temporal e redes de atenção, permite explorar de maneira mais profunda a estrutura dos dados irregulares.

A Figura 24 apresenta os valores reais e previstos para os dois modelos ao longo do conjunto de testes. Observa-se que o *Gap-Ahead* supera consistentemente a GRU-standard, confirmando sua superioridade em cenários que envolvem dados com irregularidades.

5.3 Contribuição

Os experimentos realizados reforçam a relevância do dataset P^2MOD como um benchmark robusto para tarefas de *IMTS-forecasting*. Além disso, demonstram que abordagens que consideram as irregularidades dos dados, ao invés de tratá-las por regularização, são mais adequadas para explorar a complexidade das séries temporais multivariadas irregulares. O uso de técnicas avançadas, como no modelo *Gap-Ahead*, oferece uma contribuição significativa para a área, mostrando que é possível alcançar maior precisão preditiva sem comprometer a integridade dos dados originais.

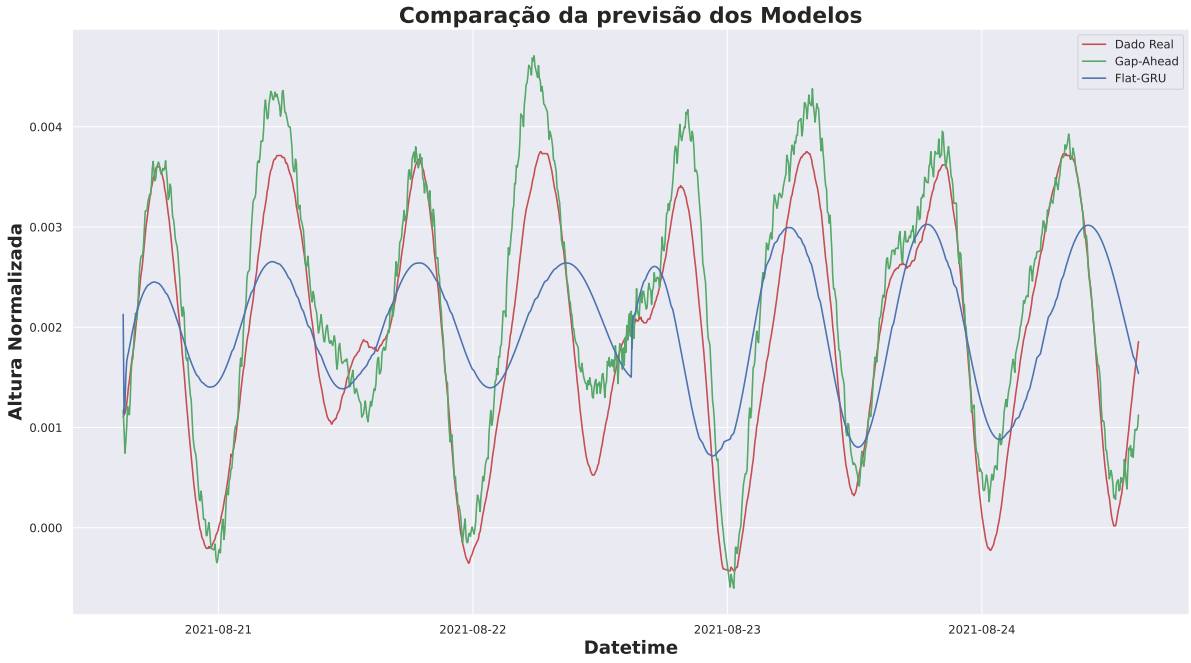


Figura 24 – Comparação dos valores SSH reais e previstos pelos modelos.

5.4 Conclusão

Os experimentos apresentados neste capítulo demonstraram claramente a importância de lidar com as irregularidades dos dados diretamente, sem recorrer a regularizações que introduzam vieses significativos. O uso do modelo *Gap-Ahead* com *Time Encoding* provou ser uma abordagem eficiente para explorar a estrutura intrínseca dos dados irregulares, superando o desempenho da GRU-standard em todas as métricas analisadas.

Além disso, os resultados destacam a relevância do dataset **P²MOD** como um benchmark para a tarefa de previsão em séries temporais multivariadas irregulares (*IMTS-forecasting*). A riqueza de variáveis, combinada com as características naturais do dataset, fornece um cenário realista e desafiador para avaliar modelos preditivos.

A análise comparativa entre os modelos demonstrou que considerar as irregularidades como parte integrante do problema, em vez de tratá-las como ruídos a serem eliminados, é uma estratégia mais eficaz. Essa abordagem não apenas preserva a integridade dos dados originais, mas também oferece percepções mais ricos sobre as dinâmicas das séries temporais, contribuindo para avanços significativos na área de previsão.

Os resultados obtidos neste estudo fornecem uma base sólida para futuras pesquisas, incentivando o desenvolvimento de novas técnicas que consigam lidar de maneira mais eficiente com as complexidades das séries temporais irregulares. Além disso, reafirmam o papel do P²MODcomo um recurso essencial para a validação de métodos inovadores em diferentes contextos aplicados.

Referências

BARBER, D. *Bayesian reasoning and machine learning*. [S.l.]: Cambridge University Press, 2012. Citado 4 vezes nas páginas 20, 21, 22 e 24.

BARROS, M. et al. Early detection of extreme storm tide events using multimodal data processing. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 38, n. 20, p. 21923–21931, mar. 2024. ISSN 2374-3468. Number: 20. Disponível em: ⟨https://ojs.aaai.org/index.php/AAAI/article/view/30194⟩. Citado 7 vezes nas páginas 26, 29, 30, 35, 36, 37 e 51.

BISHOP, C. M. et al. *Pattern recognition and machine learning*. [S.l.]: Springer, 2006. v. 4. Citado 4 vezes nas páginas 20, 21, 22 e 24.

CHO, K. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. Citado na página 33.

COOKE, M. et al. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, v. 120, n. 5, p. 2421–2424, nov. 2006. ISSN 0001-4966. Disponível em: ⟨https://doi.org/10.1121/1.2229005⟩. Citado 2 vezes nas páginas 38 e 39.

DU, S. et al. Multivariate time series forecasting via attention-based encoder–decoder framework. *Neurocomputing*, v. 388, p. 269–279, maio 2020. ISSN 09252312. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S0925231220300606⟩. Citado na página 26.

HALEVY, A. et al. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, v. 24, n. 2, p. 8–12, mar. 2009. ISSN 1941-1294. Citado na página 21.

HEBRAIL, A. B. G. *Individual Household Electric Power Consumption*. UCI Machine Learning Repository, 2006. Disponível em: ⟨https://archive.ics.uci.edu/dataset/235⟩. Citado na página 15.

HUANG, S. et al. Temporal graph benchmark for machine learning on temporal graphs. *Advances in Neural Information Processing Systems*, v. 36, 2024. Citado na página 39.

IBGE |*Biblioteca* − − − *biblioteca.ibge.gov.br*.2017.¡¿. [Accessed 24-09-2024]. Citado na página 14.

JOHNSON, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Scientific Data*, v. 3, n. 1, p. 160035, 2016. ISSN 2052-4463. Publisher: Nature Publishing Group. Disponível em: ⟨https://www.nature.com/articles/sdata201635⟩. Citado 2 vezes nas páginas 38 e 39.

LIU, X. et al. Deep time series forecasting models: A comprehensive survey. *Mathematics*, v. 12, n. 10, p. 1504, maio 2024. ISSN 2227-7390. Disponível em: ⟨https://www.mdpi.com/2227-7390/12/10/1504⟩. Citado na página 15.

MAKRIDAKIS, S. et al. Statistical and machine learning forecasting methods: Concerns and ways forward. *Plos One*, Public Library of Science, v. 13, n. 3, p. e0194889, mar. 2018. ISSN 1932-6203. Disponível em: ⟨https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889⟩. Citado na página 26.

MARLIN, B. M. A survey on principles, models and methods for learning from irregularly sampled time series. *arXiv preprint arXiv:2012.00168*, arXiv, 2020. Disponível em: ⟨https://arxiv.org/abs/2012.00168⟩. Citado 2 vezes nas páginas 28 e 29.

MARTIN, S. *An introduction to ocean remote sensing*. [S.l.]: Cambridge University Press, 2014. Citado na página 14.

MENNE, M. J. et al. *Long-Term Daily and Monthly Climate Records from Stations Across the Contiguous United States (U.S. Historical Climatology Network)*. [S.l.], 2016. Disponível em: ⟨https://www.osti.gov/dataexplorer/biblio/dataset/1394920⟩. Citado 2 vezes nas páginas 38 e 39.

MOHRI, M. et al. *Foundations of machine learning*. Cambridge, MA: MIT Press, 2012. (Adaptive computation and machine learning series). ISBN 978-0-262-01825-8. Citado 4 vezes nas páginas 20, 21, 22 e 24.

MUDELSEE, M. Tauest: a computer program for estimating persistence in unevenly spaced weather/climate time series. *Computers & Geosciences*, v. 28, n. 1, p. 69–72, fev. 2002. ISSN 00983004. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S0098300401000413⟩. Citado na página 14.

NETTO, C. F. D. et al. Modeling oceanic variables with dynamic graph neural networks. arXiv, 2022. Disponível em: ⟨https://arxiv.org/abs/2206.12746⟩. Citado na página 51.

REISS, A. et al. Introducing a new benchmarked dataset for activity monitoring. In: *2012 16th International Symposium on Wearable Computers*. Newcastle, United Kingdom: Ieee, 2012. p. 108–109. ISBN 978-0-7695-4697-1. Disponível em: ⟨http://ieeexplore.ieee.org/document/6246152/⟩. Citado 3 vezes nas páginas 15, 38 e 39.

REYES-ORTIZ, D. A. J. *Human Activity Recognition Using Smartphones*. UCI Machine Learning Repository, 2013. Disponível em: ⟨https://archive.ics.uci.edu/dataset/240⟩. Citado na página 15.

REYNA, M. A. et al. Early prediction of sepsis from clinical data: The PhysioNet/computing in cardiology challenge 2019. *Critical Care Medicine*, v. 48, n. 2, p. 210–217, 2020. ISSN 0090-3493. Disponível em: ⟨https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6964870/⟩. Citado na página 39.

SCHWARZACHER, W. An application of statistical time-series analysis of a limestone-shale sequence. *The Journal of Geology*, v. 72, n. 2, p. 195–213, mar. 1964. ISSN 0022-1376, 1537-5269. Disponível em: ⟨https://www.journals.uchicago.edu/doi/10.1086/626976⟩. Citado na página 14.

SCOTT, G. J. et al. Htidb: Hierarchical time-indexed database for efficient storage and access to irregular time-series health sensor data. In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. [S.l.]: Ieee, 2022. p. 2972–2975. Citado na página 14.

SEZER, O. B. et al. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, v. 90, p. 106181, maio 2020. ISSN 15684946. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S1568494620301216⟩. Citado na página 14.

SHIH, S.-Y. et al. Temporal pattern attention for multivariate time series forecasting. *Machine Learning*, v. 108, n. 8–9, p. 1421–1441, set. 2019. ISSN 0885-6125, 1573-0565. Disponível em: ⟨http://link.springer.com/10.1007/s10994-019-05815-0⟩. Citado na página 26.

SILVA, I. et al. Predicting in-hospital mortality of ICU patients: The PhysioNet/computing in cardiology challenge 2012. *Computing in cardiology*, v. 39, p. 245–248, 2012. Disponível em: ⟨https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965265/⟩. Citado 2 vezes nas páginas 38 e 39.

SMITH, S. R. et al. Ship-based contributions to global ocean, weather, and climate observing systems. *Frontiers in Marine Science*, Frontiers Media SA, v. 6, p. 434, 2019. Citado na página 14.

SUVRIT et al. *Optimization for machine learning*. Cambridge (Mass.): MIT press, 2012. (Neural information processing series). ISBN 978-0-262-01646-9. Citado 2 vezes nas páginas 21 e 24.

ULABY, F. et al. *Signals & Systems: Theory and Applications*. Michigan Publishing, 2018. ISBN 978-1-60785-487-6. Disponível em: ⟨https://books.google.com.br/books?id=juk4zgEACAAJ⟩. Citado 2 vezes nas páginas 24 e 25.

VASWANI, A. et al. Attention is all you need. arXiv, v. 30, 2017. Disponível em: ⟨https://arxiv.org/abs/1706.03762⟩. Citado 2 vezes nas páginas 35 e 36.

VELIcKOVIć, P. et al. Graph attention networks. arXiv, 2017. Disponível em: ⟨https://arxiv.org/abs/1710.10903⟩. Citado na página 37.

VIO, R. et al. Irregular time series in astronomy and the use of the lomb–scargle periodogram. *Astronomy and Computing*, Elsevier, v. 1, p. 5–16, 2013. Citado na página 14.

VITO, S. *Air Quality*. UCI Machine Learning Repository, 2008. Disponível em: ⟨https://archive.ics.uci.edu/dataset/360⟩. Citado na página 15.

VOOGT, J. A. et al. Thermal remote sensing of urban climates. *Remote sensing of environment*, Elsevier, v. 86, n. 3, p. 370–384, 2003. Citado na página 14.

WANG, X. et al. Heterogeneous graph attention network. In: *The World Wide Web Conference*. San Francisco CA USA: Acm, 2019. p. 2022–2032. ISBN 978-1-4503-6674-8. Disponível em: ⟨https://dl.acm.org/doi/10.1145/3308558.3313562⟩. Citado na página 37.

WEI, W. W. *Multivariate time series analysis and applications*. [S.l.]: John Wiley & Sons, 2018. 528 p. Citado 4 vezes nas páginas 24, 25, 26 e 27.

WILLMOTT, C. J. On the Validation of Models. *Physical Geography*, v. 2, n. 2, p. 184–194, jul. 1981. ISSN 0272-3646. Disponível em: ⟨https://doi.org/10.1080/02723646.1981.10642213⟩. Citado 2 vezes nas páginas 42 e 51.

XIAO, J. et al. Ivp-vae: Modeling ehr time series with initial value problem solvers. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 38, n. 14, p. 16023–16031, mar. 2024. ISSN 2374-3468, 2159-5399. Number: 14. Disponível em: ⟨https://ojs.aaai.org/index.php/AAAI/article/view/29534⟩. Citado na página 15.

YALAVARTHI, V. K. et al. Grafiti: Graphs for forecasting irregularly sampled time series. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 38, n. 15, p. 16255–16263, mar. 2024. ISSN 2374-3468, 2159-5399. Number: 15. Disponível em: ⟨https://ojs.aaai.org/index.php/AAAI/article/view/29560⟩. Citado na página 15.

ZHANG, A. et al. Monitoring meteorological drought in semiarid regions using multi-sensor microwave remote sensing data. *Remote sensing of Environment*, Elsevier, v. 134, p. 12–23, 2013. Citado 2 vezes nas páginas 14 e 39.

ZHOU, H. et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 35, n. 12, p. 11106–11115, maio 2021. ISSN 2374-3468, 2159-5399. Number: 12. Disponível em: ⟨https://ojs.aaai.org/index.php/AAAI/article/view/17325⟩. Citado na página 15.