Predição do nível de água utilizando os modelos ARIMA e Random Forest: Um Estudo de Caso da Barragem Eclusa do São Gonçalo

Paulo Ricardo B.Dutra Lima[1],**Felipe Marques**[1], **Sabrina Orth**[2]

[1]Universidade Federal de Pelotas (UFPEL)
Pelotas – RS – Brasil

[2]Instituto Federal Farroupilha (IFFAR)
São Borja – RS – Brasil

{paulo.lima,felipem}@inf.ufpel.edu.br,sabrina.orth@iffarroupilha.edu.br

Abstract. *Artificial intelligence models have been successfully applied in hydrology in several studies. The prediction of water levels in catchments and rivers is of great importance for flood protection, inland navigation and domestic water supply.The comprehensive comparison of their applicability, especially for predicting water levels, has been little explored, as has the comparison of data sets with a large amount of information. In this study, Artificial Neural Network, Recurrent Neural Networks, Random Forest and Support Vector Regression were selected for water level prediction. Then, a hybrid algorithm was proposed for water level prediction. A case study for the Canal São Gonçalo was used for illustration. The attributes were compared in terms of water level, in total there are 14369 records. According to the results obtained, Random Forest has a better prediction performance and was used as a basis for the proposed algorithm. The results show that machine learning techniques can be used to support decision making in hydrological domains.*

Resumo. *Modelos de inteligência artificial tem sido aplicados com sucesso em hidrologia em diversos estudos. A previsão do nível da água nas bacias e rios é de importância significativa para as estratégias de prevenção de inundações, navegação interior e abastecimento doméstico de água. A comparação abrangente de sua aplicabilidade, em particular para previsão de nível de água, tem sido pouco explorada, bem como em conjuntos de dados com excesso de informações. No presente estudo, Artificial Neural Network (ANN), Recurrent Neural Networks (RNN), Random Forest (RF) e Support Vector Regression (SVR) foram selecionadas para a previsão do nível de água. Posteriormente foi proposto um algoritmo híbrido para a previsão do nível de água. Como caso de estudo foi utilizado o Canal São Gonçalo. Os atributos foram comparados em relação ao nível da água, no total são 14369 registros. Conforme os resultados gerados, o modelo Random Forest teve um melhor desempenho preditivo, utilizado como base para o algoritmo proposto. Os resultados mostram que o este modelo supera as técnicas convencionais de aprendizagem de máquina.*

1. Introdução

O uso dos recursos hídricos de rios por meio de reservatórios e barragens é de fundamental importância para a geração de energia, abastecimento de água, navegação e con-

trole de inundações. Mais da metade dos principais sistemas fluviais do mundo possuem reservatórios represados, que controlam ou afetam o fluxo de um determinado rio [Joo and Kim 2015]. A gestão de reservatórios de água em rios e lagoas é, portanto, um problema crítico. Existem diversas tarefas em que a previsão do nível de água represada no reservatório é um fator preocupante, tais como: para avaliar problemas estruturais em barragens, abastecimento de água e disponibilidade de recursos, qualidade da água, bioconservação da diversidade, gestão da navegação, prevenção de desastres e otimização da produção de energia hidrelétrica.

A previsão do nível da água desempenha um papel importante no bem-estar e na subsistência econômica de uma comunidade. Mudanças no nível de água podem impulsionar a ocorrência de processos físicos em lagos, resultando em mudanças na mistura de água, portanto, podem afetar ainda mais a qualidade da água e dos ecossistemas aquáticos. A previsão do nível de água tem atraído cada vez mais a atenção de pesquisadores.

Normalmente dados hidrológicos são em sua essência uma série temporal, esta pode ser conceituada como: um conjunto de observações quantitativas ordenadas cronologicamente[Kirchgässner et al. 2012].A natureza não linear e não estacionária das séries temporais pode resultar em incertezas em certas aplicações. Todavia, alguns pesquisadores tentam combinar diferentes tipos de modelos para melhorar o desempenho preditivo. A combinação de diferentes modelos é chamado de Ensembles. Um exemplo típico da aplicação é o algoritmo Random Forest.[Joo and Kim 2015]

Os dados coletados neste trabalho são referentes a barragem Eclusa de São Gonçalo, conforme a Figura 1, que representa uma estrutura hidráulica disposta numa seção do canal São Gonçalo, construída entre 1974 e 1977, com o intuito de impedir a intrusão salina advinda da Laguna dos Patos. Trata-se de uma estrutura indispensável para o desenvolvimento regional, garantindo assim atividades consolidadas na região, tais como capitação da água doce para consumo humano e uso desse recurso para a irrigação, importante para o cenário econômico regional. Associada a esta estrutura, esta uma eclusa, que através de sua operação permite a navegação neste corpo hídrico e é atualmente gerida pela Agência da Lagoa Mirim da Universidade Federal de Pelotas (ALM/UFPel), e estão completando 45 anos de serviços prestados ao desenvolvimento regional das comunidades Brasil-Uruguai.

Figura 1. Barragem Eclusa do São Gonçalo

A série de dados disponível no conjunto de dados é do tipo diária, onde os registros iniciam-se em 1979. Logo observa-se um excesso de informações, todavia, ana-

lisando estes dados, somente os registros de nível de água estavam íntegros, ou seja, com poucos dados nulos ou faltantes, ou demais presentes como: evaporação, insolação, precipitação, temperatura máxima, média e mínima, umidade do ar e velocidade do vento estavam comprometidos, logo foram substituídos pelas informações presentes na página do Inmet(Instituto Nacional de Meteorologia).

Este trabalho pretende: apresentar alguns modelos de aprendizagem de Máquina com o modelo Auto-Regressivo ARIMA para prever o nível de água para a barragem de São Gonçalo. Tendo em vista as particularidades de cada modelo e o desempenho preditivo, foi desenvolvido uma metodologia preditiva, para prever o Nível de Água futuro tendo em vista o estudo de caso proposto.

Analisando o estado da arte do tema em questão, pouco ou nada foi abordado em conjunto de dados com grande volume de informações e excesso de dados faltantes. Outro fator a ser considerado, é a utilização não somente do nível de água como variável independente, mas sim parâmetros ambientais que podem contribuir para a previsão do nível de água. Em outras palavras, o modelo proposto neste trabalho é um avanço na forma de se predizer o nível de água e pode beneficiar a população que depende de sua predição. A implementação e os dados usados em nossos experimentos estão disponíveis em.[1].

2. Trabalhos Relacionados

Nesta seção serão abordados os trabalhos relacionados sobre a previsão do nível de água. Para a busca de estudos correlatos foram utilizados os principais motores de busca de estudos científicos, sendo eles *IEEE Explore*, *Science Direct*, *ACM* e *Google Scholar*, nos quais foram mapeados estudos completos publicados em revistas ou anais de evento. Visto que alguns motores resultaram em um número muito grande de artigos, foram analisados somente os primeiros 50 de cada motor, levando-se em consideração os seguintes critérios de exclusão: artigos incompletos ou sem resultados, revisões de literatura e inferiores à 2017. Logo, os modelos selecionados nesta trabalho, consideram o estado da arte presentes nesses artigos.

O modelo ARIMA proposto por [Box et al.] é um dos modelos estatísticos lineares mais conhecidos e eficazes para previsão de séries temporais. Em [Birylo et al. 2018], os autores utilizaram o modelo ARIMA para prever o nível do lençol freático que requer três parâmetros: precipitação, vazão superficial e evapotranspiração. Os resultados de previsão apontam que para doze meses este modelo obteve bons resultados, como parâmetros de configuração utilizaram-se os seguintes valores: p=2,d=0 e q=2. Da mesma forma, [Ghimire 2017] desenvolveu um modelo baseado no Arima para prever as vazões de rios nos EUA com um sucesso significativo em um conjunto de dados de seis anos com os seguintes parâmetros de configuraçao: estação1: p=1,d=1 e q=2 e estação2: (2,1,1).

Com o passar dos anos, métodos estatísticos de aprendizado de máquina têm contribuído para o avanço dos sistemas de previsão que fornecem soluções com bom desempenho utilizando séries históricas de dados de nível de água.[Wang and Wang 2020] comparou o desempenho de vários modelos estatísticos de aprendizado de máquina como: MLR*(Regression Linear Multiple)*, GP*(Process Gaussian)*, *Random Forest*, *M5P*, para a

[1]https://github.com/prbdl/code.git.

previsão do nível de água do Rio Erie. Foram utilizados como informações de entrada dados de nível de água e variáveis ambientais como: temperatura do ar, radiação solar, velocidade do vento e humidade relativa. Neste caso, o MLR e M5P obtiveram melhores resultados, ou seja, tendo um erro absoluto de 0.02 e 0.01 mm respectivamente.

Nos últimos anos, Redes Neurais Artificias foram utilizadas para a previsão do nível de água. As redes neurais podem capturar e representar a relação não linear entre a entrada e a saída de dados. Na pesquisa de [Hrnjica and Bonacci 2019], foi aplicado dois modelos de Redes Neurais para a previsão do nível de água no lago Vrana na Croácia sendo que o primeiro modelo aplicado foi o *Feed and Forward* e o segundo o RNN. Ambos obtiveram bons resultados para a previsão em uma série temporal de 38 anos. A previsão foi de 6 a 12 meses em diante, onde a RNN obteve uma pequena vantagem na predição, principalmente no intervalo de seis meses.

Uma RNN, segundo [Xu et al. 2019a] introduz o conceito de séries temporais na estrutura de uma rede que possui uma considerável adaptabilidade no processamento de um conjunto hidrológico. Em [Zhang et al. 2020], foi proposto um modelo baseado em RNN, para a previsão do nível de água em um rio na China, onde a quantidade de dados para treinamento foi de 70% e de teste de 30%. Para a previsão do próximo mês de nível de água foi utilizado os cinco meses anteriores e os dados faltantes foram preenchidos pela média de todos os dados. Como resultado, o modelo proposto apresenta melhor desempenho preditivo em relação a ANN e LSTM, em outras palavras, possuindo um erro absoluto de 1.19 e os demais de 1.29 e 1.37 respectivamente.

Random Forest, é uma técnica de aprendizado de máquina usada para resolver problemas de regressão e classificação. Ela Utiliza uma abordagem de aprendizagem por conjunto, mas detalhes podem ser encontrados em [James et al. 2013].

Em muitos dos casos a utilização de um único modelo preditivo não é o suficiente para a obtenção de bons resultados, tendo em vista as características lineares e não lineares de uma série temporal hidrológica. Logo, alguns modelos híbridos foram propostos, como em [Xu et al. 2019b], que utilizou um modelo híbrido combinando ARIMA e RNN para a previsão do nível de água. Os resultados experimentais indicaram que o modelo combinando pode capturar melhor a tendência geral e a flutuação de amplitude, em outras palavras, o modelo ARIMA obteve um erro absoluto de 0.047, RNN de 0.033 e o híbrido de 0.021.

Segundo [James et al. 2013] o modelo SVR possui algumas vantagens, sendo elas: não é influenciado por ruído nos dados, aprende conceitos não presentes nos dados originais, sendo utilizado para problemas de classificação e regressão e com considerável desempenho. No trabalho de [Xie and Lou 2019] foi proposto uma técnica híbrida onde se observa a conjunção do ARIMA com SVR e a utilização da transformada de *Wavelet* para a decomposição dos aspectos lineares e não lineares da série para a previsão do nível de água. Este modelo, diferente dos demais, faz a decomposição desses aspectos antes do treinamento dos dados. Onde o Modelo ARIMA apresentou um erro absoluto de 0.0241, SVR de 0.0088 e o modelo ARIMA-SVR de 0.0050.

No trabalho de [Nguyen et al. 2020], é proposto a utilização da combinação do modelo ARIMA com vários modelos de Machine Learning como: Random Forest, KNN, SVR,RNN para a previsão do nível de água em um conjunto de dados de nove anos. Os

resultados mostram um melhor desempenho para os modelos combinados com: ARIMA-RF e ARIMA-KNN. Apesar deste artigo apresentar bons modelos preditivos, apenas consideram o nível de água para a previsão, o que pode ser uma limitação. Detalhes conceituais sobre o modelo RF podem ser encontrados em [James et al. 2013]

Conforme as informações obtidas através do processo de revisão do estado da arte sobre a previsão de séries temporais hidrológicas, observa-se que este tipo de conjunto de dados apresenta características lineares e não lineares. Alguns modelos foram testados, como descrito acima, em conjunto de dados não tão extensos considerando somente o nível de água como atributo previsor. Todavia, pouco ou nada foi abordado em conjunto de dados com grande volume de dados e excesso de dados faltantes.

3. Estudo Preliminar

Visando entender o problema relacionado a predição do nível de água em barragens, um estudo preliminar foi realizado, a fim de comparar diferentes técnicas e conjuntos de dados considerando a eficiência na predição. As seções seguintes descrevem as métricas utilizadas para determinar o erro instrinssinco associado aos diferentes modelos preditivos, bem como o resultado do experimento realizado.

Resumindo as contextualizações acima temos a Tabela 1, o qual verificam-se alguns trabalhos publicados e o modelo ARIMA aparece na maioria deles, principalmente quando a sua utilização é destinada em dados lineares e com margem de previsão não tão grande, como em [Ghimire 2017] e [Yu et al. 2017].

Referência	Modelo	Entrada	Conjunto	Híbrido
Birylo,2018	Arima	Nível de Água	1979 – 2015	Não
Ghimire,2017	Arima	Nível de Água	2000 – 2006	Não
Sun, 2017	Neural Network	Nível de Água e Dados Climáticos	2007 – 2012	Não
Yu, 2017	Arima	Nível de Água	2012 – 2015	Não
Zhang, 2020	RNN	Nível de Água	91 meses	Não
Hrnjica, 2019	RNN, FFN	Nível de Água	38 anos	Não
Wang, 2020	RF,KNN,MLR,GP,M5P	Nível de Água e Dados Climáticos	2002 – 2014	Não
Xu,2019	Arima/RNN	Nível de Água e Dados Climáticos	2009 – 2018	Sim
Xie, 2019	Arima/SVR	Nível de Água	2010 – 2015	Sim
Nguyen, 2020	RF,KNN,RNN e SVR	Nível de Água	2008 – 2015	Sim

Tabela 1. Principais Modelos

Outros modelos analisados como: RF,SVR e RNN e com um considerável desempenho é relatado em [Wang and Wang 2020],[Zhang et al. 2020] e [Hrnjica and Bonacci 2019].

Mesmo com relativo sucesso na utilização destes modelos para a previsão do Nível de água, alguns fatores ambientais como velocidade do vento, insolação, entre outros possuem características não lineares e a utilização de modelos híbridos tornou-se necessário como visto nos trabalhos de [Xu et al. 2019a], [Xie and Lou 2019] e [Nguyen et al. 2020].

3.1. Métricas e Indicadores de Erros

O *Mean Absolute Error* e o *Mean Square Error* são medidas básicas de análise de desempenho em problemas de regressão. O MAE é calculado subtraindo-se todos os valores em relação à média no final somando esses resultados. A palavra absoluto se refere ao módulo que desconsidera o sinal negativo. O MSE é calculado pegando o resultado do MAE e elevando ao quadrado. O MSE, penaliza os valores que estão mais distantes da média.

Elevar ao quadrado as diferenças elimina valores negativos para as diferenças e garante que o erro quadrático médio seja sempre maior ou igual a zero. É quase sempre um valor positivo. Apenas um modelo perfeito sem erro produz um MSE de zero.E isso não ocorre na prática.

$$MSE = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{1}$$

$$MAE = \sum_{i=1}^{D} |x_i - y_i| \tag{2}$$

Observando-se as fórmulas acima, temos: x representando o valor real e y o valor predito, a distância entre esses valores elevado ao quadrado é o MSE e o módulo o MAE.

Outra medida a ser considerada em problemas de regressão é o desvio padrão que mede o grau de dispersão de um conjunto de dados, sendo x valor em uma posição no conjunto de dados, M media aritmética dos dados e n a quantidade de dados.

$$\sqrt{\sum \frac{(x_i - \bar{M})^2}{n}} \tag{3}$$

Estas métricas de desempenho descritas nesta seção serão utilizadas para avaliar os modelos estudados.

3.2. Estudo sobre Modelos Aplicados à Predição do Nível de Água

Conforme as informações obtidas através do processo de revisão do estado da arte sobre a previsão de séries temporais hidrológicas, observa-se que este tipo de conjunto de dados apresenta características lineares e não lineares. Alguns modelos foram testados, como descrito nas seções anteriores, em conjunto de dados não tão extensos. Todavia, pouco ou nada foi abordado em conjunto de dados com grande volume de dados e excesso de dados faltantes, por exemplo, insolação possui *2700* registros sem presença de informação e temperatura máxima *3296* registros. Outro fator a ser considerado, é a utilização não somente do nível de água como variável independente, mas sim parâmetros ambientais que podem contribuir para a previsão do nível de água.

Este estudo é norteado sob a hipótese de que um grande conjunto de dados de mais de 40 anos com uma considerável quantidade de dados faltantes possa ser treinado e apresentar resultados relativamente bons para a previsão do nível de água.

Para a implementação dos experimentos foram utilizados o ambiente de desenvolvimento integrado *Spyder*[2] com a linguagem *Phyton*. A biblioteca de análise de dados *Pandas*[3], também foi utilizada. Para a padronização dos dados antes do treinamento foi utilizada a funcionalidade *MinMaxScaler* [4], que transforma os dados num intervalo entre *0* e *1*.

Para que resultados preditivos tenham uma performace aceitável, alguns ajustes nos algoritmos devem ser realizados. Inicialmente os dados foram divididos em treinamento e teste, sendo o melhor ajuste foi o de *4369* dados para teste e *10.000* para o treinamento, ou seja, 70% da base de dados para o treinamento e 30% para o teste.

A linguagem utilizada para a implementação dos modelos foi *Phyton*, com a biblioteca de análise de dados Pandas.

No modelo SVR, utilizou-se o *Kernel Radial Basic Function* (RBF) [5], sendo um kernel uma função que transforma um problema não linear original em um problema linear dentro do espaço de uma dimensão superior. Foi testado também o kernel polinomial [6], todavia não apresentou resultados satisfatórios.

Outra questão sobre o SVR é o parâmetro γ, pois o seu aumento faz com que os pontos mais distantes da região de separação entre classes sejam considerados, tornando as fronteiras de decisão mais restritas e complexas, gerando assim *overfitting*. Entretanto, valores menores de γ apenas consideram os pontos próximos para o cálculo do hiperplano, permitindo alguns erros de classificação ou regressão e podendo levar ao *underfitting*. Tendo em vista esta situação, o valor mais adequando para o γ foi com o valor de *3*.

O parâmetro C é responsável por controlar o quão tolerante a erros será o modelo treinado. O valor que apresentou melhores resultados para C foi com o valor *1*, em regra geral, aumentando este valor faz com que o algoritmo treine o modelo almejando a separação completa entre classes (mesmo em problemas de maior complexidade), podendo causar *overfitting* e demandar muito tempo de treinamento por gerar fronteiras de decisão muito complexas. Por outro lado, baixos valores de C flexibilizam a etapa de treinamento e permitem fronteiras de decisão com erros, mas pode levar a um *underfitting*.

No modelo *Random Forest* foram utilizadas *200* árvores com profundidade máxima *6*. No parâmetro profundidade, na medida que se aumenta seu valor, melhor o seu resultado no seu treinamento e pior na base de testes. O número mínimo de divisões foi configurado com o valor *3*, que consiste no número mínimo de amostras que um nó interno deve conter para se dividir em outros nós, para isso ele utiliza é métrica da informação com menor ganho.

Na RNN, para célula de entrada foi utilizado o número de neurônios com o valor de *120*, o qual apresentou melhor desempenho. Este valor deve ser um número razoavelmente grande para poder capturar a total dimensão e a tendência no decorrer do período. Também foi utilizado um *dropout*, de 30%, para prevenir o *overfiting* dos dados, eliminando aleatoriamente alguns nerônios da rede. Foram adicionadas mais quatro camadas

[2]https://www.spyder-ide.org/

[3]https://pandas.pydata.org/

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

[5]https://scikit-learn.org/stable/modules/generated/sklearn.gaussianprocess.kernels.RBF.html

[6]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

internas com o mesmo *dropout*. A saída da RNN consistiu em uma camada de saída utilizando como elemento de ativação à função linear e o otimizador adam, este otimizador, é eficiente ao trabalhar com grandes problemas envolvendo muitos dados ou parâmetros. Esse modelo foi treinado por um período de *50* épocas, sendo que mais épocas não refletiram em mudanças preditivas significativas.

Para a ANN, o treinamento foi realizado com as mesmas *50* épocas e os parâmetros de configuração foram: *60* neurônios, *droupout* de 30%, como elemento de ativação foi utilizado a *tangente hyperbolica* e o otimizador utilizado foi o Adam. Em termos de comparação com a DNN, a RNN obteve um tempo de treinamento de 50 minutos, enquanto o DNN foi de 5 minutos.

SVR	Mse: 0.4012 Mae: 0.2968 Dp: 0.1599
RF	Mse: 0.3990 Mae: 0.2873 Dp: 0.1643
ANN	Mse: 0.4105 Mae: 0.2865 Dp: 0.1315
RNN	Mse: 0.4011 Mae: 0.2173 Dp: 0.5441

Tabela 2. Resultado dos Experimentos

Em uma análise introdutória da Tabela 2, verifica-se que as medidas de erros MSE e MAE são denotadas de valores reais de erros expressos em metros. Outra métrica é o desvio padrão que neste caso mostra a relação da variável independente e a dependente.

Conforme a Tabela 2, pode-se concluir que o modelo RF obteve melhores resultados, este apresenta também menor variância preditiva na base de testes. Os modelos, SVR e RNN obtiveram resultados similares. Um fator a ser considerado é o tempo de execução dos modelos que na RNN é bem superior aos demais.

4. Uma Proposta de Modelo Híbrido para Predição do Nível de Água

Nesta seção serão descritos os métodos e procedimentos propostos neste trabalho, onde a arquitetura proposta foi baseada no estudo dos artigos descritos nas seções anteriores, tomando como base o problema a ser resolvido e o estudo de caso em questão.

4.1. Visão Geral

Tendo em vista os resultados do estudo preliminar, apesar de serem animadores, carecem de uma melhor acurácia. Para tanto, é necessário a combinação de modelos, criando uma estrutura híbrida.

O modelo ARIMA é um dos modelos mais utilizados para a predição de séries temporais. Todavia, o mesmo obtém um bom desempenho em predições com um intervalo pequeno e com dados lineares. Logo, com dados climáticos, que apresentam um comportamento não linear, a sua acurácia cai consideravelmente.

Em [Xu et al. 2019a] verifica-se um modelo híbrido utilizando o modelo ARIMA com RNN. Esta arquitetura apresenta resultados interessantes, mas se aplica a um *dataset* pequeno e utiliza uma célula RNN básica.

No trabalho de [Xie and Lou 2019] foi proposto um modelo com ARIMA e SVR, que utiliza somente dados de nível de água em uma série pequena, sabe-se que dados ambientais influenciam o nível de água. Em [Nguyen et al. 2020], é testado vários modelo híbridos com o ARIMA, onde como o anterior utiliza somente dados de nível de água para a previsão em um conjunto de dados reduzido.

A arquitetura proposta neste trabalho é mostrada na Figura 2, onde existe inicialmente uma base de dados hidrológicos. Nessa base de dados o único conjunto de dados relativamente íntegro é o de nível de água, isto é, existem poucos dados faltantes. As variáveis climáticas foram descartadas, em virtude de estarem com um excesso de informações nulas. Assim sendo, foram utilizados os dados do Inmet (Instituto Nacional Meteorológico) compreendendo o mesmo período e localização. Tendo este novo conjunto de dados, o mesmo foi dividido em dois outros *datasets* o primeiro somente com o nível de água e o segundo com os dados climáticos que são: evaporação, insolação, precipitação, temperatura máxima, média e mínima, umidade do ar e velocidade do vento.

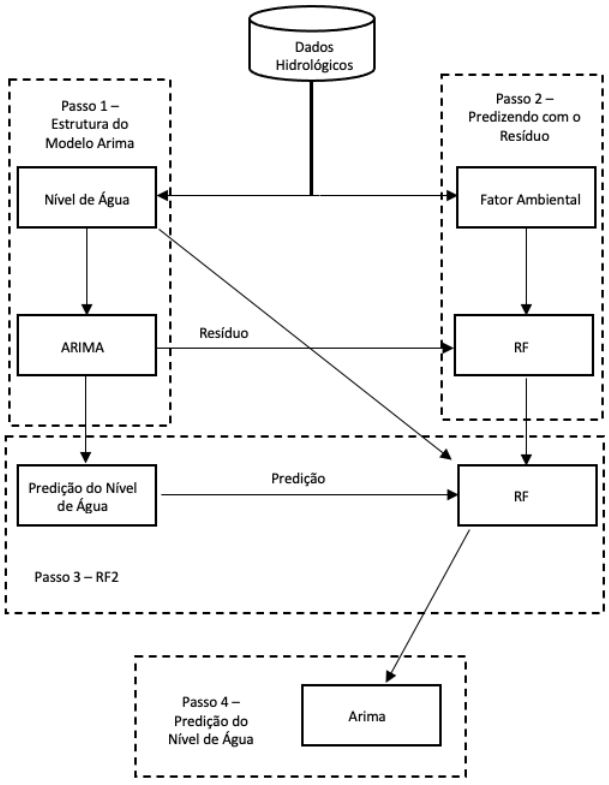

Figura 2. Arquitetura Proposta

Continuando a análise da Figura 2, no passo 1, observa-se que os dados de nível, que possuem características lineares, são submetidos ao modelo ARIMA. No segundo passo, os resíduos do modelo ARIMA, ou seja, a parte não linear, servem como entrada para o modelo *random forest* com os dados ambientais. No passo 3 temos a segunda camada com o RF, onde tem como entrada a predição do nível anterior com a predição do ARIMA, considerando os dados originais. Por fim, no passo 4, a predição do modelo anterior é utilizado como entrada para a última camada, que utiliza o modelo ARIMA.

Nesta última camada, os dados são divididos em 70% para o treinamento e 30% para o teste. Esta divisão se justifica pelo fato de apresentar um melhor desempenho preditivo nos testes executados.

No modelo proposto, os melhores valores encontrados para p,d e q foram *(2,1,2)*, ou seja, é utilizado um componente autorregressivo p de valor 2, as médias móveis q como o valor de 2 o qual tenta explicar o efeito do ruído na série. A diferenciação d, de ordem 1 torna a série estacionária em uma mesma média.

4.2. Algorítimo e Análise

Nesta seção será abordado o algoritmo proposto intitulado de ARIMA-RF.

Conforme a Figura 3, observa-se inicialmente a variável act, que recebe a autocorrelação total, servindo de base para a escolha do número de lags da série hidrológica, ou seja, quantos períodos para trás será utilizado para a previsão. Em d é aplicado a função das diferenças para deixar a base estacionária na média 0. É aplicado também a função AIC [7] para obter os melhores "p"e "q".

Use of the ARIMA-RF model to predict the water level

Input: base_h is hydrological base, base_e is environmental base
Output: predicted water level

```
//Model ARIMA
act = ACF(base_h);
d = diff(base_h);
for p=0 to pmax do
 for q=0 to qmax do
   m = ARIMA(act,(p,d,q);
   AIC = aic(m);
 end for
end for
m=ARIMA(act,(p,d,q)
pdc=m.predict();
res = residual();
//Model RF1
baserf1 = res + base_e
x = normalize(baserf1[:,0:9]);
y = normalize(baserf1[:,9]);
rf1 = RF(x,y);
//Model RF2
baserf2 = pdc + rf1 + base_h;
x = normalize(baserf2[:,0:9]);
y = normalize(baserf2[:,9]);
rf2 = RF(x,y);
//Model ARIMA2
basearima = rf2;
train_size = (basearima * 2/3)
train=basearima[:,train_size]
test=basearima[train:]
out = ARIMA(test)
```

Reverse normalization and predict water level

Figura 3. Algoritmo Proposto: ARIMA-RF

As variáveis: pdc e res armazenam a predição do modelo ARIMA e os resíduos. No modelo RF1, a variável $baserf1$ armazena os resíduos do modelo anterior com os dados de satélite. x contém os dados normalizados com as informações preditoras e y possui o dado dependente, também normalizado.

[7] https://coolstatsblog.com/2013/08/14/using-aic-to-test-arima-models-2/

O modelo RF2 possui como entradas: pdc, $rnn1$ e $base_h$, ou seja, tem como parte que compõe a variável $baserf2$, a predição do ARIMA, as predições do modelo anterior e a base de nível original. Esta etapa realiza o processo inverso da normalização, para que os dados de nível sejam expressos em $metros$ e, assim, se obtém as predições. O resultado das predições são armazenados em $rf2$.

Na última camada aplica-se o modelo ARIMA. A entrada é o nível de água do modelo anterior $rf2$ é armazenado em $basearima$. O conjunto é dividido em treino e teste, onde os 70% primeiros registros são utilizados para treinamento do modelo e 30% por cento para teste. A variável out armazena a predição final do modelo.

Como se observa na Tabela 4, o ARIMA-RF apresentou menor erro preditivo em relação ao ARIMA-RNN proposto por [Xu et al. 2019a]. Foi realizada a implementação deste modelo e os resultados sinalizam uma melhor acurácia no modelo proposto neste trabalho.

ARIMA-RF	Mse: 0.030 Mae: 0.020 Dp: 0.995
ARIMA-RNN	Mse: 0.084 Mae: 0.058 Dp: 0.9650

Figura 4. Comparação do ARIMA-RF com o ARIMA-RNN

O modelo ARIMA-RF apresenta um erro menor, atingindo uma variação de *0.054*m comparado ao ARIMA-RNN. No que se refere ao erro médio absoluto, a diferença foi de *0,038*m. Esses dados, são gerados na base de testes, onde o modelo treinado não conhece as informações. O experimento deste trabalho comprova a vantagem do modelo combinado na previsão do nível de água, mostra sua racionalidade científica e apresenta melhor desempenho preditivo. Portanto, no sistema hidrológico, este modelo pode alcançar um bom efeito na previsão do nível da água.

5. Conclusão

A previsão do nível de água em séries temporais é um problema de pesquisa bastante investigado atualmente. Com a utilização de técnicas de *machine learning* foi possível um avanço substancial na precisão deste tipo de previsão.

A proposta deste trabalho tem como base explorar técnicas preditoras, considerando uma série temporal extensa. Com isso surge a hipótese de que um modelo treinado e com algumas alterações possa prever com certa precisão o nível de água em barragens ou lagos, fornecendo assim benefícios para quem depende dessa previsão, ou seja, prevendo futuras estiagens ou grandes enchentes, por exemplo.

Para tanto, foi escolhido o modelo *Random Forest* tendo em vista o bom desempenho nesta série temporal em questão. Então, o modelo híbrido proposto utiliza uma primeira camada com o ARIMA, tendo como entrada somente o nível de água, posteriormente uma segunda camada com os resíduos do modelo anterior acrescidos com os dados ambientais, o resultado da predição do ARIMA com o resultado da predição da segunda camada é submetido novamente ao modelo *Random Forest*. Por fim, o resultado desta camada é submetido ao modelo *ARIMA* prevendo assim o nível de água, utilizando 70% da base para treinamento e o restante para testes.

Sugere-se como trabalhos futuros, testar este modelo com uma maior quantidade de dados, estes dispostos em partes distintas da lagoa, verificando assim o quanto cada sensor contribui para o nível de água na região.

Referências

Birylo, M., Rzepecka, Z., Kuczynska-Siehien, J., and Nastula, J. (2018). Analysis of water budget prediction accuracy using arima models. *Water Science and Technology: Water Supply*, 18(3):819–830.

Box, G. E., Jenkins, G. M., and Reinsel, G. C. Time series analysis.

Ghimire, B. N. (2017). Application of arima model for river discharges analysis. *Journal of Nepal Physical Society*, 4(1):27–32.

Hrnjica, B. and Bonacci, O. (2019). Lake level prediction using feed forward and recurrent neural networks. *Water Resources Management*, 33(7):2471–2484.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Joo, T. W. and Kim, S. B. (2015). Time series forecasting based on wavelet filtering. *Expert Systems with Applications*, 42(8):3868–3874.

Kirchgässner, G., Wolters, J., and Hassler, U. (2012). *Introduction to modern time series analysis*. Springer Science & Business Media.

Nguyen, X. H. et al. (2020). Combining statistical machine learning models with arima for water level forecasting: The case of the red river. *Advances in Water Resources*, 142:103656.

Wang, Q. and Wang, S. (2020). Machine learning-based water level prediction in lake erie. *Water*, 12(10):2654.

Xie, Y. and Lou, Y. (2019). Hydrological time series prediction by arima-svr combined model based on wavelet transform. In *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, pages 243–247.

Xu, G., Cheng, Y., Liu, F., Ping, P., and Sun, J. (2019a). A water level prediction model based on arima-rnn. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 221–226. IEEE.

Xu, G., Cheng, Y., Liu, F., Ping, P., and Sun, J. (2019b). A water level prediction model based on arima-rnn. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 221–226.

Yu, Z., Lei, G., Jiang, Z., and Liu, F. (2017). Arima modelling and forecasting of water level in the middle reach of the yangtze river. In *2017 4th International Conference on Transportation Information and Safety (ICTIS)*, pages 172–177. IEEE.

Zhang, J., Zhang, Z., Weng, Y., Gosling, S., Yang, H., Yang, C., Li, W., and Ma, Q. (2020). Using recurrent neural network for intelligent prediction of water level in reservoirs. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1125–1126. IEEE.