Predição de nível de rios da Região Hidrográfica do Rio Paraguai utilizando algoritmos de Aprendizado de Máquina

Rogério Alves dos Santos Antoniassi

SERVIÇO DE PÓS-GRADUAÇÃO FACOM-UFMS

Data de Depósito:

Assinatura:_____

Predição de nível de rios da Região Hidrográfica do Rio Paraguai utilizando algoritmos de Aprendizado de Máquina

Rogério Alves dos Santos Antoniassi

Orientador: *Prof. Dr. Renato Porfirio Ishii*

Dissertação apresentada ao programa de Pós-Graduação, Mestrado Acadêmico em Ciência da Computação, da Universidade Federal de Mato Grosso do Sul - UFMS, como parte dos requisitos necessários à obtenção do título de Mestre em Ciência da Computação.

UFMS - Campo Grande

Março/2022

Aos meus pais,
Dilce e Flávio,

À minha irmã,
Patrícia,

À minha esposa,
Letícia.

Agradecimentos

A Deus, por sempre estar presente em minha vida e por guiar meus passos.

Aos meus pais, Flávio e Dilce, à minha irmã, Patrícia, e minha esposa, Letícia, por todo amor, dedicação, paciência e compreensão nos momentos de ausência para a realização do presente estudo, meu eterno muito obrigado!

Meu agradecimento especial ao orientador, Renato Ishii, que me ensinou não só a pesquisa, mas também lições que levarei para a vida. Sempre disposto, me motivando e tentando me mostrar os caminhos a seguir. Obrigado pela paciência, por nunca me deixar desistir e pela dedicação nas orientações.

Ao pesquisador da Embrapa Pantanal, Carlos Padovani, pela parceria no projeto, amizade, por sempre se mostrar disponível, independentemente do horário, por me motivar e pela paciência em suas explicações.

A todos da Embrapa Pantanal, Embrapa Sudeste e INPE, por toda a ajuda e ensinamentos com as configurações do TerraMA2.

Ao professor Rafael Geraldeli Rossi, da UFMS, pelas explicações nos momentos de dúvida. Aos professores, Eduardo Mario Mendiondo e Alexandre Delbem, da USP Campus São Carlos e à Maria Clara Fava, pelas contribuições e ensinamentos tanto na área de computação, quanto na área de hidrologia.

Aos amigos de laboratório da pós-graduação: Cleiton, Pedro, Lucas, Adriano e Fabrício.

A todos os funcionários e servidores da UFMS e FACOM.

À Thalita Ribeiro, pela contribuição com seu projeto de pesquisa.

Ao IFMS, pelo afastamento concedido, que foi de extrema importância para a realização do trabalho. Aos amigos do IFMS, pelo apoio e parceria de sempre, em especial ao Douglas Toledo, Marco Aurélio Ferreira e Alex Araujo.

Abstract

The Pantanal is part of the Paraguay River Hydrographic Region (RH-Paraguay) and has, as a characteristic, the flooding of a portion of its area at certain times of the year. This flood generates a series of socioeconomic problems for the population living in its surroundings. These problems can be mitigated when their occurrence is predicted in advance. In this sense, this work investigates the application of Machine Learning (ML) techniques for the prediction of river levels in RH-Paraguay, using daily data from upstream stations to predict the level of downstream stations. From this perspective, modeling with ML techniques proves to be effective in predictions, as similar works appear in the related literature. In this way, a data source with daily level values is used, and a sample of three stations is selected. Then, the time lag between the flood peak of one station and another was removed, searching the ideal hyperparameters for the Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU) and Bidirectional Long Short Term Memory (BiLSTM) networks and submitted to the training process. Subsequently, the models with the best results of each algorithm were selected, which were compared with the Regression technique currently used. The results show that the three models tested can be used for the prediction task, in which the three present improvements in relation to the current model. The model with the GRU algorithm stood out for presenting the lowest error rates and for being 23.84% more accurate than the Regression model, while LSTM and BiLSTM are, respectively, 18.09% and 19.16% more accurate than the Regression model. The LSTM and BiLSTM models are closer to the real value in the peaks of maximum and minimum levels, when compared to the Regression and the GRU.

Keywords: Machine Learning, Time Series, Prediction, Forecast, River Level.

Resumo

O Pantanal faz parte da Região Hidrográfica do Rio Paraguai (RH-Paraguai) e possui, como característica, a inundação de uma parcela de sua área em determinadas épocas do ano. Essa inundação gera uma série de problemas socioeconômicos à população que vive em seu entorno. Problemas estes que podem ser mitigados quando há previsão de sua ocorrência com antecedência. Nesse sentido, este trabalho investiga a aplicação de técnicas de Aprendizado de Máquina (AM) para a predição de níveis dos rios da RH-Paraguai, utilizando dados diários de estações a montante para previsão do nível de estações a jusante. Nessa perspectiva, a modelagem com técnicas de AM demonstra-se eficaz nas predições, na medida que trabalhos similares aparecem na literatura correlata. Dessa maneira, é utilizada uma fonte de dados com valores de níveis diários, sendo selecionada uma amostra de três estações. Em seguida, foi retirada a defasagem do tempo entre o pico de cheia de uma estação e outra, pesquisados os hiperparâmetros ideais para as redes Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU) e Bidirectional Long Short Term Memory (BiLSTM) e submetidos ao processo de treinamento. Posteriormente, foram selecionados os modelos de melhores resultados de cada algoritmo, os quais foram comparados com a técnica de Regressão utilizada atualmente. Os resultados mostram que os três modelos testados podem ser utilizados para a tarefa de predição, em que, os três apresentam melhorias em relação ao modelo atual. O modelo com o algoritmo GRU se destacou por apresentar as menores taxas de erro e por ser $23,84\%$ mais preciso que o modelo de Regressão, já LSTM e BiLSTM são, respectivamente, $18,09\%$ e $19,16\%$ mais precisos que o modelo de Regressão. Os modelos de LSTM e BiLSTM se aproximam mais do valor real nos picos de níveis máximos e mínimos, quando

comparados ao de Regressão e ao GRU.

Palavras-chave: Aprendizado de Máquina, Séries Temporais, Predição, Previsão, Nível de Rio.

Sumário

Abstract i
Resumo ii
Sumário iv
Lista de Figuras vi
Lista de Tabelas viii
Lista de Abreviaturas ix

1 Introdução **1**
1.1 Objetivos 5
1.2 Estrutura do documento 5

2 Referencial Teórico **7**
2.1 Séries temporais 7
2.2 Aprendizado de máquina 8
2.2.1 Deep Learning 10
2.2.2 *Long Short Term Memory* 14
2.2.3 *Bidirectional Long Short Term Memory* 16
2.2.4 *Gated Recurrent Unit* 17
2.3 Métricas de desempenho 17
2.4 Considerações Finais 19

3 Trabalhos Relacionados **20**
3.1 Revisão Sistemática da Literatura 20
3.1.1 Planejamento 20
3.1.2 Execução e avaliação 23

3.2 Considerações finais 31

4 Metodologia **33**

4.1 Dados 33

4.1.1 Seleção das Amostras 40

4.1.2 Alinhamento das Séries 43

4.2 Modelagem 46

4.3 *Benchmarking* 50

5 Resultados e Discussão **52**

5.1 Resultado da busca por hiperparâmetros 52

5.2 Resultados do treinamento 55

6 Conclusão **61**

6.1 Dificuldades e principais problemas 62

6.2 Contribuições 62

6.3 Trabalhos futuros 64

Referências **65**

Lista de Figuras

1.1 RH-Paraguai, adaptado de (CPRM, 2006). 2

2.1 Série temporal contínua (Rios, 2013). 8
2.2 Série temporal discreta (Rios, 2013). 8
2.3 Esquema de um neurônio artificial (Buduma and Locascio, 2017). . . 11
2.4 Exemplo de uma *Rede Neural Artificial* (RNA) multicamadas. 12
2.5 Funcionamento de uma LSTM. 16

3.1 Quantidade de trabalhos por ano de publicação. 23

4.1 Estrutura da base de dados disponibilizada. 34
4.2 Estrutura da base de dados convertida. 35
4.3 Diferença de amostras entre as estações. 37
4.4 Dados das estações de São José do Boriréu e Porto Conceição. 38
4.5 Dados da estação de Ladário. 39
4.6 Dados da estação de Porto Esperança. 39
4.7 Quantidade de dados ausentes. 40
4.8 Localização das estações selecionadas. 41
4.9 Fluxo a montante de Ladário, (Padovani, 2010). 42
4.10 Séries selecionadas. 43
4.11 Estações lado a lado. 44
4.12 Estações lado a lado, segundo passo. 44
4.13 Estações lado a lado com defasagem ajustada. 45
4.14 Séries selecionadas após o ajuste. 45
4.15 Exemplo de janelas. 50

5.1 Resultado da Regressão. 59
5.2 Resultado da LSTM. 59
5.3 Resultado da GRU. 59
5.4 Resultado da BiLSTM. 60
5.5 Resultados das predições juntas. 60

Lista de Tabelas

3.1 Quantidade de trabalhos após a aplicação dos critérios. 24
3.2 Informações sobre os trabalhos selecionados. 29
3.3 Referências dos trabalhos selecionados. 29

4.1 Informações sobre a base de dados. 37
4.2 Opções de busca de hiperparâmetros. 48

5.1 Resultados da busca de hiperparâmetros - LSTM. 53
5.2 Resultados da busca de hiperparâmetros - GRU. 54
5.3 Resultados da busca de hiperparâmetros - BiLSTM. 54
5.4 Resultados dos modelos LSTM. 56
5.5 Resultados dos modelos GRU. 57
5.6 Resultados dos modelos BiLSTM. 57
5.7 Comparação com o método atual. 58

Lista de Abreviaturas

AM Aprendizado de Máquina

ANA Agência Nacional de Águas e Saneamento Básico

ANFIS *Adaptive Neuro-Fuzzy Inference System*

ANN *Artificial Neural Network*

ARIMA *Autoregressive Integrated Moving Average*

BiLSTM *Bidirectional Long Short Term Memory*

BPNN *Back Propagation Neural Networks*

CAPES Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CC *Coefficient Of Correlation*

CNN *Convolutional Neural Network*

CPRM Companhia de Pesquisa de Recursos Minerais

DC *Deterministic Coefficients*

DeepCNN *Deep Convolutional Neural Network*

DNN *Deep Neural Network*

DT *Decision Tree*

DTWBI *Dynamic time warping-based imputation*

FCN *Fully Connected Network*

FFBPNN *Feed Forward Back Propagation Neural Network*

FSD *Fraction of Standard Deviation*

GCN *Graph Convolutional Networks*

GRU *Gated Recurrent Unit*

HA *Historical Average*

IA Inteligência Artificial

INPE Instituto Nacional de Pesquisas Espaciais

KNN *K-Nearest Neighbors*

LSTM *Long Short Term Memory*

LSTM-S2S *Sequence-to-Sequence Long Short Term Memory*

MAE *Mean Absolute Error*

MAPE *Mean Absolute Percentage Error*

MedAE *Median Absolute Error*

MLP *MultiLayer Perceptron*

MSE *Mean Square Error*

NSE *Nash-Sutcliffe Efficiency Coefficient*

PI *Persistence Index*

PRE *Peak Runoff Error*

PTE *Peak Time Error*

QR *Ratio Of Qualified Predictions*

R^2 *Coefficient of Determination*

RBF-FA *Sadial Basis Function - Firefly Algorithm*

RBFNN *Radial Basis Function Neural Network*

RF *Random Forest*

RH-Paraguai *Região Hidrográfica do Rio Paraguai*

RMSE *Root Mean Squared Error*

RNA *Rede Neural Artificial*

RNN *Recurrent Neural Network*

SA-LSTM *Apatial Attention LSTM*

Sim *Similarity*

SNIRH Sistema Nacional de Informações sobre Recursos Hídricos

STA-LSTM *Spatio-Temporal Attention Long Short Term Memory*

SVM *Support Vector Machine*

SVM-FA *Support Vector Machine - Firefly Algorithm*

SVR *Support Vector Regression*

TA-LSTM *Temporal Attention LSTM*

WANN *Wavelet Transform Coupled With An Artificial Neural Network*

WSVM *Wavelet Transform Coupled With Support Vector Machine*

CAPÍTULO

1

Introdução

A predição de desastres naturais sempre se mostrou uma importante atividade para a tomada de decisão, permitindo, assim, que ações sejam tomadas com antecedência, objetivando a minimização dos danos, seja em relação a fatores econômicos ou até mesmo salvando vidas.

No caso de recursos hídricos não seria diferente, pois a predição demonstra a situação da região no passar dos dias, o que possibilita a realização de determinadas ações que são de extrema importância, como ocorre em regiões que são atingidas por alagamentos em determinados períodos do ano. O Pantanal, região que faz parte da *Região Hidrográfica do Rio Paraguai* (RH-Paraguai), é uma região que passa por esse processo de alagamento no período de cheias.

Segundo ANA (2018), a RH-Paraguai observada na Figura 1.1 abrange uma área de cerca de 600.000 km^2, tendo 61% de sua área total no território brasileiro, bem como parte nos territórios paraguaio e boliviano. Em sua parcela brasileira inclui uma das maiores extensões de áreas alagadas do planeta. Ainda na Figura 1.1 é possível observar as estações de observação, responsáveis pela aferição do nível dos rios.

O tempo que uma cheia leva para influenciar a cheia em outro rio, ou em outro ponto com estação de observação, varia, pois há outras questões envolvidas, como por exemplo: relevo, intensidade de chuvas, evapotranspiração e a vazão

do rio.

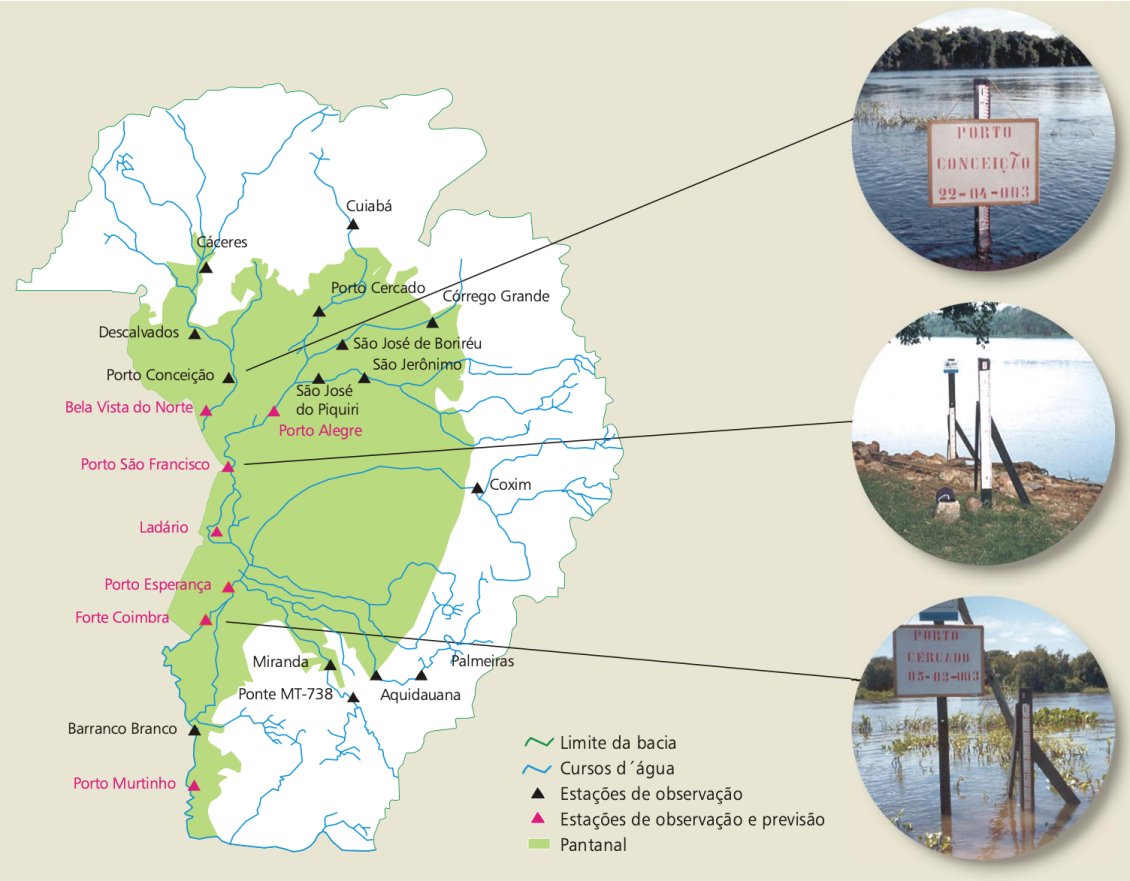

Figura 1.1: RH-Paraguai, adaptado de (CPRM, 2006).

As estações de observação hidrometeorológicas são operadas por entidades parceiras ou contratadas pela Agência Nacional de Águas e Saneamento Básico (ANA), que é a responsável pelo planejamento, normatização de procedimentos e equipamentos, fiscalização, organização dos dados hidrometeorológicos e sua publicação (ANA, 2018). A ANA, que disponibiliza os dados, mantém um portal denominado Portal HidroWeb[1], que é uma ferramenta que faz parte do Sistema Nacional de Informações sobre Recursos Hídricos (SNIRH). Os dados disponíveis nesse Portal são obtidos por meio de coleta de dados a partir de estações convencionais, por um observador e por estações automáticas, em que os dados são enviados por satélite.

A Companhia de Pesquisa de Recursos Minerais (CPRM) é uma das entidades que realiza o monitoramento, previsões e lança boletins semanais com previsões

[1] https://www.snirh.gov.br/hidroweb/apresentacao

para quatro semanas, em formato de gráficos hidrológicos.

De posse dos dados disponibilizados pelo CPRM, pela ANA e pela Marinha do Brasil, a Embrapa Pantanal realiza um importante trabalho para o qual utiliza a base de dados de nível para predições de cheias e estiagens nos rios da RH-Paraguai.

A compilação dos dados e as predições são realizadas de maneira manual em planilhas eletrônicas a partir de modelos de regressões lineares e não lineares, utilizando os dados de máximas e mínimas anuais, com horizonte de predição de um um pico de máxima e mínima. Essas previsões são feitas comparando uma ou mais estações de montante, como variáveis independentes, para uma estação de jusante, como variável dependente. Os resultados são disponibilizados à população por meio de entrevistas em mídias de notícias e uma página de rede social denominada Geohidro-Pantanal[2].

Essas predições são importantes para as comunidades que vivem próximas aos rios da RH-Paraguai, pois, em períodos de cheias, a população acaba por ficar ilhada em alguns pontos, e pecuaristas precisam retirar o seu rebanho das propriedades atingidas pela cheia. As cheias impactam, diretamente, no fator econômico, mudando a rotina das comunidades, pois dificulta a locomoção e o acesso a itens fundamentais, como médicos e escolas; em alguns casos os municípios chegam a decretar situação de emergência.

As previsões permitem que ações sejam tomadas com antecedência, na tentativa de mitigar os danos causados por essas cheias, como a retirada e a realocação de rebanhos. Para a retirada deve-se analisar a melhor rota e planejar como será efetuado o transporte, visto que, em alguns pontos, o gado deve seguir à pé devido a dificuldade de acesso, o que causa, por exemplo, o emagrecimento do gado, o que irá influenciar, diretamente, na sua venda. Nos períodos de estiagem, os alertas atendem, principalmente, à navegação de grandes barcaças de transporte de minérios.

Nesse contexto, a aplicação de algoritmos de Aprendizado de Máquina (AM), tem sido utilizado para predições, pois trabalhos aplicados em bacias e rios distintos utilizaram algoritmos de AM para a realização dessa tarefa.

Dois trabalhos recentes realizam a revisão na literatura e discutem sobre a utilização de algoritmos de AM para predição de nível de rio e inundações.

[2] https://www.facebook.com/geohidropantanal/

- Ghorpade et al. (2021) caracterizam os dados históricos de parâmetros climáticos como sendo, cada vez mais, úteis para a tarefa de previsão. Os autores demonstram os avanços recentes, o desempenho e a precisão de alguns algoritmos mais citados na literatura. Entre os algoritmos discutidos, os autores destacam que *Long Short Term Memory* (LSTM) tem apresentado bons resultados e está sendo muito utilizado para a tarefa de previsão de inundações;

- Dubey and Katarya (2021) realizam revisão bibliográfica sobre tecnologias de AM mais populares para detecção e previsão de inundações, na qual fica evidente a ampla utilização de algoritmos de AM em tarefas de previsão de nível de rio. O trabalho conclui que *Support Vector Machine* (SVM) e *Artificial Neural Network* (ANN) se destacam em trabalhos que realizam gerenciamento de inundações.

Diante disso, foram estipuladas as hipóteses que:

1. Os modelos de AM são capazes de prever o nível dos rios da RH-Paraguai de forma satisfatória, com dados de estações a montante para prever o nível de estações a jusante?

2. Apresentam melhorias em relação aos métodos atuais?

Para a realização do presente trabalho foi utilizada uma fonte de dados com valores de níveis diários, fornecida pela Embrapa Pantanal, sendo selecionada uma amostra de três estações. Em seguida foi retirada a defasagem do tempo entre o pico de cheia de uma estação em relação à outra.

Os hiperparâmetros ideais para as redes LSTM, *Gated Recurrent Unit* (GRU) e *Bidirectional Long Short Term Memory* (BiLSTM) foram selecionados com o auxílio do *framework* Optuna e, em seguida, submetidos ao processo de treinamento. Posteriormente, foram selecionados os modelos de melhores resultados de cada algoritmo, os quais foram comparados com a técnica de Regressão utilizada atualmente.

Os resultados mostram que os três modelos testados podem ser utilizados para a tarefa de predição, em que, os três apresentam melhorias na predição dos picos mínimos, máximos e em relação as métricas de avaliação, quando comparados com o modelo atual, que utiliza regressões.

O trabalho apresenta como principais contribuições:

- A comprovação de que é possível a utilização de dados a montante para a previsão de estações a jusante na RH-Paraguai, e pode ser significativa alternativa frente à falta de dados, os quais, em situações de dados ausentes, podem ser utilizadas outras estações a montante. Esse tipo de previsão não foi encontrada na literatura para a RH-Paraguai;
- Possibilidade de auxiliar a Embrapa Pantanal na tarefa de previsões de níveis dos rios da RH-Paraguai. Também pode contribuir com instituições na definição de políticas públicas e ações que visem à mitigação dos impactos causados por esses eventos extremos;
- O *benchmarking* de modelos pode servir como parâmetro para pesquisas futuras na região da RH-Paraguai;
- Disponibilização dos códigos utilizados na busca por hiperparâmetros, treinamento das redes e download de dados do *WebService* da ANA, os quais podem contribuir para o desenvolvimento de demais trabalhos.

1.1 Objetivos

A proposta deste trabalho é aplicar algoritmos de AM para a predição de nível nos rios da RH-Paraguai, utilizando dados diários de estações a montante para previsão do nível de estações a jusante.

Os objetivos específicos são:

1. Obter uma amostra da base de dados selecionada e alinhada no tempo;
2. Ter modelos preditivos com algoritmos de AM;
3. Disponibilizar as predições no sistema de alertas da Embrapa Pantanal.

1.2 Estrutura do documento

O trabalho está organizado em 6 capítulos, em que, no Capítulo 2 está o Referencial Teórico, com os conceitos necessários para o desenvolvimento do trabalho. No Capítulo 3 está descrito os trabalhos encontrados na literatura, os quais realizam estudos similares ao tema proposto. A metodologia está descrita

no Capítulo 4. No Capítulo 5 estão os resultados e discussões sobre o trabalho realizado. As conclusões, dificuldades encontradas e sugestões para possíveis estudos futuros estão no Capítulo 6.

CAPÍTULO

2

Referencial Teórico

Este capítulo é destinado a descrever os conceitos relacionados que são necessários para o desenvolvimento do trabalho.

Na Seção 2.1, estão os conceitos sobre as séries temporais, visto que os dados utilizados no presente trabalho são dessa natureza. Já na Seção 2.2, estão os conceitos sobre AM, os algoritmos e os conceitos necessários para a etapa de modelagem. Por fim, a Seção 2.3 contém as métricas de desempenho utilizadas na etapa de avaliação dos modelos.

2.1 Séries temporais

Morettin and Toloi (2006) consideram que uma série temporal é um conjunto de observações ordenadas ao longo do tempo. Representa com a notação $Z_1, Z_2 ... Z_n$ de n observações ou com notação vetorial $Z = \{Z_1, ..., Z_n\}$.

Séries temporais estão presentes nas mais variadas áreas e aplicações, como número mensal de casos de dengue, índice de precipitação diário, mensal ou anual em uma determinada cidade e temperaturas diárias.

O intervalo das observações da série temporal, ou seja, o tempo que se leva para a ocorrência dos dados, pode classificá-la como contínua ou discreta.

Quando as observações são coletadas em todos os instantes do tempo, a sé-

rie temporal é referida como contínua. Por outro lado, uma série temporal é chamada de discreta quando suas observações são amostradas em intervalos de tempo (Rios, 2013).

Nas Figuras 2.1 e 2.2 podem ser observadas, respectivamente, séries temporais contínuas e discretas.

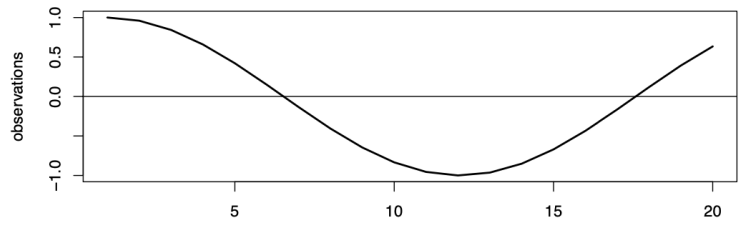

Figura 2.1: Série temporal contínua (Rios, 2013).

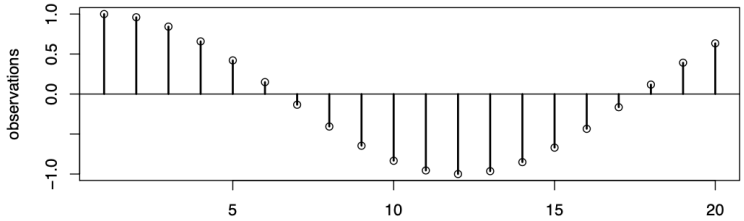

Figura 2.2: Série temporal discreta (Rios, 2013).

Uma série temporal pode apresentar características como:

- Sazonalidade: Diz respeito a comportamentos que podem se repetir em períodos específicos de tempo, como exemplo a venda de chocolates que aumentam no período da Páscoa;
- Tendência: Quando a série possui um determinado comportamento ao longo do seu tempo, podendo ser crescente, decrescente ou contínua;
- Aleatoriedade: Quando não há um padrão de sazonalidade ou tendência, a série varia sem um padrão específico.

2.2 Aprendizado de máquina

Aprendizado de Máquina é um campo da inteligência artificial que aplica e treina algoritmos que possam aprender com os dados e o contexto em que es-

tão inseridos, sem a necessidade de que seu comportamento seja previamente programado. Goodfellow et al. (2016) definem que:

> "Um algoritmo de aprendizado de máquina é um algoritmo capaz de aprender com os dados."

Para Awad and Khanna (2015), a aprendizagem é fundamental na generalização do problema, agindo de acordo com sua experiência histórica. O conjunto de dados de treinamento possui um domínio de problema existente, que pode ser usado para criar um modelo geral sobre esse domínio, o que permite que o modelo gere resultados precisos em novos casos.

O treinamento permitirá que a rede aprenda comportamentos a partir dos dados, os quais podem ser extrapolados para outros conjuntos, como um conjunto de dados de teste, por exemplo.

Géron (2019) apresenta um exemplo da aplicação de AM em um filtro de spam, no qual se aprende, automaticamente, as palavras e frases que são boas preditoras de spam. Caso o remetente perceba que os e-mails contendo "4U" estão sendo bloqueados, podem alterar a grafia para "*For* U", com o objetivo de burlar o filtro de spam. Um filtro, baseado nas técnicas de AM, nota que o termo "For U", tornou-se um termo que possui frequente sinalização como spam, por usuários, e passa a sinalizá-lo sem a sua interação.

Quando considerado o aspecto de supervisão humana, os algoritmos de AM são divididos em quatro categorias:

1. Aprendizado supervisionado: Consiste no aprendizado que contém um objeto-alvo, também conhecido como objeto rotulado, ou seja, o valor que se deseja na saída é conhecido. Esse tipo de aprendizado pode ser utilizado para classificação, na qual se separam os dados de acordo com o seu rótulo, ou regressão, em que se prevê a saída de um objeto-alvo de acordo com os dados passados. A tarefa de regressão é muito utilizada em séries temporais.

2. Aprendizado não supervisionado: Esse aprendizado busca estruturas desconhecidas em dados não rotulados, ou seja, não se tem um objeto-alvo, sendo assim, não se conhece a saída desejada. Esse tipo de aprendizado é muito utilizado para agrupamento de instâncias desconhecidas.

3. Aprendizado semissupervisionado: Aprendizado no qual se utiliza uma pequena quantidade de dados rotulados e um conjunto de dados não rotu-

lados, ou seja, são definidos poucos rótulos a serem utilizados na base de dados não rotulada.

4. Aprendizado por reforço: Esse aprendizado busca definir a melhor ação a ser tomada diante da incerteza. Trabalha-se com a ideia de ganhar recompensas ou punições a cada escolha, de forma que possa aprimorar as escolhas no futuro.

Nas próximas Seções estarão os principais conceitos utilizados para o treinamento de algoritmos de AM utilizados na tarefa de predição, que é o foco do presente trabalho.

2.2.1 Deep Learning

Deep Learning, ou seja, aprendizado profundo, é uma forma de AM muito utilizada na resolução de problemas grandes ou complexos, que exigem extrema capacidade de aprendizado em uma escala maior, devido à dimensão dos dados ou complexidade do problema.

De acordo com Awad and Khanna (2015), o aprendizado profundo está em ascensão porque as arquiteturas tradicionais de aprendizado mostraram-se impróprias para as tarefas mais desafiadoras de AM e de Inteligência Artificial (IA).

Goodfellow et al. (2016) definem que *Deep Learning* permite que os computadores aprendam seguindo uma hierarquia de conceitos, partindo de conceitos mais simples. Nessa hierarquia, os conceitos são construídos um sobre o outro, os quais, se colocados em um gráfico, será possível observar que este ficará com muitas camadas ou profundo, diante disso, considera-se a denominação de aprendizado profundo.

Uma *Deep Learning* busca construir conceitos complexos partindo de conceitos mais simples, utilizando RNA, no qual o aprendizado é dividido em hierarquias ou camadas, o que permite que o aprendizado ocorra em várias etapas. Quanto maior a quantidade de etapas ou profundidade dessas, maior a quantidade de instruções que podem ser executadas em sequência.

Segundo Goodfellow et al. (2016), a profundidade do modelo pode ser medida de duas maneiras principais:

- Considerando o número de instruções sequenciais que podem ser executadas, ou seja, medindo o comprimento do caminho mais longo através de

um fluxograma que descreve o cálculo em que se considera cada saída do modelo de acordo com suas entradas.

- Utilizando modelos probabilísticos profundos, dos quais não se considera a profundidade do gráfico computacional, mas a profundidade do gráfico que descreve como os conceitos se relacionam entre si.

Uma RNA opera de maneira semelhante ao cérebro humano, através da interconexão dos neurônios, ou seja, simulam os neurônios biológicos.

Tan et al. (2005) consideram que as RNAs foram inspiradas na tentativa de simular sistemas neurais biológicos, onde os neurônios são ligados a outros neurônios por meio de axônios que transmitem impulsos nervosos de um neurônio para outro sempre que esses neurônios são estimulados. Um neurônio está conectado aos axônios de outros neurônios por meio de dendritos, que são extensões do corpo celular do neurônio. O ponto de contato entre um dendrito e um axônio é chamado de sinapse.

Um *Perceptron* ou neurônio artificial é o que simula o funcionamento do neurônio humano; quando esses neurônios são utilizados em conjunto criam diferentes configurações de RNAs e aprendizados.

Conforme descrito por Buduma and Locascio (2017) e observado, na Figura 2.3, que o neurônio artificial recebe um número de entradas, $x_1, x_2, ..., x_n$, onde cada uma é multiplicada por um peso específico, $w_1, w_2, ..., w_n$. Em seguida, a soma dessas entradas ponderadas é passada para a função de ativação do neurônio, em que a saída desta função determina quais neurônios serão ativados, sendo que a saída pode ser transmitida para outros neurônios.

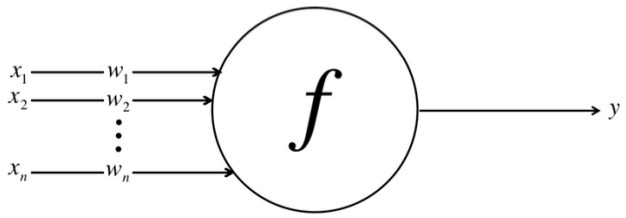

Figura 2.3: Esquema de um neurônio artificial (Buduma and Locascio, 2017).

Na Figura 2.4 pode ser observado um exemplo de uma RNA, as redes podem ser configuradas com uma ou várias entradas, podendo ter uma ou várias camadas ocultas *Hidden Layers* que aumentam a complexidade da rede quando

necessário. As saídas também podem ser configuradas com uma ou mais saídas, dependendo do problema a ser resolvido.

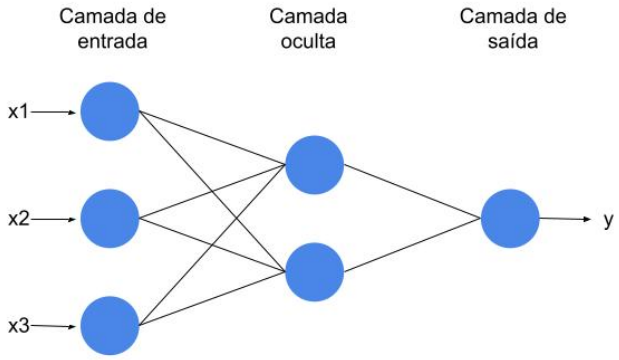

Figura 2.4: Exemplo de uma RNA multicamadas.

Função de Ativação

Em AM uma função de ativação é usada para introduzir não linearidades na rede neural (Pomerat et al., 2019).

As funções de ativação desempenham o papel de transformar a saída de um neurônio de acordo com sua equação, determinando os neurônios a serem ativados.

Cada neurônio artificial possui uma função de ativação para decidir a taxa de disparo dos neurônios a serem ativados (Lau and Hann Lim, 2018).

A escolha das funções de ativação depende do problema a ser resolvido, pois cada uma apresenta uma resolução específica.

Géron (2019) apresenta alguns exemplos de funções de ativação, entre eles:

- Função tangente hiperbólica (tanh) que, assim como a função logística, é em forma de S, contínua e diferenciável, mas seu valor de saída varia de –1 a 1 (em vez de 0 a 1 no caso da função logística). Esse intervalo tende a tornar a saída de cada camada mais ou menos centrada em torno de 0 no início do treinamento, o que geralmente ajuda a acelerar a convergência;

- Função ReLU é contínua, não diferenciável em $z = 0$ e sua derivada é 0 para $z < 0$, ou seja, se o valor for maior ou igual a zero, retorna o valor, senão retorna zero. Na prática, tem a vantagem de ser rápido de calcular, assim, se tornou o padrão, e o fato de não possuir um valor máximo de saída, ajuda a reduzir alguns problemas durante a descida do gradiente.

Existem outras funções de ativação, como por exemplo: linear, sigmoid, softmax, em que cada uma deve ser selecionada para o problema específico que se queira resolver.

Otimizadores

O processo de treinamento em AM se torna cada vez mais complexo e demorado, diante do aumento exponencial de dados e limitações com recursos computacionais. Diante disso, os algoritmos de otimização atuam para acelerar e facilitar o processo de convergência.

Goodfellow et al. (2016) descrevem alguns algoritmos de otimização, entre eles: RMSProp, Adagrad e Adam, que alteram a taxa de aprendizado no decorrer do treinamento.

O AdaGrad reduz a taxa de aprendizado de acordo com todo o histórico do gradiente, e pode tornar a taxa de aprendizado muito pequena.

RMSProp altera a acumulação de gradiente em uma média móvel exponencialmente ponderada. Usa uma média exponencialmente decrescente para descartar o histórico do passado extremo, para que possa convergir rapidamente.

Adam pode ser visto como uma variante da combinação de RMSProp e Momentum. Esse algoritmo inclui correções de viés nas estimativas dos momentos de primeira ordem e dos momentos de segunda ordem (não centrados) para explicar sua inicialização na origem. É, geralmente, considerado bastante robusto para a escolha de hiperparâmetros.

Dropout

Dropout é conhecido como uma técnica de regularização que previne a ocorrência de *overfitting*.

Segundo Géron (2019) o *Dropout* trabalha com o principio de que os neurônios da camada oculta não podem co-adaptar com seus neurônios vizinhos, eles devem ser tão úteis quanto possível e por conta própria. Não podem confiar em apenas alguns neurônios de entrada, devem prestar atenção a cada um dos seus neurônios de entrada, dessa forma, tem-se uma rede mais robusta.

O *dropout* executa a probabilidade de eliminar um ou mais nós da camada oculta na rede neural, dessa forma, são eliminados os nós seguintes conectados a eles. Assim, a rede tem seu tamanho reduzido e, então, o treino é efetuado, em

seguida são eliminados outros conjuntos de nós e, assim, sucessivamente, para cada exemplo de treino.

Para Hinton et al. (2012) no momento do teste é usada a "rede média", que contém todas as unidades ocultas, mas, com seus pesos de saída reduzidos pela metade, para compensar o fato de que o dobro deles está ativo. Na prática, isso oferece um desempenho muito semelhante à média de um grande número de desistências nas redes.

Cross-Validation

A Cross-validation (validação cruzada) é uma forma de verificar se o modelo pode ser generalizável, pois trabalha com o treinamento e testes de amostras de comprimentos variáveis, nas quais se utiliza o erro médio para avaliação do modelo.

De acordo com Goodfellow et al. (2016), o método se baseia na ideia de repetir a computação de treinamento e teste em diferentes subconjuntos ou divisões do conjunto de dados original escolhidos aleatoriamente. O mais comum deles é o procedimento de validação cruzada k-fold, no qual uma partição do conjunto de dados é formada dividindo-o em k subconjuntos não sobrepostos. O erro de teste pode, então, ser estimado tomando o erro de teste médio em k tentativas.

Para séries temporais, esses subconjuntos ou divisões do conjunto de dados não podem ser escolhidos aleatoriamente, para manter as propriedades da série temporal. Diante disso, é utilizada a função TimeSeriesSplit, da biblioteca *Sklearn*, que é específica para realizar essas divisões do conjunto de dados, mantendo as propriedades da série temporal.

2.2.2 Long Short Term Memory

Recurrent Neural Network (RNN)s são Redes Neurais para dados sequenciais, as quais são aplicadas em séries temporais. As RNNs utilizam não apenas os dados de entrada, mas também as saídas anteriores para fazer a previsão atual (Petneházi, 2019).

Segundo Aggarwal (2018), LSTM é um aprimoramento da arquitetura de RNN, que busca solucionar o problema de explosão ou fuga do gradiente. Em LSTM são alteradas as condições de recorrência de como os estados ocultos são propagados. Para atingir este objetivo, é adicionada uma memória de longo prazo, a

qual retém, pelo menos, uma parte da informação de estados anteriores.

Dessa forma, LSTM é um tipo de RNN capaz de armazenar valores ao longo do tempo através do seu barramento de memória. É o que faz essa técnica ser muito utilizada em problemas que envolvam séries temporais.

De acordo com Buduma and Locascio (2017), a memória da LSTM armazena informações críticas que aprendeu ao longo do tempo. Essas informações podem ser mantidas ao longo de várias etapas e, a cada passo, a unidade LSTM modifica a memória com novas informações. Essa característica define que LSTM pode aprender relacionamentos de longo prazo com mais eficácia do que uma RNN comum.

Na prática, uma rede LSTM procura aprender o que é importante e deve ser armazenado a longo prazo, o que pode ser descartado e o que deve ser recuperado para leitura.

Géron (2019) resume seu funcionamento como descrito e explicado abaixo, sendo que trabalha com três tipos de componentes, denominados portões ou portas, sendo:

- De esquecimento, que controla quais partes de estado de longo prazo podem ser descartadas;
- De entrada, que controla quais partes devem ser adicionadas ao estado de longo prazo;
- De saída, que controla as partes do estado de longo prazo que devem ser lidas.

O funcionamento do estado de longo prazo se inicia passando por um porta de esquecimento a qual descarta algumas memórias que não são mais úteis. Em seguida, são adicionadas as memórias que foram selecionadas pela porta de entrada. Dessa forma, a cada passo algumas memórias são descartadas e outras são adicionadas.

Além disso, após a operação de adição de novas entradas, o estado de longo prazo é copiado e o resultado é filtrado pela porta de saída, produzindo o estado de curto prazo, que é igual à saída da memória para esta etapa do tempo, como pode ser observado na Figura 2.5.

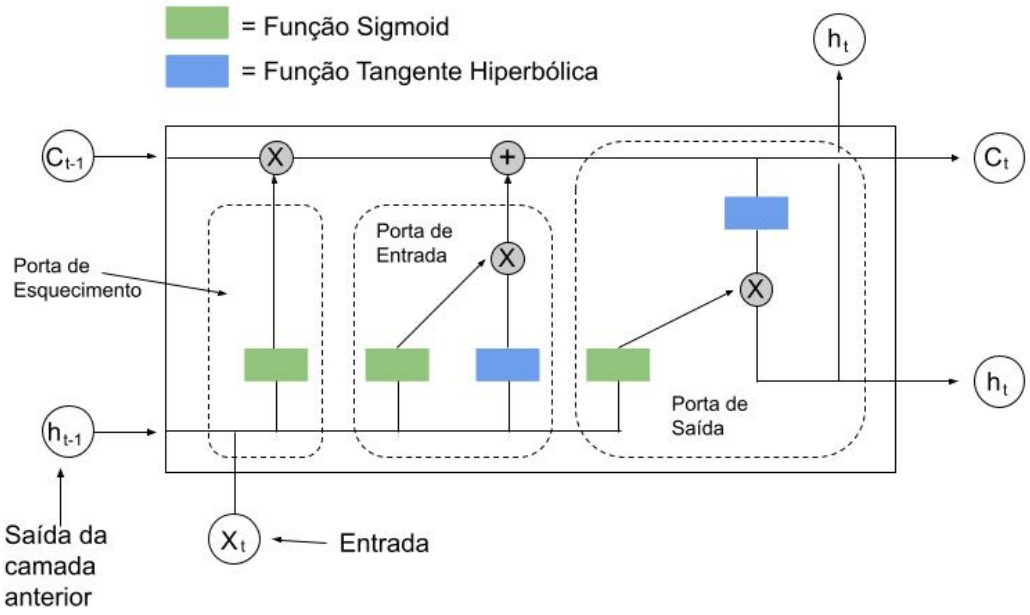

Figura 2.5: Funcionamento de uma LSTM.

2.2.3 *Bidirectional Long Short Term Memory*

Uma RNN, como a LSTM, trabalha considerando o evento de causa e efeito, ou seja, considera dados passados pra estimar o futuro, fato que qualifica esse tipo de rede como ótima alternativa para séries temporais.

As RNNs bidirecionais combinam uma RNN que avança no tempo, começando no início da sequência, com outra RNN a qual retrocede no tempo, começando no final da sequência (Goodfellow et al., 2016).

Segundo Aggarwal (2018), uma observação imediata sobre os estados ocultos, na direção para frente e para trás, é que eles não interagem entre si. Os estados de saída são calculados a partir dos estados ocultos nas duas direções.

As redes LSTMs podem ser combinadas com a implementação Bidirecional, sendo considerada BiLSTM, a qual segue o mesmo conceito, mas, neste caso, combinando uma rede LSTM, que começa no início, com outra que retrocede no tempo.

Siami-Namini et al. (2019), concluíram em seu estudo que os modelos BiLSTM forneceram melhores previsões em séries temporais em comparação aos modelos *Autoregressive Integrated Moving Average* (ARIMA) e LSTM, pois esses permitem treinamento adicional percorrendo os dados de entrada duas vezes. Da esquerda

para a direita e da direita para a esquerda.

2.2.4 Gated Recurrent Unit

De acordo com Aggarwal (2018), uma GRU pode ser vista como uma simplificação da LSTM, que não usa estados de célula explícitos e não controla, diretamente, a quantidade de informações alteradas no estado oculto, usando portas separadas de esquecimento e saída, como ocorre na LSTM.

Uma rede GRU altera a arquitetura mas mantém, praticamente, a mesma ideia de manter a memória, como ocorre na rede LSTM, alterando e simplificando a forma de trabalho.

Uma célula GRU típica contém a porta de redefinição e a porta de atualização, portanto, compreendendo menos parâmetros, o que implica cálculos mais diretos (Wang et al., 2020).

Géron (2019) descreve as principais alterações em relação à LSTM:

- Ambos os vetores de estado ("c"e "h") são fundidos em um único vetor h(t);
- Um controlador de porta controla tanto a porta de esquecimento quanto a porta de entrada. Se o controlador de porta emitir um 1, o esquecimento está aberto (= 1) e a porta de entrada está fechado (1 – 1 = 0). Se emitir um 0, acontece o oposto;
- Não há porta de saída, então o vetor de estado completo é gerado a cada passo de tempo. Entretanto, existe um controlador de porta de redefinição que controla qual parte do estado anterior será mostrada para a camada principal.

A diminuição de portas e a junção dos barramentos (vetores de estado), tornam a GRU uma rede de treinamento mais rápida devido os cálculos serem mais diretos e, consequentemente, não possuírem tantos pesos como a LSTM.

2.3 Métricas de desempenho

Tarefa tão importante quanto criar e treinar modelos em AM é testar o aprendizado, ou seja, verificar o quanto ele está acertando. Neste sentido, é necessário

definir e aplicar as métricas de desempenho, as quais possibilitam que se compare o resultado alcançado com o resultado esperado.

Algumas métricas comumente encontradas em trabalhos de predição são *Root Mean Squared Error* (RMSE) e *Mean Absolute Error* (MAE), além delas, outras duas se mostram interessantes para a avaliação dos resultados, que são: *Mean Square Error* (MSE) e *Mean Absolute Percentage Error* (MAPE).

Géron Géron (2019) apresenta a equação de RMSE observada na Equação 2.1, na qual considera que RMSE trabalha com base na norma Euclidiana.

$$RMSE(\mathbf{X},h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(\mathbf{x^{(i)}}) - y^{(i)})^2} \tag{2.1}$$

Já o MSE Equação 2.2 utiliza o mesmo conceito de RMSE, porém não se extrai a raiz, mantendo a unidade de medida quadrática.

$$MSE(\mathbf{X},h) = \frac{1}{m}\sum_{i=1}^{m}(h(\mathbf{x^{(i)}}) - y^{(i)})^2 \tag{2.2}$$

Na Equação 2.3 pode ser observada a equação de MAE, que trabalha com base na norma de *Manhattan*.

$$MAE(\mathbf{X},h) = \frac{1}{m}\sum_{i=1}^{m}|h(\mathbf{x^{(i)}}) - y^{(i)}| \tag{2.3}$$

Onde:

- $\mathbf{X}$ representa a matriz com todos os valores, de todas as instâncias, excluindo os rótulos.
- h representa a função de previsão, ou hipótese.
- m representa o número de instâncias do conjunto de dados.
- $\mathbf{x^{(i)}}$ representa um vetor de todos os valores, excluindo o rótulo da i-ésima instância.
- $y^{(i)}$ representa o seu rótulo, ou seja, o valor de saída desejado dessa instância.

De acordo com Malhotra and Jain (2011), MAE mede a distância que as estimativas ou previsões estão dos valores reais, e considera que RMSE prioriza a detecção de erros grandes, ou seja, a valores discrepantes.

MAPE segue o mesmo princípio que MAE, porém apresenta o erro em forma de porcentagem, fato que o torna visualmente de fácil interpretação.

Entende-se que, no caso das métricas mencionadas, quanto menor o seu valor, melhor o desempenho do modelo.

2.4 Considerações Finais

Para o desenvolvimento do presente trabalho é fundamental que se tenha conhecimento teórico das séries com as quais se trabalha, conceitos sobre AM e *Deep Learning*.

Não é possível realizar a modelagem com algoritmos de AM sem que se conheça os componentes necessários, seus parâmetros e como trabalham. Para a avaliação do modelo, é importante entender como as métricas são calculadas, para que se torne possível avaliar, corretamente, os modelos, bem como verificar se a métrica é adequada para o problema em questão.

Por mais que existam ferramentas de procura e seleção de parâmetros e hiperparâmetros, é imprescindível que se tenha bem fundamentado os conceitos sobre o funcionamento do que se manipula, permitindo a correta configuração da busca e interpretação dos resultados.

Sendo assim, este Capítulo contribui com os conceitos de entendimento necessários para o desenvolvimento do trabalho, bem como para a utilização de algoritmos LSTMs e suas variações, as quais se mostram promissoras para o trabalho preditivo com séries temporais de dados hidrológicos.

O Capítulo 3, a seguir, contém a revisão da literatura com trabalhos similares ao presente estudo.

CAPÍTULO

3

Trabalhos Relacionados

Neste Capítulo é apresentada a revisão sistemática da literatura, com trabalhos que realizam predições de nível de rios de maneira semelhante ao presente estudo.

3.1 Revisão Sistemática da Literatura

A partir da revisão sistemática da literatura, foram encontrados trabalhos que discutem o uso de técnicas de AM em contextos similares ao presente trabalho.

De acordo com Biolchini et al. (2005), a revisão sistemática é uma metodologia de pesquisa com etapas bem definidas, na qual o objetivo é reunir e avaliar evidências sobre um tema específico.

Para tal, foram executadas as etapas de planejamento, execução e avaliação, recomendadas por Felizardo et al. (2017) e descritas a seguir.

3.1.1 Planejamento

Essa etapa iniciou-se com a identificação da necessidade de uma revisão, sendo que essa se faz necessário para investigar quais práticas estão sendo adotadas para a previsão de níveis de rios.

Em seguida, foram formulados os questionamentos a serem esclarecidos ou respondidos após a execução da revisão, ou seja, o foco de interesse da revisão, elencados abaixo:

1. Quais algoritmos e ferramentas têm sido investigados e aplicados para a predição de níveis de rios?
2. Qual a quantidade de fontes de dados utilizadas?
3. Qual a periodicidade dos dados?
4. Quais métricas têm sido utilizadas para a avaliação dos resultados?

As questões acima possuem o grupo que se pretende observar (população), o que será observado no contexto da revisão (intervenção) e os resultados esperados ao final da revisão.

Neste caso, a população são os trabalhos que realizam a predição de níveis de rios, já a intervenção definiu-se como sendo os algoritmos, métricas de avaliação e fontes de dados. Como resultados, espera-se identificar o que está sendo empregado no âmbito de predição de níveis de rios.

O próximo passo do planejamento consiste em definir e selecionar as fontes, as quais foram consideradas fontes com cobertura de conferências e *journals*, que tenham:

1. Interface amigável;
2. Atualizações regulares de publicações na base;
3. Disponibilizem o texto completo;
4. Possibilite a exportação do conteúdo de maneira prática;
5. Tenham precisão na busca de acordo com os critérios de busca, permitindo a utilização de *strings* de buscas personalizáveis.

De acordo com os critérios de seleção das fontes foram investigadas as bases do Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)[1], as quais possibilitam a busca de bases por área e subáreas do conhecimento.

[1] https://www-periodicos-capes-gov-br.ezl.periodicos.capes.gov.br/index.php?

Para a presente revisão foi aplicado o filtro de Ciências Exatas e da Terra para a área do conhecimento e Ciência da Computação para a subárea, pelo qual retornaram o total de 100 resultados. Dos 100 resultados retornados, alguns já foram descartados de imediato por serem referentes a livros, estatísticas, patentes, materiais audiovisuais e repositórios institucionais.

O objetivo é a busca por bases ou buscadores que indexem a maior quantidade de trabalhos possíveis e tenha alcance de nível global, por esse motivo, e de acordo com os critérios de seleção de fontes citados acima, foi selecionado o Scopus[2] .

Scopus permite a utilização de *string* de busca única em bases conhecidas pela comunidade científica e que também estão presentes na lista de bases do Periódicos CAPES, como:

- ACM;
- Elsevier;
- IEEE;
- Springer.

A lista completa das bases, a qual pode ser utilizada a mesma *string* através do Scopus está disponível em: https://www.elsevier.com/?a=91122.

Definida a base, é essencial que se crie a *string* de busca, formulada para responder às questões elencadas no planejamento. A *string* foi definada como:

("water level prediction"OR "flood prediction"OR "flood forecasting"OR "river prediction"OR "river forecasting") AND ("machine learning")).

Junto à definição da *string* acima, foram planejadas aplicações de filtros de:

- Idioma: Inglês, por ser aceito internacionalmente;
- Período: Entre os anos de 2011 a 2021, buscando as técnicas mais atuais para a resolução do problema.

Ainda na etapa de planejamento foram definidos os critérios de inclusão e exclusão dos trabalhos, visando encontrar os trabalhos mais relacionados, sendo:

[2] https://www.scopus.com/home.uri

- Inclusão: utilizar AM, apresentar as fontes de dados utilizadas, apresentar as métricas utilizadas para verificação dos resultados, apresentar os rios ou bacias, objeto do estudo, deixar claro os algoritmos utilizados;
- Exclusão: não utilizar AM, não indicar os algoritmos utilizados, não apresentar os resultados da predição de forma clara, não realizar nenhum comparativo com outras técnicas ou algoritmos, não trabalhar com os dados de série temporal e não descrever com detalhes os dados utilizados.

Por fim, foram definidos os critérios de qualidade, listados abaixo:

- O objetivo da pesquisa é similar ao contexto do presente trabalho;
- Efetua a avaliação dos resultados de forma clara e imparcial;
- Justifica a escolha das fontes de dados;
- Apresenta, de maneira clara, os algoritmos utilizados e o motivo para essa escolha.

3.1.2 Execução e avaliação

Após a etapa de planejamento, a *string* de busca foi aplicada na base, em que se obteve como resultado 200 trabalhos.

Na Figura 3.1, fica evidente que o tema está sendo discutido com mais ênfase nos últimos anos, em que 60% das publicações foram realizadas nos dois últimos anos, 2020 e 2021.

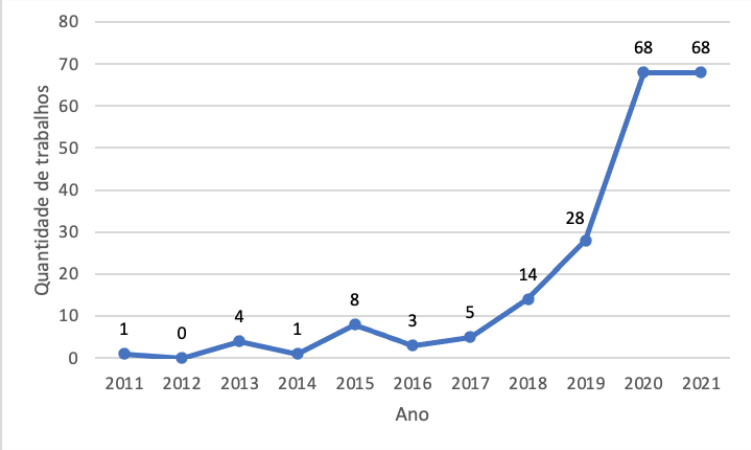

Figura 3.1: Quantidade de trabalhos por ano de publicação.

O próximo passo na execução da revisão foi o de pré-seleção, que consiste na leitura dos títulos e resumos dos trabalhos. Após a leitura, foram aplicados os critérios de inclusão e exclusão, com o objetivo de definir os trabalhos que deveriam ser lidos na íntegra, para análise mais refinada de inclusão e exclusão.

Após a leitura na íntegra e aplicação dos critérios de inclusão e exclusão, foi obtido, como resultado final, o total de 17 trabalhos, considerados como os mais relacionados, como pode ser observado na Tabela 3.1, na qual é apresentada a quantidade de trabalhos selecionados após a aplicação de cada critério.

Critério	**Total de Trabalhos**
Total de trabalhos	200
Leitura do título e resumo - inclusão	105
Leitura do título e resumo - exclusão	75
Leitura na íntegra - inclusão	**26**
Leitura na íntegra - exclusão	**17**

Tabela 3.1: Quantidade de trabalhos após a aplicação dos critérios.

Com o objetivo de responder os questionamentos do planejamento da revisão, foi elaborada a Tabela 3.2, referente aos trabalhos selecionados, a qual contém as colunas:

- ID, identificador para relacionar os trabalhos da Tabela 3.2 com suas referências na Tabela 3.3;
- Algoritmos, com os algoritmos utilizados para a tarefa de predição. Os que foram superiores nas comparações foram destacados em negrito;
- Dados, com as características da base de dados utilizada;
- Avaliação, com as métricas utilizadas.

ID	Algoritmos	Dados	Avaliação
1	**LSTM** e *Adaptive Neuro-Fuzzy Inference System* (ANFIS)	O tamanho médio máximo da bacia ultrapassa 12.000 km^2 no período chuvosa. Foram utilizados dados diários de vazão e nível de água para as estações denominadas Kratie e Chaktomuk. Para a estação do porto de Phnom Penh os dados não estavam disponíveis e foram modelados com modelos hidrológicos. Dados de 2000 a 2012 (4.749 amostras), 2013 a 2016 (1.461 amostras) e 2017 a junho de 2020 (1.276 amostras) foram divididos em treinamento, validação e teste, respectivamente. Os autores trabalharam com alguns algoritmos de classificação para preparação dos dados antes da modelagem preditiva, devido ao evento de fluxo reverso da bacia.	RMSE, *Coefficient of Determination* (R^2) e *Nash-Sutcliffe Efficiency Coefficient* (NSE)
2	*Back Propagation Neural Networks* (BPNN), SVM, LSTM e **Attention-LSTM**	A área de estudo é localizada na área montanhosa do Sul de Anhui, na China, a área de captação da bacia hidrográfica de Tunxi é de 2.670 km^2. Os dados que foram utilizados eram de 1991 a 1999, tratava-se de observações horárias que foram coletados de uma estação que monitora vazão, e dez estações que monitoram precipitação. Totalizando 19.535 registros de vazão e mais 19.535 registros de precipitação, os quais, no último ano, os dados foram destinados a testes do modelo, e o restante destinados ao treinamento.	RMSE, R^2 e NSE
3	SVM, LSTM e ***Sequence-to-Sequence Long Short Term Memory* (LSTM-S2S)**	Utilizaram dados de vazão observados em 5 estações a montante na bacia do rio Russo, na Califórnia, EUA. A área de drenagem da bacia é de 3.846 km^2. Os dados eram de observações horárias, no período de 2015 a 2019. Para treinamento, foram utilizados 80% dos dados e 20% para teste. Valores ausentes foram substituídos por valores médios, considerando o valor anterior e posterior.	RMSE, R^2, NSE, *Peak Runoff Error* (PRE) e *Peak Time Error* (PTE)
4	***Support Vector Machine - Firefly Algorithm* (SVM-FA)**, *Sadial Basis Function - Firefly Algorithm* (RBF-FA), *Radial Basis Function Neural Network* (RBFNN), *Feed Forward Back Propagation Neural Network* (FFBPNN) e **SVM**	Dados mensais de vazão das estações Silchar e Dholai, localizadas no distrito de Cachar de Assam, na Índia, foram utilizados no estudo. Primeiramente, foram coletados dados diários de vazão, para os meses de maio a outubro, de 1992 a 2019, em seguida, os dados diários foram convertidos em dados mensais e utilizou-se 70% do conjunto de dados para treinamento e 30% para teste.	RMSE, R^2 e MSE

ID	Algoritmos	Dados	Avaliação
5	*Support Vector Regression* (SVR), *MultiLayer Perceptron* (MLP) e **LSTM**	O estudo ocorreu no rio Parma, na Itália, o qual tem, aproximadamente, 92 km^2 de extensão, desde a cabeceira até a confluência com o rio Po. O estudo focou no trecho inferior de 32 km^2 de comprimento, Os dados das estações de Colorno, Ponte Verdi e Casalmaggiore(Rio Po) foram coletados de 2012 a 2020, em intervalos de 10 min. Os dados originais foram reduzidos para valores horários. No total, 47 eventos de inundação foram extraídos desta série temporal. Os eventos foram categorizados em 4 grupos: pequenas, médias e grandes inundações.	RMSE, MAE, *Coefficient Of Correlation* (CC), NSE e MAPE
6	MLP, **LSTM** e GRU	O trabalho considerou dados de 4 estações a montante e 1 estação a jusante da bacia hidrográfica Brandywine-Christina, na Pensilvânia, com uma área de drenagem de 1.929 km^2. Coletou-se dados históricos de precipitação e vazão de 1 de julho de 2013 a 30 de junho de 2019, com um total de 401.896 amostras que tinham o intervalo de tempo de 15 minutos para 4 estações. Utilizou-se os dados de chuva observados, a montante, como dados de entrada e os dados de vazão, a jusante, como dados de saída do modelo. O conjunto de dados utilizado foi dos anos de 2013 a 2018 e foi dividido em 75 % para treinamento, 15% para validação e 15% para teste do modelo. Por fim, dados do ano de 2019 foram utilizados para testar, prever e avaliar o desempenho do modelo como dados em tempo real.	NSE, RMSE e Regression R
7	ARIMA e **LSTM**	O estudo considerou o rio Meramec, em Valley Park, no estado americano de Missouri. Os dados foram dos anos de 2016 a 2019, são dados de nível com um intervalo de 15 minutos, totalizando 113.994 amostras. Foi realizada a divisão de 80% para treinamento e 20% para teste do modelo.	RMSE e MAE
8	ANN, *Random Forest* (RF) e **SVR**	O estudo coletou características meteorológicas, hidrológicas e naturais da bacia hidrográfica de Tich-Bui, no Vietnã. Foram coletados dados de 8 estações hidrológicas e meteorológicas para a previsão de nível em uma outra estação. Os dados de chuvas observadas coletadas correspondem a períodos de 3 meses e meio, a partir da 01h00 do dia 15/06. Os dados de níveis de água foram coletados no período de 01/05 a 31/11 dos anos de, 2014 a 2017, com intervalo de tempo de 6 horas.	NSE e RMSE

ID	Algoritmos	Dados	Avaliação
9	***Spatio-Temporal Attention Long Short Term Memory* (STA-LSTM)**, *Fully Connected Network* (FCN), SVM e LSTM	A área de estudo foi a bacia do rio Lech, na Europa. O rio Lech tem comprimento total de 263 km^2 e a área da bacia é de 4.126 km^2. A base de dados possuía 7.360 dados hidrológicos, que corresponderam a momentos de aumento significativo no volume de escoamento a jusante. Os dados foram coletados entre os anos de 2002 a 2018, em que 174 amostras do total de dados representaram múltiplas cheias na bacia, nos períodos de maio de 2002, agosto de 2013 e janeiro de 2018 e foram utilizados como conjunto de teste. Os autores utilizaram suavização exponencial dupla de Holt-Winters para a diminuição dos ruídos dos dados brutos.	RMSE e *Deterministic Coefficients* (DC)
10	**ANN**, *Decision Tree* (DT), RF e SVM	O estudo ocorreu na zona úmida de Upo, que é a maior zona úmida do interior da Coréia do Sul. O conjunto de dados continha observações do nível de água de 1° de abril de 2009 a 31 de março de 2015. Entre 2009 e 2010, os níveis de água foram medidos a cada 10 minutos e entre 2011 e 2015, os níveis de água foram medidos a cada 15 minutos. Os dados foram convertidos em dados de nível de água médio diário e usados como variáveis dependentes. Os dados de treinamento corresponderam ao período de 1° de abril de 2009 a 31 de março de 2013 e o período de teste foi de 1° de abril de 2013 a 31 de março de 2015. As variáveis independentes foram: temperatura média diária, temperatura mínima diária, temperatura máxima diária, precipitação diária, velocidade do vento instantânea máxima diária, velocidade média diária do vento de uma estação a, aproximadamente, 6 km^2 da zona úmida de Upo, além de dados de nível de água, medidos em duas estações a montante.	CC, NSE, RMSE e *Persistence Index* (PI)
11	SVR, ANN, ***Deep Neural Network* (DNN)**,	O estudo coletou 10 anos de dados entre 2007 e 2017, o conjunto de dados continha 3 atributos principais, sendo data/hora, nível médio de maré por hora e nível médio de maré diário. Os dados foram pré-processados para derivar alguns atributos preditores como: altura mínima e máxima da maré, desvio da altura da maré, média móvel de 1,2,5 e 7 dias e exponencial média móvel. Por fim, a base de dados foi dividida em 70% para treinamento e 30% para teste.	RMSE, MAE,
12	SVM e **LSTM**	O estudo foi realizado na bacia hidrográfica de Tunxi, que é uma pequena bacia hidrográfica de 2.696,76 km^2, foram coletados cerca de 19.000 amostras de dados hidrológicos entre 1991 a 1999 no período de cheias. O conjunto de treinamento continha 15.000 amostras e o restante foi destinado ao conjunto de teste. O objetivo foi a previsão de vazão.	RMSE, *Median Absolute Error* (MedAE) e R^2

ID	Algoritmos	Dados	Avaliação
13	ANN, SVM, *Wavelet Transform Coupled With An Artificial Neural Network* (WANN), *Wavelet Transform Coupled With Support Vector Machine* (WSVM) e ***Deep Convolutional Neural Network* (DeepCNN)**	O trabalho foi realizado considerando o rio Shannon, na Irlanda, em que o conjunto de dados era de 11.392 amostras correspondentes a amostras diárias em um período de 30 anos, de 1983 a 2013. Os dados foram coletados em três estações hidrométricas diferentes. As entrados dos modelos foram: temperatura máxima e mínima, escoamento superficial. As saídas foram nível de água e vazão.	RMSE, MAE e R^2
14	ANN, **GRU** e LSTM	O estudo ocorreu considerando a bacia de controle da estação de Yutan, província de Fujian, sudeste da China. A base de dados foi constituída por dados de vazão horária de uma estação, e dados horários de precipitação de três estações a montante. O período dos dados foi de 2000 a 2014, esses dados foram tratados e selecionados 8.022 registros horários de chuva e vazão. Para o treinamento foram destinadas 6.383 amostras e, para testes, 1.639 amostras.	NSE e MAE
15	LSTM, **STA-LSTM**, *Temporal Attention LSTM* (TA-LSTM), *Apatial Attention LSTM* (SA-LSTM), *Historical Average* (HA), FCN, *Convolutional Neural Network* (CNN), *Graph Convolutional Networks* (GCN),	Foram utilizados dados da bacia de Tunxi, bacia de Changhua e bacia de Heihe, todas localizadas na China. Os dados incluíam dados de precipitação horária de todas as estações hidrológicas e dados de vazão horária de estações de descarga nessas bacias. Tunxy: 12 estações pluviométricas e 1 estação de vazão, sendo o total de dados de 49.532, em que, 39.625 foi destinado ao treinamento e 9.907 para o teste. Período dos dados: 27/06/1981 a 18/03/ 2007. Changhua: 8 estações pluviométricas e 1 estação de vazão, sendo o total de dados 9.354, em que 7.483 foi destinado ao treinamento e 1.871 para o teste. Período dos dados: 07/04/1998 a 20/07/2010. Heihe: 10 estações pluviométricas e 1 estação de vazão, sendo o total de dados de 5.423, em que 4.338 para o treinamento e 1.085 para o teste. Período dos dados: 01/04/2003 a 10/11/ 2010.	RMSE, MAE, MAPE, R^2 e *Ratio Of Qualified Predictions* (QR)

ID	Algoritmos	Dados	Avaliação
16	ARIMA, LSTM, ARIMA_LSTM, RF, **ARIMA_RF**, *K-Nearest Neighbors* (KNN), **ARIMA_KNN**, SVM e ARIMA_SVM	O estudo utilizou dados de nível de água de três estações hidrológicas do rio Vermelho, no Vietnã, e efetuou um preenchimento de dados ausentes quando necessário. Os dados foram: Estações de Vu Quang e Hanoi: Dados de 01/01/2008 a 31/12/2017 com diferentes frequências e com 29.224 amostras para cada uma. Hung Yen: Dados horários de 01/01/2008 a 23/04/2015 com 64.061 amostras.	MAE, RMSE, NSE, *Fraction of Standard Deviation* (FSD), *Similarity* (Sim) e R-Score
17	ANN, ANFIS e **LSTM**	No estudo, dados cumulativos de chuva foram coletados de oito estações pluviométricas a montante, para previsão do nível em uma estação a jusante. Os dados de precipitação acumulada foram coletados do período de 04/01/2011 a 30/11/2019 para todas as estações. Os dados foram dimensionados e depois divididos em 75% para treinamento e 25% para teste.	RMSE, NSE, MAE

Tabela 3.2: Informações sobre os trabalhos selecionados.

ID	Referências
1	(Morovati et al., 2021)
2	(Yan et al., 2021)
3	(Han et al., 2021)
4	(Sahoo et al., 2021)
5	(Dazzi et al., 2021)
6	(Mousavi et al., 2021)
7	(Gude et al., 2020)
8	(Nguyen and Le, 2019)
9	(Ding et al., 2019)
10	(Choi et al., 2020)
11	(Rasel et al., 2018)
12	(Yan et al., 2019)
13	(Assem et al., 2017)
14	(Gao et al., 2020)
15	(Ding et al., 2020)
16	(Phan and Nguyen, 2020)
17	(Khairudin et al., 2021)

Tabela 3.3: Referências dos trabalhos selecionados.

Em resposta às questões a serem respondidas pela presente revisão, tem-se:

- Resposta 1: O algoritmo mais citado nos trabalhos que foram selecionados é a LSTM, que esteve presente em 12 trabalhos, nos quais obteve o melhor desempenho na maioria dos testes realizados. Destaca-se, também, a quantidade de variações de implementações da LSTM que também apresentam bons resultados. Outro algoritmo muito frequente nos testes foi a SVM/SVR, o qual alguns autores apresentam como SVM, mas para tarefas de regressão é utilizado o algoritmo SVR, o qual esteve presente em 11 trabalhos, e apresentou resultados interessantes;

- Resposta 2: A maioria dos trabalhos apresentam, no mínimo, mais de uma fonte de dados, unindo dados de chuva, vazão, nível e até mesmo mapeando outros fatores ambientais, mas, em sua maioria, utilizam abordagens de pré-processamento desses dados, removendo possíveis ruídos ou deixando-os na mesma periodicidade;

- Resposta 3: A periodicidade dos dados varia de trabalho para trabalho, mas, no geral, são dados horários ou diários, o que deixa evidente que a maioria dos trabalhos possuem quantidade considerável de dados. Em poucos trabalhos foram considerados dados mensais ou apenas poucos dados;

- Resposta 4: As métricas mais utilizadas nos trabalhos foram RMSE, presente em 11 trabalhos, seguido por NSE, presente em 10 trabalhos e MAE, presente em 8 trabalhos. Entre as métricas citadas tem-se RMSE que penaliza erros maiores, enquanto MAE tende a medir o erro médio absoluto entre o valor real e o predito, as quais podem ser boas alternativas para a avaliação de modelos preditivos.

Essas informações são importantes para a etapa da modelagem, pois saber o que está sendo aplicado em contextos semelhantes e com bons resultados pode ajudar no planejamento da modelagem.

Morovati et al. (2021) utilizaram modelos hidrológicos para modelar dados não disponíveis de uma estação, já Han et al. (2021) e Mousavi et al. (2021) preencheram os valores ausentes com valores médios, considerando o valor anterior e o valor posterior.

Phan and Nguyen (2020) utilizaram interpolação linear e média móvel para preenchimento de poucos dados ausentes, e para preenchimento de maior quantidade de dados ausentes, utilizaram o método *Dynamic time warping-based im-*

putation (DTWBI), discutido por Phan et al. (2020), que levam em consideração o dinamismo dos dados.

Foram utilizados nos trabalho de Gude et al. (2020) e de Ding et al. (2020), ferramentas distintas para a procura e escolha de parâmetros e hiperparâmetros da rede, fato este que pode contribuir com a tarefa de modelagem; já o trabalho de Gao et al. (2020) descreve a utilização método de tentativa e erro para a escolha dos hiperparâmetros, testando várias combinações, o que pode ser um trabalho demorado e pode não gerar um resultado ótimo.

Gao et al. (2020) e Mousavi et al. (2021) utilizaram uma RNN denominada GRU, descrita como uma variação de redes LSTMs, porém, com estruturas um pouco menos complexas, o que chamou a atenção pelos resultados quando comparados à LSTM convencional, e por ter o processo de treinamento mais rápido, devido a sua estrutura.

3.2 Considerações finais

Neste Capítulo foram apresentados trabalhos relacionados a predições de nível de rios, os quais mostram casos de sucesso em diferentes cenários e motivam a realização deste trabalho na RH-Paraguai.

Os trabalhos relacionados colaboraram por garantir a necessária percepção de algoritmos mais utilizados e com melhores resultados, métricas que podem ser utilizadas no estudo, configuração de parâmetros e divisão de dados para treinamento e teste do modelo.

Os estudos deixam clarividente à importância de buscar o mapeamento dos aspectos hidrológicos e físicos no modelo, e que, alguns casos, são utilizadas abordagens híbridas com modelos estatísticos ou hidrológicos.

Um fato que não foi muito explorado nos trabalhos aqui analisados foi a questão de dados ausentes: poucos trabalhos descreveram como trataram os dados ausentes, ademais, sem muitos detalhes. Fato que pode ser um desafio para o presente trabalho, devido à quantidade de dados ausentes na série de dados.

Outro fator que pode ser um desafio para a aplicação de previsão a jusante, utilizando dados a montante, é a distância entre as estações e as influências que recebem em cada localidade.

Dos trabalhos mencionados na presente revisão, a maioria trabalha com ba-

cias ou porções relativamente pequenas em relação à bacia do presente estudo, ou com estações próximas uma das outras; quando são estações mais distantes, possuem maior variedade e quantidade de dados na entrada do modelo, aumentando a capacidade de simulação do fenômeno natural de subida e descida dos níveis de água, fato que impossibilita a replicação no presente estudo.

O presente estudo visa enfrentar esses fatores desafiadores e, na RH-Paraguai, serão utilizados dados diários de nível. Algumas estações estão em ambientes totalmente diferentes, umas no planalto outras no interior da bacia, portanto, recebem diferentes influências de chuva, escoamento, influência de águas de outros rios, relevo, entre outros fatores.

No Capítulo 4, a seguir, estão descritas as etapas para o desenvolvimento do presente estudo.

CAPÍTULO

4

Metodologia

Este Capítulo contém os passos seguidos para o desenvolvimento do trabalho e para o alcance dos objetivos propostos. Na Seção 4.1 está descrito o primeiro passo, que foi o de análise e entendimento da base de dados, seguida pela Seção 4.1.1, na qual é discutida a seleção das amostras na base de dados. O alinhamento das séries é apresentado na Seção 4.1.2, e as etapas de modelagem e *benchmarking* são descritas nas Seções 4.2 e 4.3, respectivamente.

4.1 Dados

Os dados utilizados no presente estudo são os que contém os níveis diários dos rios, coletados pelo CPRM nas vinte e duas estações de observação espalhadas nos rios da RH-Paraguai em sua porção brasileira. Os dados são os mesmos utilizados pela Embrapa Pantanal na elaboração de análises e previsões.

A base de dados foi disponibilizada em formato de planilhas eletrônicas, as quais cada planilha corresponde a uma estação de medição. Cada planilha possui a estrutura observada na Figura 4.1.

- Primeira linha não possui informações;
- Segunda linha possui apenas o título da planilha: “Dados obs”;

		Dados obs		
DIA	...	2017	2018	...
01/01	...	nível em cm	nível em cm	...
02/01	...	nível em cm	nível em cm	...
⋮	⋮	⋮	⋮	⋮
31/12	...	nível em cm	nível em cm	...
ANO	...	**2017**	**2018**	...
Nº de dados	...	**Total de dados no ano**	**Total de dados no ano**	...
DIA MÍN	...	dia pico mínimo (dia/mês)	dia pico mínimo (dia/mês)	...
MÍN	...	pico mínimo em cm	pico mínimo em cm	...
MÉD	...	média anual em cm	média anual em cm	...
MÁX	...	pico máximo em cm	pico máximo em cm	...
DIA MÁX	...	dia pico máximo (dia/mês)	dia pico máximo (dia/mês)	...

Figura 4.1: Estrutura da base de dados disponibilizada.

- Primeira coluna, intitulada de "dia", é correspondente ao dia da aferição no formato dia/mês, tendo o período de primeiro de janeiro a trinta e um de dezembro, incluindo o dia vinte e nove de fevereiro, no qual são preenchidos apenas nos anos bissextos;
- As demais colunas possuem como título o ano da aferição, as quais possuem em suas linhas os valores de níveis do rio em centímetros;
- Após o dia trinta e um de dezembro é apresentada uma linha intitulada "ANO", com os anos novamente apresentados, que servem para melhor visualização do resumo dos dados apresentados nas próximas linhas;
- "Nº de dados", totaliza o número de dados em cada coluna (cada ano), valor que pode variar de ano para ano, de acordo com a disponibilidade dos dados;
- "DIA MÍN", possui o dia que apresenta o pico mínimo de nível de cada ano, no formato dia/mês;
- "MÍN", representa o valor em cm do pico mínimo do dia, indicado na linha "DIA MÍN";
- "MÉD", possui a média anual do nível em cm;

- "MÁX", representa o valor em cm do pico máximo do dia, indicado na linha abaixo, intitulada "DIA MÁX";
- "DIA MÁX", possui o dia que apresenta o pico máximo de nível de cada ano.

Com o objetivo de realizar uma segunda análise dos dados e prepará-los para o treinamento com algoritmos de AM, os dados foram convertidos, com o auxílio de um *script* desenvolvido na linguagem de programação `Python`, como pode ser observado na Figura 4.2. Em seguida os dados foram armazenados no formato `csv`.

dia	**...**	**nome estação**	**nome estação**	**...**
dia/mês/Ano	...	nível em cm	nível em cm	...
dia/mês/Ano	...		nível em cm	...
⋮	⋮	⋮	⋮	⋮
dia/mês/Ano	...	nível em cm	nível em cm	...

Figura 4.2: Estrutura da base de dados convertida.

- Primeira coluna, intitulada “dia” contém o dia da aferição, no formato dia/mês/ano, já organizada de maneira cronológica, respeitando os anos bissextos;
- As demais colunas possuem o nome da estação, em que cada linha de cada coluna possui o valor em cm da aferição do nível do rio. Entre esses valores podem aparecer níveis com valor zero, ou até mesmo valores negativos, considerados volume morto do rio, ou seja, volume abaixo do nível operacional, sendo assim, o *script* considerou o valor da linha como vazio apenas quando não continha dados, exemplo destacado em azul na Figura 4.2.

Após a conversão, foi realizada análise exploratória nos dados, sobre os quais foram levantadas algumas informações para melhor entendimento da base de dados, descritos na Tabela 4.1.

Estação	**Ano de início**	**Ano de término**	**Qtd. de amostras**	**Qtd. de dados ausentes**	**Média**	**Desvio padrão**	**Valor mínimo**	**Valor máximo**
Cáceres	1965	2018	19.192	133	294,12	122,98	68	640

Descalvados	1967	2006	14.081	609	416	75,71	230	545
Porto Conceição	1967	2018	18.408	3.725	414,81	75,67	207	573
Bela Vista do Norte	1967	2018	18.398	1.254	420,60	95,77	200	712
Cuiabá	1933	2018	31.226	1.377	213,11	161,67	-11	1.083
Porto Cercado	1969	2018	17.673	1.787	264,87	134,21	14	524
Córrego Grande	1969	2018	17.782	2.143	186,29	110,23	14	575
São José do Boriréu	1968	2018	18.131	4.964	229,31	78,29	66	352
São Jerônimo	1967	2018	18.366	1.914	276,83	74,95	149	522
São José do Piquiri	1969	2018	18.060	4.004	293,87	105,47	141	592
Porto Zé Viana	1967	2018	18.485	1.119	477,03	91,56	229	645
Porto São Francisco	1967	2018	18.492	2.176	576,83	135,79	206	878
Ladário	1900	2018	43.279	6	273,42	151,21	-61	664
Coxim	1966	2018	19.173	1.376	301,47	81	116	597
Palmeiras	1965	2018	19.436	2.444	212,25	91,72	83	891
Aquidauana	1968	2018	18.431	1.077	329,77	131,49	91	1.000
Estrada MT-738	1969	2018	17.775	1.493	183,65	126,19	78	1.081
Miranda	1965	2018	19.239	1.499	323,43	174,34	76	790
Porto Esperança	1963	2018	19.917	0	266,11	164,23	0	654
Forte Coimbra	1961	2018	20.999	1.298	227,64	150,18	-84	543
Barranco Branco	1965	2017	19.141	2.046	416,36	189,14	19	980

Porto Murtinho	1939	2018	29.035	945	426	165,15	73	971

Tabela 4.1: Informações sobre a base de dados.

Foi identificado que cada estação possui uma quantidade diferente de amostras, no decorrer do tempo, e que há quatro estações que se mostram com uma diferença considerável em relação às demais, sendo:

- Ladário, que possui a maior série histórica, com 43.279 amostras;
- Cuiabá, que possui a segunda maior série histórica, com 31.226 amostras;
- Porto Murtinho, que possui a terceira maior série histórica, com 29.035 amostras;
- Descalvados, que possui a menor série histórica, com 14.081 amostras.

As demais estações não apresentam diferenças exorbitantes de amostras entre elas, como observado na Figura 4.3.

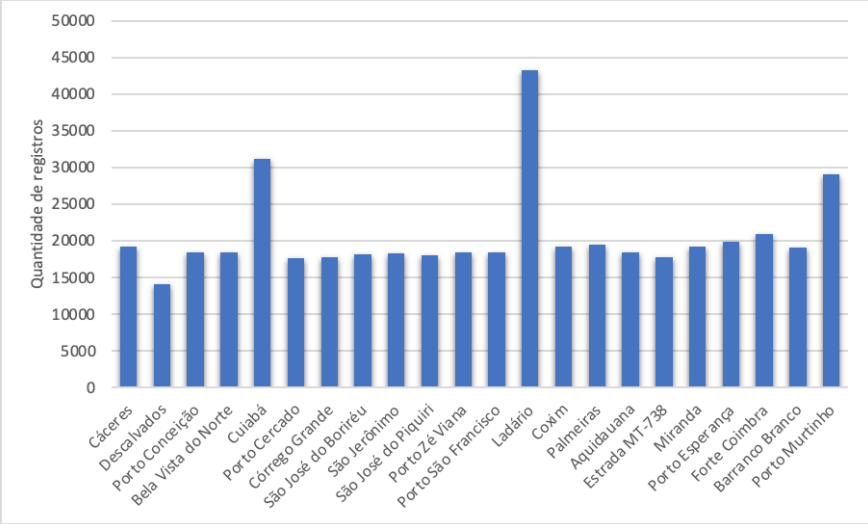

Figura 4.3: Diferença de amostras entre as estações.

A pouca diferença na quantidade de amostras de algumas séries não significa que seja possível criar uma base de dados com o número total de amostras, visto que essas possuem diferenças entre a data de início e a data de término da série.

A maioria das séries possuem término no ano de 2018, em alguns casos em dias e meses diferentes; já em relação ao início da série, é possível identificar na Tabela 4.1 que a maioria possui uma proximidade do ano de início, mas, ainda assim, há uma variação, além dos casos de maior quantidade de amostras, como Ladário, que se inicia no ano de 1900.

Com o objetivo de montar uma base de dados representativa, com as estações disponibilizadas, as estações foram plotadas e analisadas uma a uma, isso permitiu identificar as ausências de dados dentro do período da série. Alguns casos são mínimos, como ocorre na estação de Ladário, e, em outros casos, são preocupantes, pois assumem espaços de anos de ausência de dados.

Em alguns casos, as ausências interrompem sequências de dados que formariam uma série de dados considerável. Para melhor visualização, é possível observar as estações de Porto Conceição e São José do Boriréu, na Figura 4.4, esses são os dois casos mais críticos.

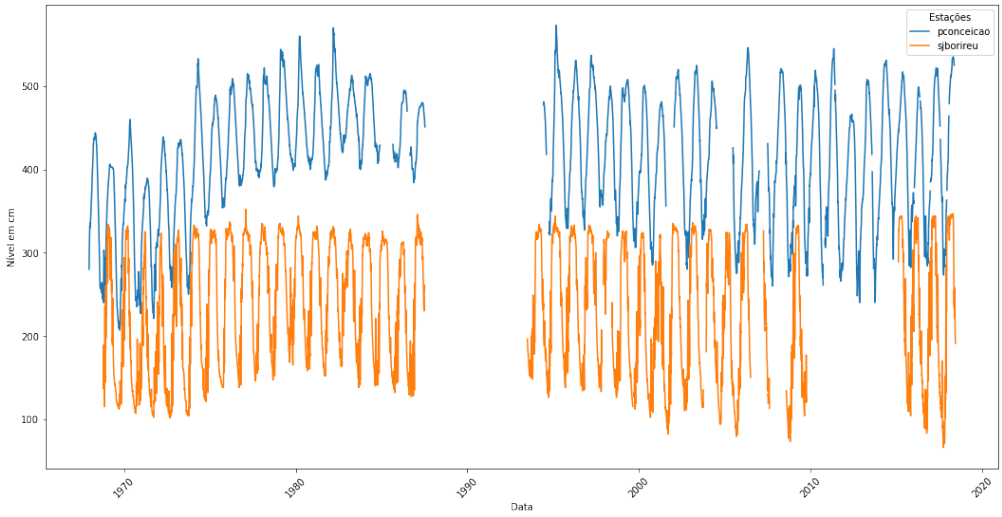

Figura 4.4: Dados das estações de São José do Boriréu e Porto Conceição.

Nas Figuras 4.5 e 4.6 podem ser observadas, respectivamente, as estações de Ladário com apenas 6 dados ausentes e Porto Esperança com nenhum dado ausente. Trata-se dos dois melhores casos entre as vinte e duas estações.

Outro aspecto nítido, nas séries apresentadas, é a representação dos períodos de cheias e estiagens severas. Se analisados por períodos, é possível observar que, antes de 1974 foi o período mais seco da RH-Paraguai, e o comportamento entre subidas e descidas do nível vai passando por todas as estações, começando com as descidas das águas das estações que se situam no planalto, percorrendo

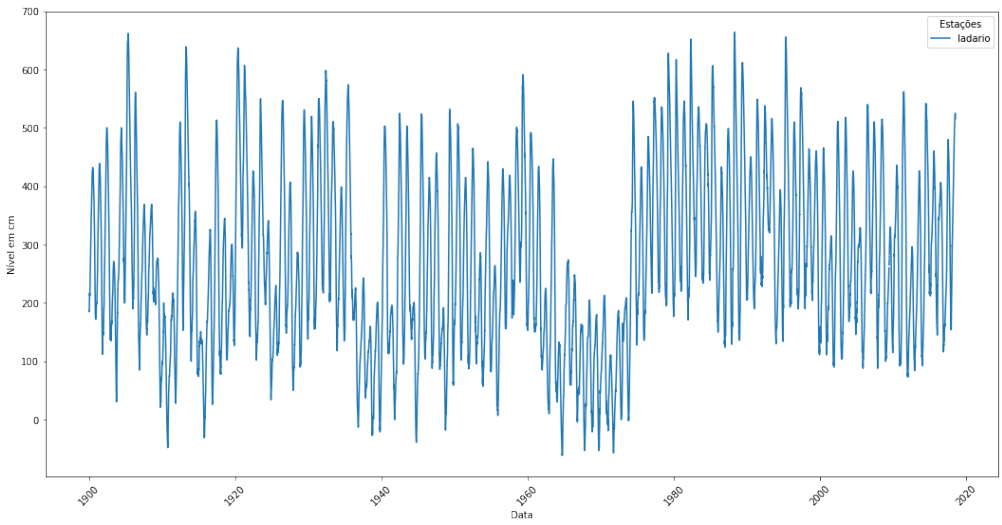

Figura 4.5: Dados da estação de Ladário.

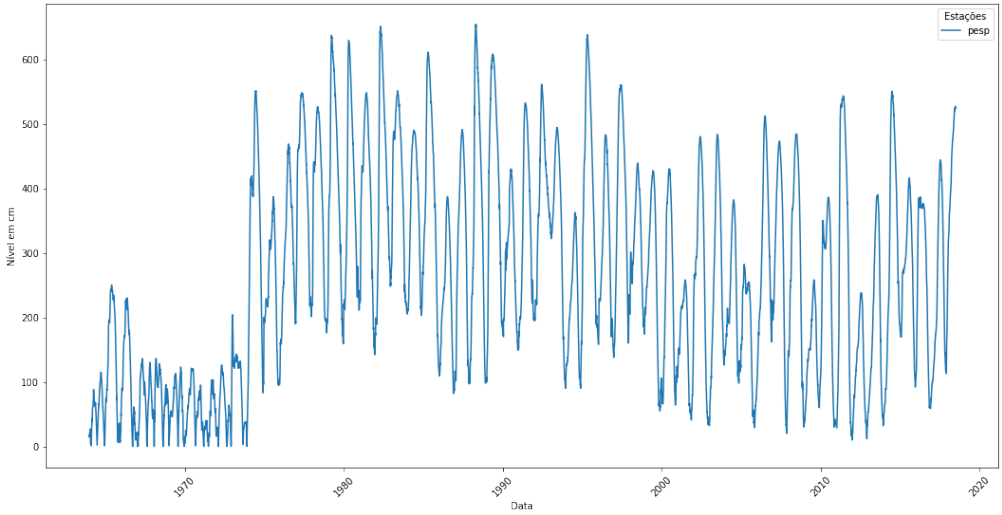

Figura 4.6: Dados da estação de Porto Esperança.

a bacia.

Na Figura 4.7, observa-se a quantidade de dados ausentes nas séries das estações, fato preocupante, pois inviabiliza a utilização de algumas séries.

Na Seção 4.1.1, a seguir, está descrito o processo de seleção das amostras de dados para o desenvolvimento do estudo.

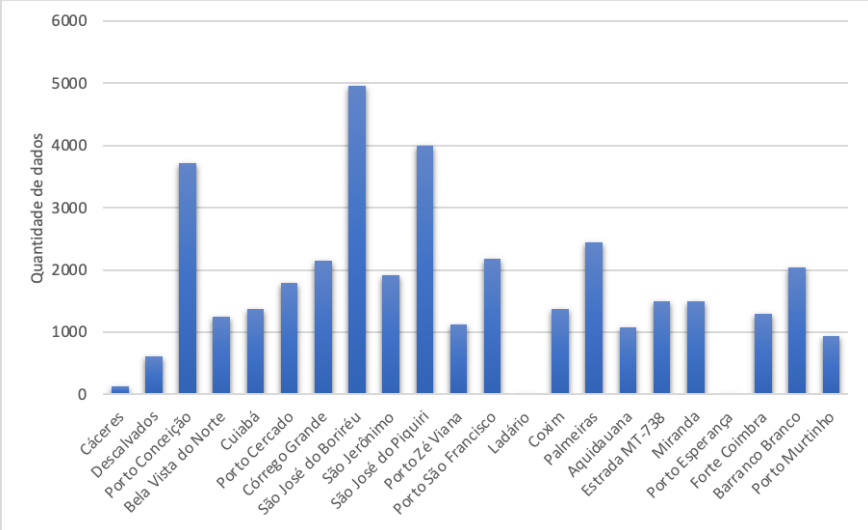

Figura 4.7: Quantidade de dados ausentes.

4.1.1 Seleção das Amostras

Os dados ausentes comprometem o tamanho das séries e, consequentemente, sua utilização na modelagem. Por esse motivo, em um primeiro momento, foi realizada a tentativa de preenchimento, utilizando média móvel e o algoritmo KNN, que é conhecido por considerar os vizinhos mais próximos para realizar suas operações.

Porém, como os intervalos de dados faltantes consecutivos são grandes, os períodos de dados ausentes acabam se tornando uma reta.

Como já discutido no Capítulo 3, os trabalhos similares ao presente estudo não descrevem, com detalhes, como foram trabalhados os dados ausentes. Em sua grande maioria, apenas descrevem os dados ou períodos que utilizaram na modelagem.

Quando utilizam o preenchimento para grande quantidade de dados ausentes, os pesquisadores não deixam clarividente a dimensão dessas ausências de dados, nem detalham a utilização do método, fato que geram dúvidas se isso poderia ser aplicado no presente trabalho.

As previsões com a utilização de dados a montante podem se tornar uma alternativa; há previsões que se utilizam de dados da própria série, sendo que pode ser feita a seleção de estações que contenham dados a montante para a realização da previsão das estações a jusante com dados ausentes.

Foram selecionadas para este estudo as estações de Bela Vista do Norte, Porto

Zé Viana e Ladário, por serem estações que formam o efeito de um funil da RH-Paraguai Norte para dentro da bacia (rio Paraguai), como pode ser observado nas linhas em vermelho na Figura 4.8. Bela Vista do Norte e Porto Zé Viana recebem toda a influência das águas que chegam das estações a montante; vale destacar, também, que Porto Zé Viana recebe águas dos rios Cuiabá e São Lourenço.

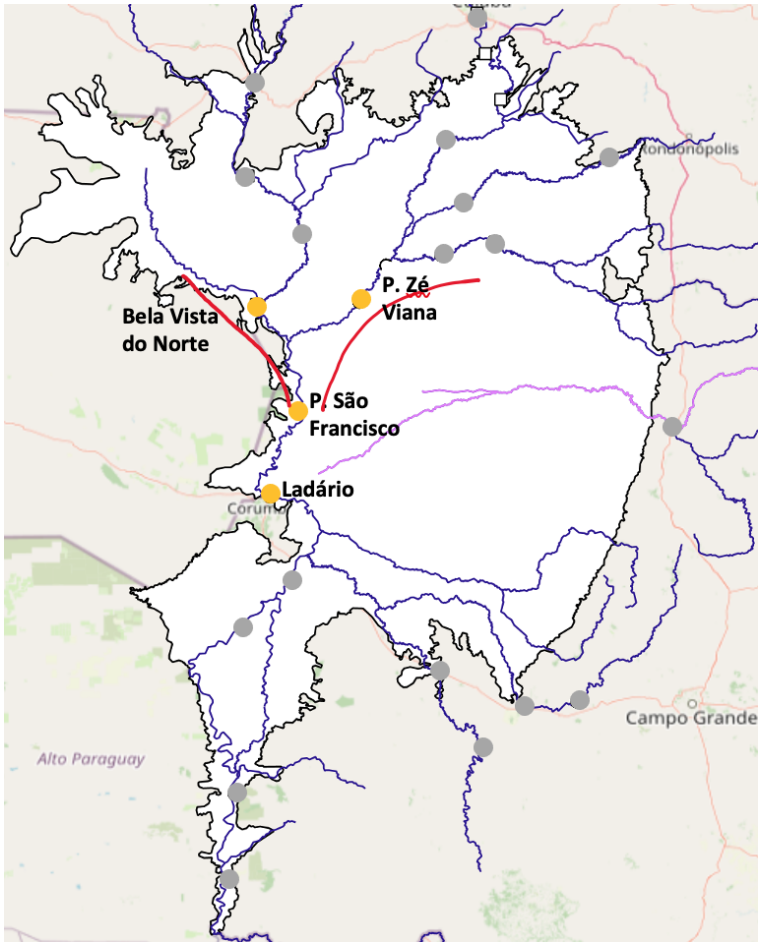

Figura 4.8: Localização das estações selecionadas.

Na Figura 4.9 é apresentada a direção do escoamento das águas para dentro da RH-Paraguai, onde é possível observar que toda a influência das águas passa por Ladário antes de seguir o curso da bacia.

A estação de Ladário se localiza a jusante da estação de São Francisco e foi selecionada, em vez da estação de Porto São Francisco, por possuir apenas 6 dados ausentes, enquanto Porto São Francisco possui 2.176 dados ausentes.

Ladário é uma estação de referência para o restante da bacia, pois, quando o

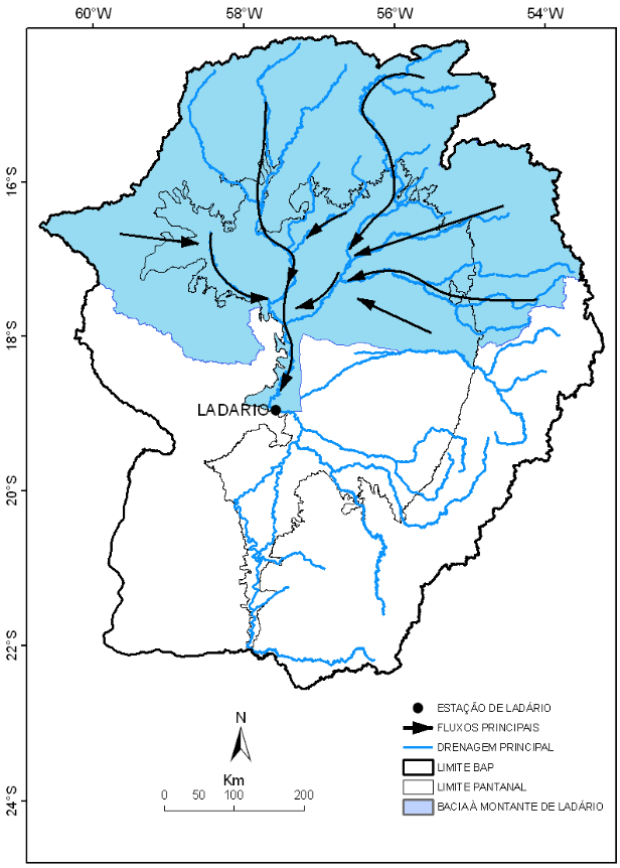

Figura 4.9: Fluxo a montante de Ladário, (Padovani, 2010).

nível da estação de Ladário começa a subir, pecuaristas e pantaneiros já preveem os efeitos através de conhecimento empírico.

A estação de Ladário também é importante para a navegação e demais atividades, pois a água que chegará no restante da bacia passa por essa estação, então a previsão do nível dessa estação se mostra importante para determinar a quantidade de dias que as cheias ou secas extremas chegarão ao restante da bacia.

Tendo em vista as estações definidas para o experimento, se fez necessária a procura por um período sem dados ausentes nas três estações mencionadas, sendo o maior período encontrado entre os anos de 1978 a 1988, Figura 4.10, totalizando 4.018 amostras.

Ao observar a Figura 4.10, é perceptível que há significativa defasagem no tempo entre os picos de níveis mínimos e máximos, os quais o movimento de subida e descida de nível ocorre em diferença de dias ou até meses entre uma estação e outra. Essa diferença de tempo é justamente o período que a água

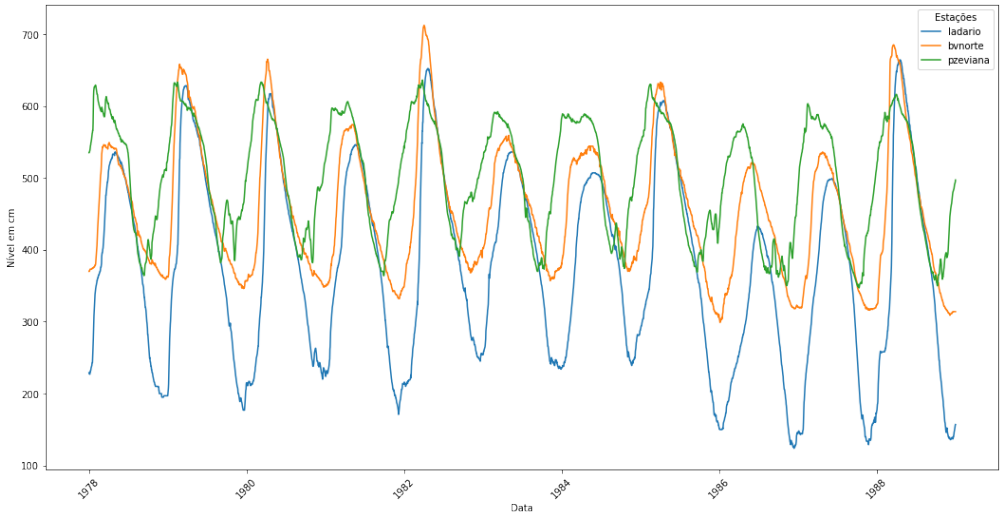

Figura 4.10: Séries selecionadas.

que passou por aquela estação leva para chegar até a próxima estação a jusante, seguindo o sentido do fluxo da bacia.

Na Seção 4.1.2, a seguir, é discutido o alinhamento das séries em relação à defasagem que sofrem no tempo.

4.1.2 Alinhamento das Séries

A Embrapa Pantanal realiza o alinhamento das séries no tempo, pois, atualmente, realiza previsões sem essa defasagem, sendo possível estimar o valor do nível e a quantidade de dias que a água chegará ao nível estimado.

DIAS et al. (2011) trabalharam com dados quinzenais e utilizou a correlação cruzada para determinar a defasagem entre as séries temporais do rio Paraguai, em Ladário, e a área inundada no Pantanal, concluindo que a defasagem, nesse caso, é de 45 dias.

O processo de alinhamento utilizado no presente projeto foi o mesmo utilizado pela Embrapa Pantanal, no qual foi reproduzido o processo realizado atualmente.

O alinhamento consiste em manter as estações lado a lado e ir calculando a correlação entre elas.

O primeiro passo foi manter os dados das estações lado a lado, como no exemplo, com valores fictícios da Figura 4.11, e, em seguida, foi calculada a correlação entre as duas, verificando o R^2, para saber se é uma correlação forte.

Em seguida, os dados da Estação 2 foram deslizando, de modo que o pri-

Dia	Estação 1	Estação 2
01/01/1978	109	209
02/01/1978	110	210
03/01/1978	111	211
04/01/1978	112	212
05/01/1978	113	213
⋮	⋮	⋮
30/12/1988	300	400
31/12/1988	301	401

Figura 4.11: Estações lado a lado.

meiro dia fique sem valor e o último valor sem data. Novamente foi calculada a correlação, verificando se o R^2 aumentou, como na Figura 4.12.

Dia	Estação 1	Estação 2
01/01/1978	109	
02/01/1978	110	209
03/01/1978	111	210
04/01/1978	112	211
05/01/1978	113	212
⋮	⋮	213
30/12/1988	300	⋮
31/12/1988	301	400
		401

Figura 4.12: Estações lado a lado, segundo passo.

Esse processo foi repetido até que o R^2 comece a cair, dessa forma, foi utilizado o ponto em que a correlação foi mais forte, ou seja, maior R^2.

Foi selecionado o intervalo que contém dados em ambas as estações. A quantidade de dias que a água demora para sair da Estação 2, a montante, e chegar na Estação 1, à jusante, é a quantidade de dias que a Estação 2 deslizou para baixo, na Figura 4.13, os exemplos são três dias. A área com borda vermelha indica o período da série alinhada, que será utilizada na modelagem.

No presente trabalho, a série selecionada de Ladário foi alinhada com as séries de Bela Vista do Norte e Porto Zé Viana, tendo, como resultado, a defasagem de 71 dias e um total de 3.947 amostras de níveis.

Dia	Estação 1	Estação 2
01/01/1978	109	
02/01/1978	110	
03/01/1978	111	
04/01/1978	112	209
05/01/1978	113	210
⋮	⋮	211
30/12/1988	300	212
31/12/1988	301	213
		⋮
		400
		401

Figura 4.13: Estações lado a lado com defasagem ajustada.

O resultado são os picos de mínimas e máximas das estações alinhados no tempo, ou seja, os picos de níveis mínimos estão no mesmo período para as três estações, e o mesmo processo ocorre com os picos de níveis máximos, Figura 4.14.

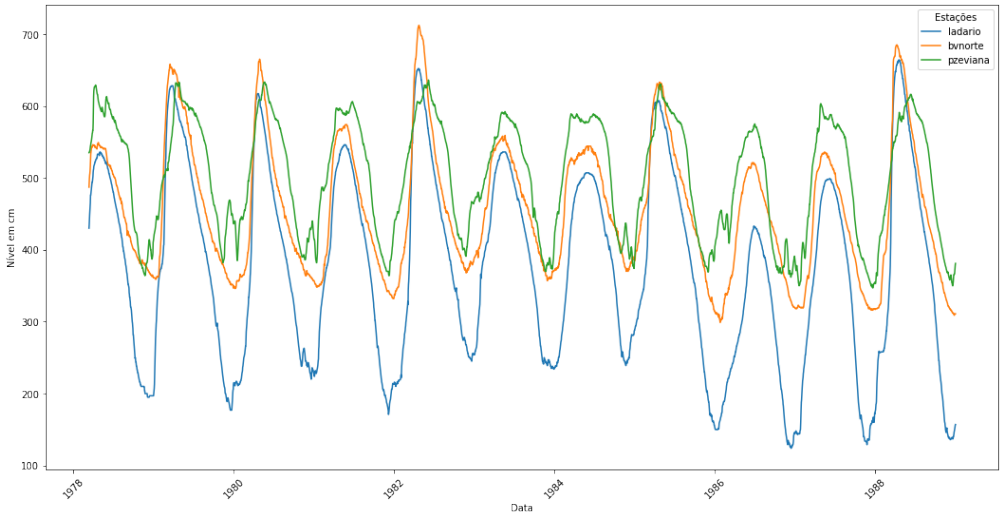

Figura 4.14: Séries selecionadas após o ajuste.

Na Seção 4.2, a seguir, são apresentadas as etapas e o planejamento para a execução da modelagem.

4.2 Modelagem

Definidos os dados a serem utilizados após o alinhamento das séries, os quais resultaram em um total de 3.947 amostras de níveis, se faz necessário o planejamento da modelagem com os algoritmos de AM.

Primeiramente, foram definidos os algoritmos a serem utilizados, listados abaixo:

- LSTM: Por ser o mais utilizado, atualmente, para previsões em trabalhos similares, como já discutido no Capítulo 3;
- GRU: Dentre as variações de LSTM, presentes no Capítulo 3, GRU chamou a atenção por apresentar resultados superiores à LSTM em um dos trabalhos discutidos na revisão da literatura. O trabalho não possuía muitas amostras de dados em relação aos demais trabalhos, fato que, também, ocorre no presente trabalho;
- BiLSTM: Por ser uma abordagem que, mesmo muito utilizada para outras finalidades, apresenta bons resultados na literatura com séries temporais, espera-se que, no caso de uma série com poucas amostras, sejam alcançados bons desempenhos.

Ainda sobre GRU e LSTM, GRU é mais simples e tem a vantagem de maior facilidade de implementação e eficiência, enquanto LSTM seja preferível com uma quantidade maior de dados (Aggarwal, 2018).

Um dos desafios na modelagem é a seleção dos parâmetros e hiperparâmetros das redes, de forma que se obtenha o melhor desempenho que a rede possa oferecer.

Uma das formas de definir os hiperparâmetros da rede é a tentativa e erro, fato que gera um amplo tempo de testes e, dificilmente, chega-se a uma configuração ótima.

Uma forma de resolver esse problema é a utilização de ferramentas e a aplicação dos chamados *tunings* dos hiperparâmetros, em que a própria ferramenta executa a busca das melhores configurações, de acordo com um objetivo pré-definido.

Aggarwal (2018) menciona três formas de busca de hiperparâmetros: a primeira é busca em grade; a segunda, busca aleatória, e a última, com o uso de otimização Bayesiana.

A busca em grade é o método que irá testar todas as combinações de hiperparâmetros do intervalo definido para cada hiperparâmetro. Esse tipo de técnica tende a demorar muito, o que torna a sua prática inviável para problemas maiores ou quando se pretende fazer ampla busca para obter a melhor configuração possível.

Já a busca aleatória visa testar combinações aleatórias de hiperparâmetros, considerando a quantidade máxima de combinações que se deseja testar, o que pode gerar, no teste, muitas combinações desnecessárias, então, sabe-se que não geram bons resultados.

A otimização Bayesiana trabalha com informações já conhecidas, em que, de início, testam-se algumas combinações de hiperparâmetros, em seguida, aplicam-se os resultados em uma função gaussiana, para atualizar informações posteriores, de modo a determinar a melhor combinação de hiperparâmetros a testar.

O processo gaussiano é altamente flexível e fácil de manusear, portanto, a otimização Bayesiana aplica o processo gaussiano para ajustar os dados e atualizar a distribuição posterior (Wu et al., 2019).

Esse processo pode ser muito útil quando se possui problemas complexos, pois contribui para ser mais objetivo nas escolhas dos conjuntos de hiperparâmetros testados.

O *Framework* Optuna[1] foi utilizado para a busca de parâmetros e hiperparâmetros no presente estudo. Optuna implementa algoritmos similares e melhorados da otimização Bayesiana e constrói, de forma dinâmica, o espaço de busca.

Optuna refere-se a cada processo de otimização como um estudo, e a cada avaliação da função objetivo como uma tentativa (Akiba et al., 2019).

De acordo com Géron (2019), Keras fornece uma camada GRU, usá-la é apenas uma questão de substituir uma RNN simples ou LSTM por GRU.

As redes são muito parecidas em questão de hiperparâmetros, por esse motivo, na utilização dos três algoritmos foram definidos os mesmos intervalos de busca.

Os parâmetros e os hiperparâmetros configurados para otimização através do

[1] https://optuna.org/

Optuna seguem um apanhado geral do que foi encontrado e utilizado na literatura, com os intervalos ou as opções de busca listados na Tabela 4.2.

Hiperparâmetro	**Opções**
N° de camadas ocultas	1, 2, 4, 8, 16
Neurônios da camada oculta	1, 4, 8, 16, 32, 64, 128
Neurônios da camada LSTM	1, 4, 8, 16, 32, 64, 128
Taxa de Dropout	De 0 a 0.5
Função de ativação	relu, sigmoid, tanh, linear
batchsize	32, 64, 128, 256

Tabela 4.2: Opções de busca de hiperparâmetros.

Algumas configurações foram fixadas, como por exemplo:

- Otimizador Adam, por ter demonstrado melhor desempenho em testes preliminares;
- Taxa de aprendizado com o valor inicial de $0,001$, por ser o valor padrão do otimizador Adam, já que esse otimizador executa o decaimento da taxa de aprendizado;
- Quantidade máxima de épocas, foi fixada em 100;
- Horizonte de predição, foi definido com um dia.

Para a escolha do melhor conjunto de hiperparâmetros, foi configurada, como função objetivo do Optuna, a métrica MSE, buscando o valor mínimo, ou seja, o conjunto de hiperparâmetros que apresentar o valor mínimo de MSE no treinamento e teste será o escolhido. A quantidade máxima de modelos a serem testadas foi configurada com o valor de 200, ou seja, serão testados 200 modelos.

Vale ressaltar que o valor de MSE é uma média das execuções da validação cruzada, pois foi aplicada a validação cruzada para séries temporais, pelos quais o número de *splits* foi definido com o valor 10, ou seja, 10 partes.

A validação cruzada foi configurada utilizando a função TimeSeriesSplit[2], da biblioteca `Sklearn`. Da mesma biblioteca foi utilizada a função `MinMaxScaler`[3], para a normalização dos dados.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html/

[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Para cada iteração da validação cruzada foi levado em conta que a base não conta com uma grande quantidade de amostras, dessa forma, essa base de dados foi dividida da seguinte forma:

- 200 registros do final da série para teste, devido à pouca quantidade de dados, sendo o número um inteiro, como pede a ferramenta utilizada;
- Do restante da base que se refere ao início da série essa foi dividida em 90% para treinamento;
- 10% para validação.

A divisão ocorre em cada um dos 10 passos, definidos na função TimeSeriesSplit.

Além da busca por hiperparâmetros ideais, outros testes foram realizados no intuito de encontrar o melhor modelo possível com a utilização de cada algoritmo, então foram testadas as configurações de:

- Janelas deslizantes de 30, 45 e 60 dias.
- Para cada janela, a rede foi treinada com paradas antecipadas de 10 e 20 épocas, além de mais um treinamento executando as 100 épocas.

O conceito de janelas deslizantes diz respeito à utilização de um grupo de dados passados para a previsão de um determinado dado futuro, em que os dados vão deslizando no tempo, de forma que o valor que é o alvo em um momento faça parte da janela no próximo instante. Figura 4.15.

De acordo com Khan et al. (2019), ao prever o próximo valor $x(t+1)$, a ideia é alimentar a rede não apenas com o valor $x(t)$, mas também com os valores $x(t-1)$, $x(t-2)$ e assim por diante.

Segundo Akbar et al. (2017), o tamanho da janela ideal é uma tarefa desafiadora e sem solução genérica. Um tamanho de janela grande pode ter resultados mais precisos, mas aumenta a complexidade do modelo, enquanto um tamanho de janela pequeno pode resultar em um erro maior.

Definir a quantidade de épocas de treinamento para um modelo é um outro desafio na modelagem, pois a utilização de poucas ou muitas épocas de treinamento podem ocasionar problemas de *underfitting* ou *overfitting*, pois não se

Figura 4.15: Exemplo de janelas.

aprendeu o necessário para generalizar o modelo, ou treinou a ponto de se ajustar demais aos dados.

Para resolver esse problema é utilizada a técnica de *Early Stop*, ou parada antecipada, que se trata de uma técnica de regularização.

De acordo com Géron (2019), à medida que as épocas passam, o algoritmo aprende, e seu erro de previsão, no conjunto de treinamento e validação, diminui naturalmente. No entanto, depois de um tempo, o erro de validação para de diminuir e, então, começa a voltar. Isso indica que o modelo começou a superajustar os dados de treinamento. Com a parada antecipada interrompe-se o treinamento assim que o erro de validação atinge o mínimo.

Dessa forma, quando o modelo passa a quantidade de épocas especificadas sem convergir o treinamento é interrompido.

Diante disso, foram executadas 9 buscas de hiperparâmetros e treinamento por algoritmo.

Todas as etapas, desde a escolha dos parâmetros, treinamento e teste dos modelos, foram realizadas por meio da linguagem de programação Python, *Framework* TensorFlow com Keras e o ambiente Jupyter Notebook.

Na Seção 4.3, a seguir, está descrita a forma de comparação e avaliação dos modelos.

4.3 Benchmarking

Após a modelagem, a próxima etapa é o *benchmarking*, que visa avaliar e comparar os resultados obtidos.

Para o processo de *benchmarking*, foram consideradas as métricas de desem-

penho descritas na Seção 2.3, que foram levantadas na literatura e utilizadas nos trabalhos similares.

Como descrito na Seção anterior, para cada algoritmo foi gerado um total de 9 modelos, com as variações de hiperparâmetros, janelas e parada antecipada.

Os 9 modelos de cada algoritmo são avaliados entre si, levando em consideração o desempenho aferido através das métricas estabelecidas. Ao final será selecionado o modelo de melhor desempenho para cada algoritmo.

Dessa forma, ao final tem-se um modelo de LSTM, um modelo de GRU e um modelo de BiLSTM, os quais serão comparados e avaliados entre si e em relação ao modelo de regressões que é utilizado hoje pela Embrapa Pantanal.

O modelo de regressão utilizado atualmente pela Embrapa Pantanal foi implementado na linguagem de programação Python, porém utilizando dados diários e não as máximas anuais, como é utilizado pela Embrapa Pantanal.

A utilização de dados diários, e a mesma divisão dos dados utilizados na modelagem com os algoritmos de AM, foram necessários para que ocorra a comparação justa, em que seja possibilitada a correta avaliação dos modelos.

As regressões foram implementadas comparando os dados das estações a montante, que são em Bela Vista do Norte e Porto Zé Viana, como variáveis independentes para uma estação de jusante, nesse caso Ladário, como variável dependente.

Serão apresentados os valores médios de erro das métricas MAE, MAPE, MSE e RMSE, resultantes das 10 execuções da validação cruzada, em que o modelo que apresentar o menor valor para as métricas será o modelo de melhor desempenho. Além da avaliação em relação às métricas, serão avaliados quais modelos se aproximam mais do valor real nos picos de níveis mínimos e máximos.

No Capítulo 5, a seguir, são apresentados os resultados e as discussões, tanto da busca por hiperparâmetros, quanto do treinamento e teste dos modelos.

CAPÍTULO

5

Resultados e Discussão

Neste Capítulo são apresentadas as discussão dos resultados, a comparação e avaliação entre eles.

Na Seção 5.1 são apresentados os resultados da etapa de busca por hiperparâmetros, com o *framework* Optuna, seguida pela Seção 5.2, em que são discutidos os resultados do treinamento e avaliação dos modelos.

5.1 Resultado da busca por hiperparâmetros

Os resultados obtidos com o *framework* Optuna foram bem distintos de um modelo para outro. O treinamento para a busca de hiperparâmetros seguiu o mesmo critério para todos os modelos em que foram testadas 200 combinações de hiperparâmetros por modelo.

Nas Tabelas 5.1, 5.2 e 5.3, podem ser observados os resultados da busca por hiperparâmetros para os algoritmos LSTM, GRU e BiLSTM respectivamente.

Nas tabelas, é possível observar que, para cada algoritmo foi gerado um total de 9 combinações de hiperparâmetros, correspondentes a 9 modelos. Foram alternadas as janelas de 30, 45 e 60 dias, e, para cada janela foi alternado o treinamento, com parada antecipada de 10 épocas, 20 épocas e sem a parada antecipada.

	Camadas ocultas	**Neurônios camada oculta**	**Neurônios camada LSTM**	***Dropout***	**Ativação**	***batchsize***
Janela de 30 dias						
P. antecipada de 10 épocas	2	64	32	0.207951	tanh	256
P. antecipada de 20 épocas	4	64	64	0.492409	tanh	128
Sem parada antecipada	2	8	16	0.165835	linear	128
Janela de 45 dias						
P. antecipada de 10 épocas	2	8	32	0.217986	tanh	64
P. antecipada de 20 épocas	1	32	64	0.498921	relu	256
Sem parada antecipada	2	16	16	0.307022	linear	32
Janela de 60 dias						
P. antecipada de 10 épocas	4	64	64	0.073952	linear	64
P. antecipada de 20 épocas	2	64	128	0.266607	linear	256
Sem parada antecipada	16	64	1	0.047828	linear	32

Tabela 5.1: Resultados da busca de hiperparâmetros - LSTM.

	Camadas ocultas	**Neurônios camada oculta**	**Neurônios camada LSTM**	***Dropout***	**Ativação**	***batchsize***
Janela de 30 dias						
P. antecipada de 10 épocas	1	8	64	0.438800	linear	256
P. antecipada de 20 épocas	4	16	1	0.210781	tanh	64
Sem parada antecipada	1	128	8	0.498262	linear	128
Janela de 45 dias						
P. antecipada de 10 épocas	1	128	8	0.481052	relu	128
P. antecipada	4	16	4	0.084492	relu	256

	Camadas ocultas	**Neurônios camada oculta**	**Neurônios camada LSTM**	***Dropout***	**Ativação**	***batchsize***
de 20 épocas						
Sem parada antecipada	1	16	64	0.356184	sigmoid	256
Janela de 60 dias						
P. antecipada de 10 épocas	4	64	8	0.434138	tanh	256
P. antecipada de 20 épocas	1	4	64	0.170924	tanh	128
Sem parada antecipada	2	8	4	0.200056	tanh	64

Tabela 5.2: Resultados da busca de hiperparâmetros - GRU.

	Camadas ocultas	**Neurônios camada oculta**	**Neurônios camada LSTM**	***Dropout***	**Ativação**	***batchsize***
Janela de 30 dias						
P. antecipada de 10 épocas	4	32	64	0.374494	tanh	128
P. antecipada de 20 épocas	4	32	32	0.314470	tanh	64
Sem parada antecipada	1	64	4	0.477819	linear	256
Janela de 45 dias						
P. antecipada de 10 épocas	2	32	8	0.325609	tanh	32
P. antecipada de 20 épocas	8	64	64	0.151479	linear	256
Sem parada antecipada	8	128	16	0.214337	tanh	256
Janela de 60 dias						
P. antecipada de 10 épocas	1	16	16	0.175605	tanh	128
P. antecipada de 20 épocas	1	8	32	0.069649	relu	128
Sem parada antecipada	2	32	1	0.022527	linear	256

Tabela 5.3: Resultados da busca de hiperparâmetros - BiLSTM.

5.2 Resultados do treinamento

Após a seleção dos hiperparâmetros foram efetuados treinamento e teste dos modelos. Na Tabela 5.4 são apresentadas as taxas médias de erros das métricas de desempenho e o desvio padrão resultantes dos modelos que utilizaram o algoritmo LSTM.

Além das taxas, é apresentada a quantidade média de épocas que cada modelo foi treinado, pois foram realizados testes com parada antecipada de 10 épocas, 20 épocas e um modelo sem a parada antecipada, o qual foi treinado por 100 épocas.

É evidente que, com exceção da execução da GRU, com janela de 30 dias e parada antecipada de 10 épocas, o restante dos modelos tende a melhorar o resultado à medida que se aumenta o tamanho da janela, isso demonstra que a quantidade maior de dados pode melhorar o treinamento da rede.

O modelo que obteve melhor desempenho com LSTM está destacado em negrito, sendo que possui janela de 60 dias, foi treinado por 33 épocas com parada antecipada de 20 épocas e obteve as menores taxas de erro.

Modelo	**MAE**	**MAPE**	**MSE**	**RMSE**	**Épocas**
Janela de 30 dias					
P. antecipada	23.3190	8.5580%	908.0780	27.4920	16
de 10 épocas	±11.3364	±5.5227%	±803.6078	±13.0086	
P. antecipada	21.3080	8.1820%	876.8120	26.7110	40
de 20 épocas	±11.1765	±5.9276%	±797.4777	±13.4683	
Sem parada	21.4900	8.3220%	875.2259	26.8470	100
antecipada	±11.1105	±5.8225%	±778.8440	±13.0992	
Janela de 45 dias					
P. antecipada	21.9130	8.4590%	840.3319	26.7870	17
de 10 épocas	±9.9597	±5.7827%	±728.01374	±11.6816	
P. antecipada	22.7360	8.6600%	907.0570	27.8780	36
de 20 épocas	±10.2606	±5.3165%	±736.2566	±12.0155	
Sem parada	23.6010	9.0530%	970.8500	28.1030	100
antecipada	±12.4992	±6.0529%	±923.4221	±14.1879	
Janela de 60 dias					

Modelo	**MAE**	**MAPE**	**MSE**	**RMSE**	**Épocas**
P. antecipada	22.2760	8.2890%	966.8980	27.3740	14
de 10 épocas	±13.0318	±5.7433%	±960.5148	±15.5483	
P. antecipada	**21.1070**	**8.0090%**	**921.3659**	**26.6750**	**33**
de 20 épocas	**±13.0552**	**±5.9974%**	**±929.5767**	**±15.2662**	
Sem parada	23.9470	8.5410%	991.8190	28.4920	100
antecipada	±12.4597	±5.4136%	±861.3734	±14.1429	

Tabela 5.4: Resultados dos modelos LSTM.

Seguindo o mesmo critério, na Tabela 5.5 são apresentadas as taxa de erros das métricas de desempenho, resultantes dos modelos que utilizaram o algoritmo GRU.

Para os modelos que utilizaram o algoritmo GRU, o melhor resultado é do treinamento com a janela de 30 dias, porém, nesse caso, o melhor desempenho foi configurado com parada antecipada de 10 épocas e foi treinado por 23 épocas destacado em negrito.

Modelo	**MAE**	**MAPE**	**MSE**	**RMSE**	**Épocas**
Janela de 30 dias					
P. antecipada	**19.6250**	**7.5460%**	**710.5329**	**24.7610**	**23**
de 10 épocas	**±8.9090**	**±5.1163%**	**±581.6853**	**±10.4025**	
P. antecipada	22.5450	8.9270%	906.7190	27.9240	62
de 20 épocas	±10.2568	±5.9141%	±769.1883	±11.8767	
Sem parada	22.3720	8.2390%	962.6050	27.1050	100
antecipada	±13.6737	±5.8092%	±980.7326	±15.9126	
Janela de 45 dias					
P. antecipada	21.8760	8.0110%	883.7510	26.4980	25
de 10 épocas	±11.6164	±5.5364%	±848.1394	±14.2020	
P. antecipada	25.2790	9.2860%	1065.5999	30.0300	52
de 20 épocas	±11.9848	±5.9092%	±870.4440	±13.4918	
Sem parada	22.3580	8.8140%	995.8539	28.6150	100
antecipada	±10.8130	±5.9405%	±905.8177	±14.0248	
Janela de 60 dias					
P. antecipada	21.8260	8.7020%	855.8849	27.0770	19
de 10 épocas	±10.7360	±5.7405%	±718.0127	±11.6767	
P. antecipada	20.6620	8.0760%	836.3109	25.6980	30
de 20 épocas	±12.2189	±5.9370%	±846.1081	±13.9817	

Modelo	MAE	MAPE	MSE	RMSE	Épocas
Sem parada	21.0600	8.1560%	876.4000	26.4710	100
antecipada	±11.3681	±5.4662%	±835.2319	±13.9682	

Tabela 5.5: Resultados dos modelos GRU.

Os resultados dos modelos que utilizaram o algoritmo BiLSTM estão presentes na Tabela 5.6.

Para esse algoritmo, o melhor resultado está no modelo que utilizou a janela de dados de 60 dias, porém, nesse caso, o modelo foi configurado com parada antecipada de 10 épocas e foi treinado por 23 épocas, também destacado em negrito.

Modelo	MAE	MAPE	MSE	RMSE	Épocas
Janela de 30 dias					
P. antecipada	22.9350	8.8120%	896.9570	27.7720	19
de 10 épocas	±10.3577	±5.6566%	±744.2680	±11.8176	
P. antecipada	21.9990	8.3260%	916.6950	27.6130	35
de 20 épocas	±10.4700	±5.0853%	±794.4707	±13.0925	
Sem parada	21.7340	8.1620%	904.2629	26.9100	100
antecipada	±12.0455	±5.1741%	±846.4226	±14.1504	
Janela de 45 dias					
P. antecipada	21.6510	8.0620%	849.9220	26.5300	15
de 10 épocas	±10.2310	±4.7639%	±727.3168	±12.7410	
P. antecipada	22.8870	8.2390%	958.0379	27.3770	33
de 20 épocas	±13.8222	±6.2729%	±919.6807	±15.2242	
Sem parada	23.2530	8.9590%	1050.6720	28.5340	100
antecipada	±12.5354	±6.8532%	±1090.1050	±16.2131	
Janela de 60 dias					
P. antecipada	**20.8300**	**7.9460%**	**865.1890**	**26.2370**	**23**
de 10 épocas	**±10.9492**	**±5.6368%**	**±845.5858**	**±14.0220**	
P. antecipada	24.5810	9.3190%	1068.7570	29.6450	37
de 20 épocas	±12.0181	±6.4857%	±1011.0868	±14.5253	
Sem parada	22.5560	8.1710%	983.1220	27.8360	100
antecipada	±13.6439	±5.3365%	±931.8115	±15.2101	

Tabela 5.6: Resultados dos modelos BiLSTM.

No geral, todos os modelos apresentaram resultados muito próximos, por mais que tenham diferentes configurações de hiperparâmetros, não há variações muito grandes nos resultados; um item de destaque é o fato de dois algoritmos LSTM e BiLSTM terem o seu melhor resultado com modelos de janelas maiores, possibilitando que as redes tenham períodos maiores para aprenderem o comportamento dos dados.

Tendo selecionado o melhor modelo de cada algoritmo, o próximo passo foi comparar os resultados de cada um deles com o método atual de regressões.

Na Tabela 5.7 são apresentados os resultado da Regressão, do melhor modelo de LSTM, GRU e BiLSTM.

Os resultados das métricas de avaliação deixam evidente que os três modelos superaram o método utilizado, atualmente, que é a Regressão, e que o modelo GRU obteve superioridade em relação aos demais, também destacado em negrito.

Modelo	**MAE**	**MAPE**	**MSE**	**RMSE**
Regressão	25.7700	8.9090%	1084.7380	31.1790
	±9.2203	±4.2350%	±718.4998	±11.1861
LSTM	21.1070	8.0090%	921.3659	26.6750
	±13.0552	±5.9974%	±929.5767	±15.2662
GRU	**19.6250**	**7.5460%**	**710.5329**	**24.7610**
	±8.9090	**±5.1163%**	**±581.6853**	**±10.4025**
BiLSTM	20.8300	7.9460%	865.1890	26.2370
	±10.9492	±5.6368%	±845.5858	±14.0220

Tabela 5.7: Comparação com o método atual.

Todos as figuras em que são mostrados os valores originais da série de Ladário e os valores preditos pelos modelos, apresentam o desvio padrão considerando os resultados da métrica MAE.

Na Figura 5.1 é possível observar o resultado da regressão em suas 10 execuções da validação cruzada, em que se nota que todos os picos de máximas foram superestimados, já os picos mínimos apresentam valores maiores em sua maioria. A regressão não estimou corretamente algumas subidas, enquanto as descidas se aproximaram do valor original.

Já na Figura 5.2 são apresentados os resultados da LSTM em suas 10 execuções da validação cruzada, em que é possível observar que os picos de máximas e de mínimas, em sua maioria, ficam mais próximos do valor original se comparado à regressão. O modelo tem uma performance melhor nas subidas quando

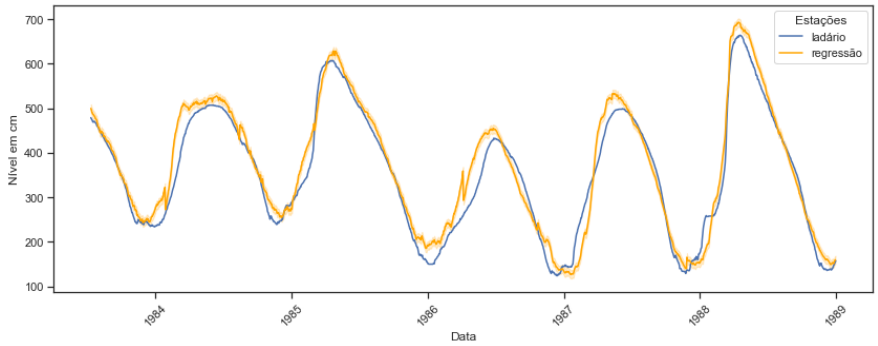

Figura 5.1: Resultado da Regressão.

comparadas com as descidas de níveis.

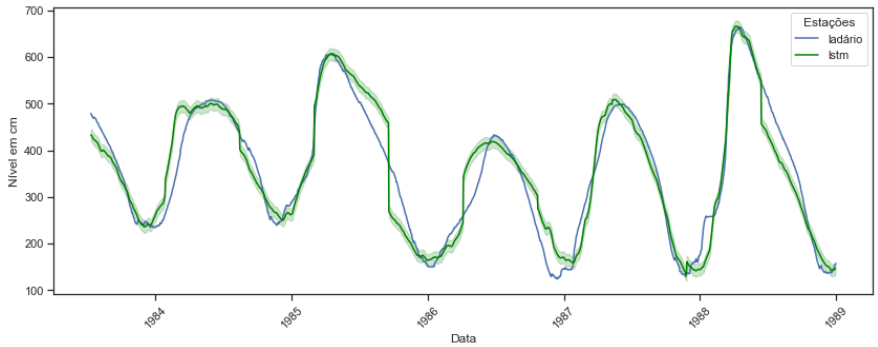

Figura 5.2: Resultado da LSTM.

Na Figura 5.3 estão os resultados da GRU, também em suas 10 execuções da validação cruzada, em que deixa evidente que o modelo não se aproximou tanto do valor real nos picos máximos e mínimos quando comparado com a LSTM. O modelo se aproxima mais do valor real na primeira e penúltima subida em comparação com a LSTM.

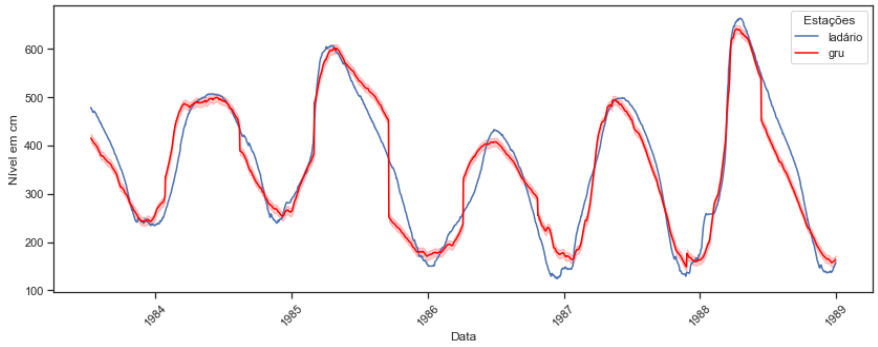

Figura 5.3: Resultado da GRU.

Na Figura 5.4 são apresentados os resultados da BiLSTM em suas 10 execuções da validação cruzada, em que o modelo se aproxima mais do valor real nos picos máximos quando comparado a GRU. O modelo se aproxima muito do resultado apresentado na LSTM, em que as subidas e descidas de nível, tanto na LSTM, quanto na BiLSTM, mostram-se semelhantes.

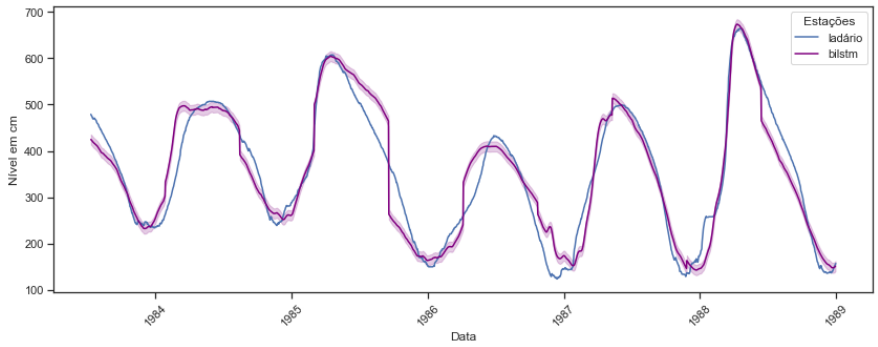

Figura 5.4: Resultado da BiLSTM.

Por fim, na Figura 5.5 estão presentes os resultados das predições da Regressão, da LSTM, GRU e BiLSTM juntas, com o intuito de visualizar as diferenças percebidas entre elas.

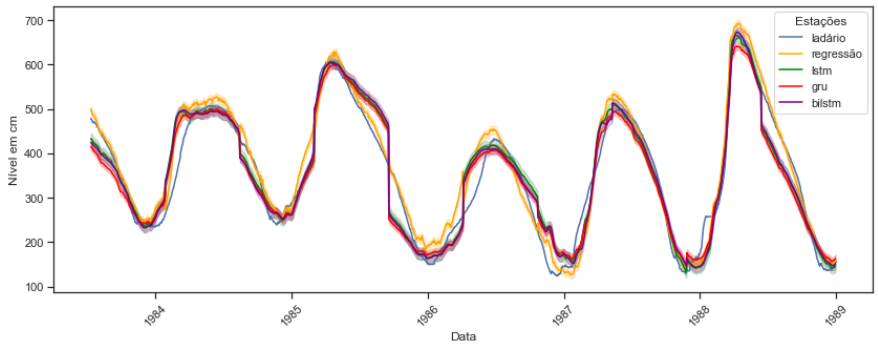

Figura 5.5: Resultados das predições juntas.

No Capítulo 6, a seguir, estão presentes as conclusões, principais contribuições, dificuldades e sugestões de trabalhos futuros.

CAPÍTULO

6

Conclusão

O presente trabalho teve como objetivo principal realizar a predição de níveis de rios da RH-Paraguai utilizando dados de nível de estações a montante para prever o nível de estações a jusante.

Para essa tarefa foram selecionadas amostras sem dados ausentes disponíveis, retirada a defasagem no tempo entre os dados das estações selecionadas, realizada a busca por hiperparâmetros ideais para a etapa de modelagem, efetuado o treinamento e avaliação dos modelos.

Após os passos citados acima, conclui-se que os três modelos testados podem realizar a predição de nível de rios da RH-Paraguai utilizando dados de nível de estações a montante para a previsão de estações a jusante.

Os três modelos apresentam resultados satisfatórios, inclusive com melhorias em relação ao modelo atual que utiliza regressões. O modelo com o algoritmo GRU se destacou por apresentar as menores taxas de erro e por ser $23,84\%$ mais preciso que o modelo de Regressão, já LSTM e BiLSTM são, respectivamente, $18,09\%$ e $19,16\%$ mais precisos que o modelo de Regressão. O modelo de LSTM e BiLSTM se aproximam mais do valor real nos picos de níveis máximos e mínimos quando comparados com a Regressão e com GRU.

Nas Seções seguintes são apresentadas as dificuldades e principais problemas encontrados no desenvolvimento do trabalho, as contribuições para o público

geral e comunidade científica e, por fim, sugestões de trabalhos futuros.

6.1 Dificuldades e principais problemas

Dentre as dificuldades encontradas, a maior se diz respeito à grande quantidade de dados ausentes, o que exigiu que fosse selecionada uma amostra menor sem dados ausentes.

Devido à quantidade de dados reduzida, não foi possível realizar testes com janelas de dados maiores, o que poderia melhorar o modelo.

Em relação à busca por hiperparâmetros, mostrou-se uma atividade desafiadora, que requer planejamento sistemático e pode levar longo período para que se encontre o resultado ideal. Como alternativa, foi utilizado o *framework* Optuna, que, mesmo facilitando o processo, levou tempo considerável para apresentar os resultados.

Foi realizado um estudo da Plataforma de Monitoramento, Análise e Alerta a Extremos Ambientais, denominada TerraMA2, desenvolvida pelo Instituto Nacional de Pesquisas Espaciais (INPE). Essa plataforma foi estudada e configurada em um servidor da Embrapa Pantanal, porém não foi possível a implementação dos modelos para a predição automatizada devido aos problemas com os dados.

A ANA disponibiliza um *WebService* com os dados de nível das estações da RH-Paraguai. Os dados possuem periodicidade de 15 minutos, sendo assim, foi implementado um *script* na linguagem de programação Python que realiza o download e configura os dados em um formato aceito pelo TerraMA2, porém os dados atuais também possuem muitos dados ausentes, o que impossibilitou a implementação das previsões na plataforma.

Por fim, uma limitação do modelo é que, para essa abordagem, a quantidade de dias da previsão é a quantidade de dias da defasagem retirada, ou seja, se foram retirados 71 dias de defasagem, quer dizer que depois de 71 dias a estação a jusante estará com o nível previsto pelos modelos.

6.2 Contribuições

Apesar das dificuldades, o trabalho apresenta contribuições que podem auxiliar, tanto a comunidade científica, quanto a comunidade em geral.

O trabalho comprova que é possível a utilização de dados a montante para a previsão de estações a jusante, o que pode ser uma alternativa frente à falta de dados, em que, em situações de dados faltantes, podem ser utilizadas outras estações a montante. Esse tipo de previsão não foi encontrada na literatura para a RH-Paraguai.

Outra contribuição é referente ao TerraMA2, pois, mesmo que não tenha sido possível a utilização dos modelos na Plataforma, ela se encontra configurada, sendo alimentada com dados diários de nível das estações da RH-Paraguai através do *WebService* da ANA. Mesmo que se encontre períodos com dados ausentes, a plataforma já se encontra configurada e disponível para a consulta da população e instituições interessadas.

A Plataforma também foi alimentada com dados estáticos de mapeamento da RH-Paraguai e está em constante processo de estudo de viabilidade para inclusão de novos dados.

Plataforma disponível em: http://geohidropantanal.cpap.embrapa.br/terrama2/monitor/.

O trabalho pode auxiliar a Embrapa Pantanal na tarefa de previsões de níveis dos rios da RH-Paraguai. Também pode contribuir com instituições na definição de políticas públicas e ações que visem à mitigação dos impactos causados por esses eventos extremos.

Há uma contribuição referente à busca por hiperparâmetros, pois os trabalhos encontrados na literatura não descrevem com clareza o processo de busca por hiperparâmetros, já o presente trabalho deixa clarividente os critérios utilizados e a forma como foi utilizado o *framework* para a busca de hiperparâmetros.

Por fim, tem-se uma contribuição que também não foi encontrada na maioria dos trabalhos revisados na literatura, que é a de disponibilizar os códigos utilizados na busca por hiperparâmetros, treinamento das redes e download de dados do *WebService* da ANA, os quais podem contribuir para o desenvolvimento de demais trabalhos. Os códigos estão disponíveis em git@github.com:rogerioalves/predicao_rh_paraguai.git e o acesso deve ser solicitado via email: rogerio.antoniassi@ifms.edu.br.

6.3 Trabalhos futuros

No desenvolvimento do trabalho foram surgindo possibilidades que servem como sugestões para a realização de trabalhos futuros:

1. Desenvolver ou aplicar técnicas de preenchimento de dados ausentes nos dados da RH-Paraguai;
2. Com a melhoria nos dados ausentes, replicar o estudo para a previsão de nível das demais estações da bacia (22 no total);
3. Realizar estudos sobre o alinhamento das séries no tempo (retirando a defasagem do tempo), com técnicas mais sofisticadas e que possam se tornar operacionais no sistema de alertas da Embrapa Pantanal, tal como a técnica DTW utilizada por Marcacini et al. (2016) e similares;
4. Incorporar mais dados de fatores hidrológicos, ou que contribuem para a ocorrência das cheias severas na RH-Paraguai, tais como chuva e vazão;
5. Testar o algoritmo SVR/SVM que também é bem discutido na literatura para problemas hidrológicos, como nos trabalhos de Nguyen and Le (2019), Sahoo et al. (2021) e Dubey and Katarya (2021).

Todas as possibilidades de estudos citadas acima possivelmente tornarão os modelos mais robustos e, no caso do SVR, investigar novas soluções para as previsões. Espera-se que essas possibilidades possam contribuir, ainda mais, para a proposta de previsão de nível dos rios da RH-Paraguai e abrir novos campos de estudo.

Referências Bibliográficas

Aggarwal, C. C. (2018). *Neural Networks and Deep Learning: A Textbook*. Springer. 14, 16, 17, 46

Akbar, A., Khan, A., Carrez, F., and Moessner, K. (2017). Predictive analytics for complex iot data streams. *IEEE Internet of Things Journal*, 4(5):1571–1582. 49

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery. 47

ANA (2018). *Plano de Recursos Hídricos da Região Hidrográfica do Rio Paraguai - Relatório Final*. Agência Nacional de Águas. http://arquivos.ana.gov.br/portal/RH-Paraguai/Plano.pdf, (Acesso em 20 de outubro de 2019). 1, 2

Assem, H., Ghariba, S., Makrai, G., Johnston, P., Gill, L., and Pilla, F. (2017). *Urban Water Flow and Water Level Prediction Based on Deep Learning*, volume 10536 LNAI of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Cited By :21. 29

Awad, M. and Khanna, R. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress. 9, 10

Biolchini, J., Mian, P. G., Natali, A. C. C., and Travassos, G. H. (2005). Systematic review in software engineering. *System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES*, 679(05):45. 20

Buduma, N. and Locascio, N. (2017). *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms*. O'Reilly Media, 1 edition. vi, 11, 15

Choi, C., Kim, J., Han, H., Han, D., and Kim, H. S. (2020). Development of water level prediction models using machine learning in wetlands: A case study of upo wetland in south korea. *Water (Switzerland)*, 12(1). Cited By :23. 29

CPRM (2006). *Recursos Hídricos*. CPRM - Companhia de Pesquisa de Recursos Minerais/Serviço Geológico do Brasil(CPRM/SGB). http://www.cprm.gov.br/publique/media/informacao_publica/recursoshidricos.pdf, (Acesso em 21 de março de 2020). vi, 2

Dazzi, S., Vacondio, R., and Mignosa, P. (2021). Flood stage forecasting using machine-learning methods: A case study on the parma river (italy). *Water (Switzerland)*, 13(12). Cited By :3. 29

DIAS, R. d. P., DIAS, C. d. S., Padovani, C. R., and Vettorazzi, C. A. (2011). Correlação cruzada e modelo de regressão não linear logístico na determinação das relações hidrológicas do pantanal. In *Embrapa Pantanal-Artigo em anais de congresso (ALICE)*. Revista da Estatística da Universidade Federal de Ouro Preto, v. 1, n. 1, p 43

Ding, Y., Zhu, Y., Feng, J., Zhang, P., and Cheng, Z. (2020). Interpretable spatio-temporal attention lstm model for flood forecasting. *Neurocomputing*, 403:348–359. Cited By :38. 29, 31

Ding, Y., Zhu, Y., Wu, Y., Jun, F., and Cheng, Z. (2019). Spatio-temporal attention lstm model for flood forecasting. In *Proceedings - 2019 IEEE International Congress on Cybermatics: 12th IEEE International Conference on Internet of Things, 15th IEEE International Conference on Green Computing and Communications, 12th IEEE International Conference on Cyber, Physical and Social Computing and 5th IEEE International Conference on Smart Data, iThings/GreenCom/CPSCom/SmartData 2019*, pages 458–465. Cited By :8. 29

Dubey, V. and Katarya, R. (2021). *An analysis of machine learning techniques for flood mitigation*, volume 1166 of *Advances in Intelligent Systems and Computing*. 4, 64

Felizardo, K. R., Nakagawa, E. Y., Fabbri, S. C. P. F., and Ferrari, F. C. (2017). *Revisão sistemática da literatura em engenharia de software: teoria e prática*. Elsevier. 20

Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M., and Lin, Q. (2020). Short-term runoff prediction with gru and lstm networks without requiring time step optimization during sample generation. *Journal of Hydrology*, 589. Cited By :51. 29, 31

Ghorpade, P., Gadge, A., Lende, A., Chordiya, H., Gosavi, G., Mishra, A., Hooli, B., Ingle, Y. S., and Shaikh, N. (2021). Flood forecasting using machine learning: A review. In *2021 8th International Conference on Smart Computing and Communications: Artificial Intelligence, AI Driven Applications for a Smart World, ICSCC 2021*, pages 32–36. 3

Goodfellow, I., Yoshua, J. B., and Courville, A. (2016). *Deep learning: adaptive computation and machine learning*. Adaptive Computation and Machine Learning series. The MIT Press. 9, 10, 13, 14, 16

Gude, V., Corns, S., and Long, S. (2020). Flood prediction and uncertainty estimation using deep learning. *Water (Switzerland)*, 12(3). Cited By :7. 29, 31

Géron, A. (2019). *Hands on Machine Learning with Scikit Learn Keras and TensorFlow*, volume 2nd Edition. O'Reilly Media, early release edition. 9, 12, 13, 15, 17, 18, 47, 50

Han, H., Choi, C., Jung, J., and Kim, H. S. (2021). Deep learning with long short term memory based sequence-to-sequence model for rainfall-runoff simulation. *Water (Switzerland)*, 13(4). Cited By :3. 29, 30

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*. 14

Khairudin, N. M., Mustapha, N., Aris, T. N. M., and Zolkepli, M. (2021). A study to investigate the effect of different time-series scales towards flood forecasting using machine learning. *Journal of Theoretical and Applied Information Technology*, 99(24):5687–5699. 29

Khan, I. A., Akber, A., and Xu, Y. (2019). Sliding window regression based short-term load forecasting of a multi-area power system. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pages 1–5. 49

Lau, M. M. and Hann Lim, K. (2018). Review of adaptive activation function in deep neural network. In *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 686–690. 12

Malhotra, R. and Jain, A. (2011). Software effort prediction using statistical and machine learning methods. *International Journal of Advanced Computer Science and Applications*, 2(1):145–152. 18

Marcacini, R. M., Carnevali, J. C., and Domingos, J. (2016). On combining web-sensors and dtw distance for knn time series forecasting. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2521–2525. IEEE. 64

Morettin, P. A. and Toloi, C. M. (2006). *Análise de Séries Temporais*. BLUCHER. 7

Morovati, K., Nakhaei, P., Tian, F., Tudaji, M., and Hou, S. (2021). A machine learning framework to predict reverse flow and water level: A case study of tonle sap lake. *Journal of Hydrology*, 603. 29, 30

Mousavi, F. S., Yousefi, S., Abghari, H., and Ghasemzadeh, A. (2021). Design of an iot-based flood early detection system using machine learning. In *26th International Computer Conference, Computer Society of Iran, CSICC 2021*. Cited By :1. 29, 30, 31

Nguyen, T. . and Le, H. T. T. (2019). Water level prediction at tich-bui river in vietnam using support vector regression. In *Proceedings - International Conference on Machine Learning and Cybernetics*, volume 2019-July. Cited By :1. 29, 64

Padovani, C. R. (2010). *Dinâmica espaço-temporal das inundações do Pantanal*. PhD thesis, Universidade de São Paulo. vi, 42

Petneházi, G. (2019). Recurrent neural networks for time series forecasting. *arXiv:1901.00069*. 14

Phan, T. . . and Nguyen, X. H. (2020). Combining statistical machine learning models with arima for water level forecasting: The case of the red river. *Advances in Water Resources*, 142. Cited By :21. 29, 30

Phan, T.-T.-H., Émilie Poisson Caillault, Lefebvre, A., and Bigand, A. (2020). Dynamic time warping-based imputation for univariate time series data. *Pattern Recognition Letters*, 139:139–147. 31

Pomerat, J., Segev, A., and Datta, R. (2019). On neural network activation functions and optimizers in relation to polynomial regression. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 6183–6185. 12

Rasel, R. I., Nizam Uddin, M., Islam, F., and Haroon, A. (2018). Application of deep neural network for predicting river tide level. In *2018 International Conference on Innovations in Science, Engineering and Technology, ICISET 2018*, pages 311–316. Cited By :1. 29

Rios, R. A. (2013). *Modelagem de séries temporais por meio da decomposição e análise de influências estocásticas e determinísticas.* PhD thesis, University of São Paulo, Brazil. vi, 8

Sahoo, A., Samantaray, S., and Ghose, D. K. (2021). Prediction of flood in barak river using hybrid machine learning approaches: A case study. *Journal of the Geological Society of India*, 97(2):186–198. Cited By :2. 29, 64

Siami-Namini, S., Tavakoli, N., and Namin, A. S. (2019). The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 3285–3292. 16

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Pearson. 11

Wang, S., Zhao, J., Shao, C., Dong, C., and Yin, C. (2020). Truck traffic flow prediction based on lstm and gru methods with sampled gps data. *IEEE Access*, 8:208158–208169. 17

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., and Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on bayesian optimizationb. *Journal of Electronic Science and Technology*, 17(1):26–40. 47

Yan, L., Chen, C., Hang, T., and Hu, Y. (2021). A stream prediction model based on attention-lstm. *Earth Science Informatics*, 14(2):723–733. Cited By :1. 29

Yan, L., Feng, J., and Hang, T. (2019). *Small watershed stream-flow forecasting based on LSTM*, volume 935 of *Advances in Intelligent Systems and Computing*. Cited By :9. 29