Daily streamflow forecasting for Paraíba do Sul river using machine learning methods with hydrologic inputs

Yulia Gorodetskaya[1*], Leonardo Goliatt da Fonseca[2], Gisele Goulart Tavares[1], Celso Bandeira de Melo Ribeiro[3]

[1]Programa de Pós-Graduação em Modelagem Computacional
Universidade Federal de Juiz de Fora - Juiz de Fora, MG - Brasil

[2]Departamento de Mecânica Aplicada e Computacional, Faculdade de Engenharia
Universidade Federal de Juiz de Fora - Juiz de Fora, MG - Brasil

[3]Departamento de Engenharia Sanitária e Ambiental, Faculdade de Engenharia
Universidade Federal de Juiz de Fora - Juiz de Fora, MG - Brasil

[*]yu.gorodetskaya@gmail.com

Abstract. *The Paraíba do Sul river flows through the most important industrial region of Brazil and its basin is characterized by conflicts of multiple uses of its water resources. The prediction of its natural flow has strategic value for water management in this basin. This research investigates the applicability of the two machine learning methods (Random Forest and Artificial Neural Networks) for daily streamflow forecasting of the Paraíba do Sul River at lead times of 1-7 days. The impact of fluviometric and pluviometric data from other basin sites on the quality of the forecast is also evaluated.*

Resumo. *O rio Paraíba do Sul flui através da mais importante região industrial do Brasil e sua bacia é caracterizada por conflitos de usos múltiplos de seus recursos hídricos. A previsão de sua vazão natural possui valor estratégico para a gestão da água nesta bacia. Essa pesquisa investiga a aplicabilidade dos métodos Random Forest e Redes Neurais Artificiais na modelagem de vazão média diária de uma estação fluviométrica do rio Paraíba do Sul com um horizonte de previsão de até 7 dias. Avalia-se o impacto dos dados fluviométricos e pluviométricos de outros locais da bacia na qualidade da previsão.*

1. Introdução

O rio Paraíba do Sul flui através da mais importante região industrial do Brasil, entre as cidades do Rio de Janeiro e de São Paulo e abastece a cidade do Rio de Janeiro e a região do Grande Rio. De acordo com a Agência Nacional de Águas (ANA), a bacia do rio Paraíba do Sul é caracterizada por conflitos de usos múltiplos de recursos hídricos (abastecimento urbano, diluição de esgotos, irrigação e geração de energia hidrelétrica) e pelo desvio de suas águas para o rio Guandú, responsável pelo abastecimento de cerca de 9 milhões de pessoas na região metropolitana do Rio de Janeiro. Em função da sua importância, a aplicação de ferramentas de auxílio à previsão de vazão natural do rio pode assumir valor estratégico para a gestão da quantidade e da qualidade da água nesta bacia.

Um caminho para a modelagem da previsão de vazão é o uso do histórico de registros de vazão, retirando da concepção de modelagem uma descrição completa dos

princípios físicos e exigindo menos requisitos de dados que no processo de modelagem dos processos físicos [Shafaei and Kisi 2016]. Desta forma, a utilização de métodos de estimativas de vazões a partir de dados históricos provenientes de séries temporais parece atraente embora a obtenção de dados confiáveis e de longa duração não seja uma tarefa fácil.

Diversas técnicas de modelagem vêm sendo empregadas para estimar vazões. Porém, as restrições baseadas em simplificações dos fenômenos hidráulicos e hidrológicos naturais, que se relacionam de forma não-linear, tem sido o principal problema nesta modelagem [da Silva et al. 2011]. As técnicas de modelagem podem ser classificadas em duas categorias: $theory\text{-}driven$ é a abordagem baseada em teoria (conceitual e física) e $data\text{-}driven$ é a baseada em dados (empírica e $black\ box$). Neste contexto, os modelos $data\text{-}driven$ podem ser preferíveis para descobrir relacionamentos a partir de dados de entrada e saída, mesmo quando o usuário não possui uma compreensão física completa dos processos subjacentes. Embora esses modelos sejam muito úteis para a previsão do fluxo do rio, a principal preocupação é obter previsões precisas de vazão em locais específicos do rio [Asadi et al. 2013]. As principais vantagens das técnicas do aprendizado de maquina em relação às técnicas de regressão estatística são que os modelos resultantes: são mais resistentes à multicolinearidade e valores extremos; incluem métodos que reduzem $overfitting$ do modelo; melhor identificação das variáveis importantes das relações não-lineares [Povak et al. 2013]. Dentre os modelos empíricos, modelos clássicos de séries temporais, como a Média Móvel Integrada de Regressão Automática (ARIMA), são amplamente aplicados na previsão de séries temporais hidrológicas. Entretanto, estes modelos são essencialmente lineares, consideram os dados como estacionários e possuem uma capacidade limitada de capturar não-estacionaridades e não-linearidades em dados hidrológicos[Asadi et al. 2013]. [Patel and Ramachandran 2015] concluíram que este fato pode ser a principal razão para o desempenho superior dos métodos do aprendizado de maquina em comparação das técnicas de regressão estatística. Os métodos de aprendizado de máquina têm sido aplicados com sucesso para resolver problemas não-lineares em hidrologia [Carlisle et al. 2010, Bhagwat and Maity 2012, Rasouli et al. 2012, Zhao et al. 2012, Povak et al. 2013, Wang et al. 2015, Li et al. 2016, Shafaei and Kisi 2016, Shortridge et al. 2016, Khair et al. 2017].

Considerando os esforços encontrados na literatura, este trabalho visa contribuir na comparação das técnicas do aprendizado de maquina na modelagem que se destacaram na previsão de vazão e verificação de seus desempenhos em dados do rio Paraíba do Sul. O objetivo dessa pesquisa é investigar a aplicabilidade dos métodos Random Forest (RF) e Redes Neurais Artificiais (ANN) na modelagem de vazão média diária na estação fluviométrica localizada no rio Paraíba do Sul com um horizonte de previsão de 1-7 dias. Além disso, avaliar a influência dos dados fluviométricos e pluviométricos de outros locais da bacia na qualidade da previsão.

2. Materiais e Métodos

2.1. Área de estudo e seleção de dados

Para este estudo de caso a sub-bacia Baixo Paraíba do Sul da bacia do rio Paraíba do Sul foi escolhida para análise. É caracterizada como uma bacia de maior porte por possuir muitos rios afluentes e uma vazão no rio principal elevada, sendo estas características

determinantes para uma dinâmica mais lenta desta bacia.

As previsões de vazão diária foram realizadas para a estação fluviométrica CAMPOS - PONTE MUNICIPAL (código ANA: 58974000), com área de drenagem 55.500 km^2, localizada na sub-bacia Baixo Paraíba do Sul. Na modelagem de vazão utilizaram-se as informações históricas dos dados de vazões diárias médias (em m^3/s) e precipitações diárias totais (em mm) observadas no próprio local e nas estações pluviométricas e fluviométricas à montante, provenientes pela ANA, referentes ao período de 1 de janeiro de 2000 a 31 de dezembro de 2016. Os dados foram coletados nas estações pluviométricas de códigos: 2141003, 2141005, 2141006, 2141007, 2142002, 2142022; e fluviométricas: 58974000, 58960000, 58880001, 58874000, 58795000, 58920000, 58857000.

Na seleção das estações buscou-se aquelas que se encontram atualmente em operação, com séries históricas sem muitas falhas e localizadas espacialmente próximas da estação analisada. Contudo, no processo de modelagem foi necessário lidar com problemas nas bases de dados. As séries temporais utilizadas possuíam inconsistência em seus valores, principalmente valores faltantes e lacunas temporais sem registro. Optou-se por não considerar esses dados pois, além da incerteza devido às medições, o modelo passaria a conter incertezas com tentativas de previsão desses dados, dos quais não se tem nenhuma informação, inserindo imprecisão na modelagem.

2.2. Modelagem de previsão da vazão

Neste estudo, os métodos de aprendizado da máquina Random Forest e redes neurais artificiais do tipo MLP foram aplicados na modelagem de vazão diária com um horizonte de previsão de 1-7 dias. As observações de vazão média diária e precipitação total diária de períodos passados das estações fluviométricas e pluviométricas à montante foram usadas como variáveis de entrada nos modelos implementados. Para investigar a aplicabilidade dos métodos na previsão de vazões e a influência dos dados climáticos de outros locais da bacia na qualidade da previsão foram criados cinco modelos de entrada com combinações de precipitações e vazões diferentes. Em todas configurações de entrada foi utilizada a mesma arquitetura de ANN e RF.

Conforme a Tabela 1, o modelo de entrada mais simples é o modelo 1, denominado "mod 1". Neste cenário, a única variável conhecida é a vazão no local onde é realizada a previsão. No modelo 2, a previsão se baseia na vazão no próprio local e nas precipitações das regiões próximas. Nos modelos 3 e 4, a vazão prevista é baseada nas vazões no próprio local e à montante do rio. Por fim, o modelo 5 contempla a vazão do próprio local e as precipitações e vazões à montante.

No caso de modelo de entrada "mod 3", o modelo de previsão recebe dados de três séries históricas provenientes das estações fluviométricas 58218000, 58204000 e 58183000, constituídos por 4678 valores de dias. Para treinamento do modelo serão escolhidos somente dias de interseção das três séries históricas. Ou seja, se existe um período faltante numa série, este período será descartado de outras séries da entrada.

Desta forma, como pode ser observado na Figura 1, para realizar a previsão para 7 dias subsequentes, o modelo de previsão recebe nas suas entradas as amostras ininterruptas de 10 valores de vazão de cada série (no total são 30 valores de três séries históricas), e retorna 7 valores correspondentes às vazões estimadas. Consideramos, portanto, como uma amostra de dados, um conjunto com informações ininterruptas (X, y), onde X =

$(10(V_{58974000}),10(V_{58960000}),10(V_{58880001}))$, composto por 30 dias antecedentes de vazões e $y = (7(V_{58974000}))$, composto por 7 dias seguintes de vazão. Desta forma, chamaremos por conjunto de dados de entrada inteiro todas as amostras $(X_i, y_i),\ i = 1, \ldots, N$, onde N é o número de todas as amostras selecionadas nas séries históricas de modelo de entrada, como mostrado na Figura 1.

Tabela 1. Estações pluviométricas e fluviométricas usadas nas modelos de entrada para a modelagem da CAMPOS - PONTE MUNICIPAL (58974000).

Modelo	Dias	Estações fluviométricas e pluviométricas que fornecem variáveis de entrada (precipitação e vazão)
mod 1	4977	58974000
mod 2	4415	58974000, 2141003, 2141005, 2141006, 2141007, 2142002, 2142022
mod 3	4678	58974000, 58960000, 58880001
mod 4	4415	58974000, 58960000, 58880001, 58874000, 58795000, 58920000, 58857000
mod 5	4415	58974000, 58960000, 2141003, 58880001, 2141005, 58874000, 2141006, 58795000, 2141007, 58920000, 2142002, 58857000, 2142022

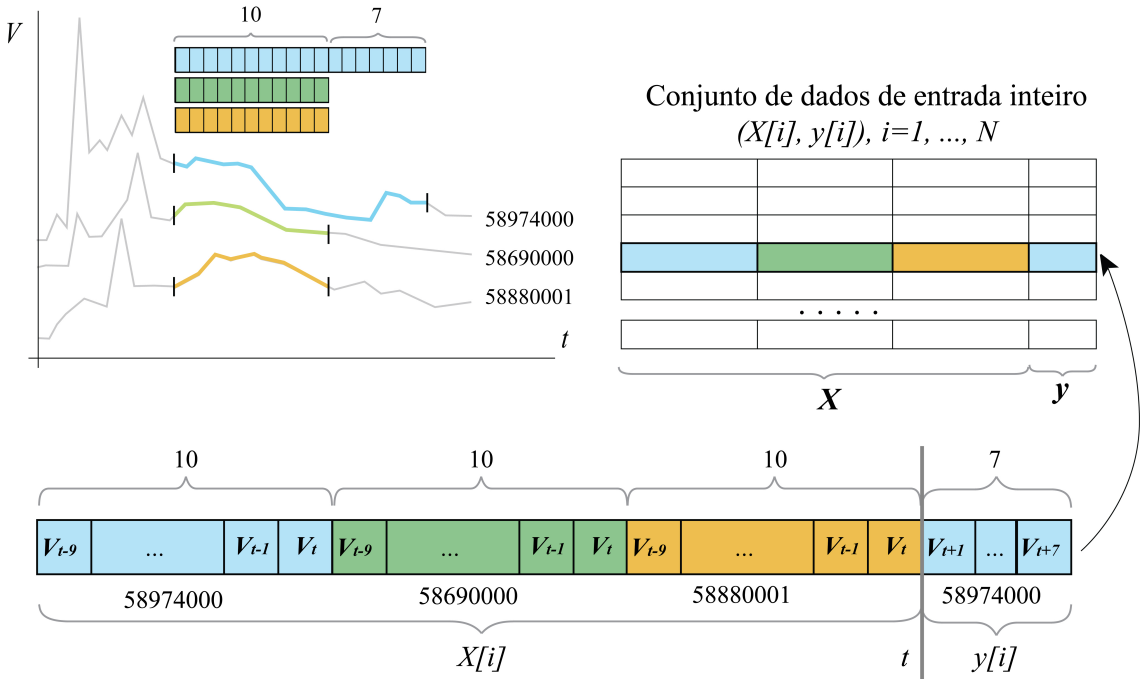

Figura 1. Amostra de dados com um conjunto de informações ininterruptas de vazão.

De um modo geral, o modelo previsor resultante tem a seguinte forma:

$$(V_{t+1}, V_{t+2}, ..., V_{t+7}) = F(V_t, V_{t-1}, ..., V_{t-9}, ..., P_t, P_{t-1}, ..., P_{t-9}),$$

onde a função F representa a relação entre as vazões e precipitações dos 10 dias anteriores e as vazões estimadas para os 7 dias seguintes. V_{t+1}, V_{t+2}, ... , V_{t+7} denotam valores

correspondentes às vazões estimadas nos dias $t+1$, $t+2$, ... , $t+7$; V_t, V_{t-1}, V_{t-2}, ... , V_{t-9} denotam diferentes defasagens temporais da vazão observada numa estação fluviométrica nos dias $t, t-1, t-2, ..., t-9$ que antecedem a previsão; P_t, P_{t-1}, P_{t-2}, ... , P_{t-9} denotam diferentes defasagens temporais de precipitação observada numa estação pluviométrica nos dias t, $t-1$, $t-2$, ... , $t-9$, respectivamente.

Nesta pesquisa, a opção por escolher uma janela de 10 dias antecedentes se deu apos uma série de testes. A aplicação de mais dias (12 e 14) de dados antecedentes ao modelo não melhorou os resultados e resultou em maior consumo de recursos computacionais. De outra forma, o uso de 7 dias de dados anteriores não mostrou resultados satisfatórios.

Cabe salientar que, em trabalhos correlatos [Shafaei and Kisi 2016, Ribeiro et al. 2014, Bhagwat and Maity 2012, Karimi et al. 2016, Rasouli et al. 2012], os autores optam por construir e treinar um modelo para cada dia subsequente. Alternativamente à maioria dos modelos encontrados na literatura, neste artigo um mesmo modelo foi usado para realizar a previsão de todos os 7 dias subsequentes. Essa característica facilita o treinamento e a interpretabilidade do modelo, porém pode resultar em perda de qualidade quando comparado com as previsões de modelos dedicados para unicamente um determinado dia subsequente.

2.3. Métodos do Aprendizado de Máquina

2.3.1. Random Forest

Random Forest, um algoritmo de aprendizado de máquina proposto por Breiman [Breiman 2001] é um método de combinação de classificação baseado na teoria de aprendizado estatístico. A RF manipula dados não-lineares e não Gaussianos, é passível de interpretação de modelos e está livre de problemas excessivos à medida que o número de árvores aumenta. Além disso, a RF fornece a importância relativa dos descritores, que podem ser ainda utilizados na seleção de variáveis [Li et al. 2016]. O modelo RF emprega como estratégia a seleção randômica de um subconjunto de m preditores para gerar uma árvore binária, onde cada árvore é gerada em uma amostra $bootstrap$ do conjunto de treinamento. Para cada árvore, os dados de resposta são agrupados em dois nós descendentes para maximizar a homogeneidade e então a melhor divisão binária é escolhida. Cada nó descendente da divisão selecionada é tratado similarmente ao nó original e o processo continua recursivamente até que um critério de parada seja alcançado. Todas as árvores são geradas até seu tamanho máximo e as predições finais são obtidas dos resultados médios [Breiman 2001]. No modelo RF, o número de árvores deve ser definido ($ntree$)[Li et al. 2016]. A partir de testes preliminares optou-se por utilizar 250 árvores neste estudo .

2.3.2. Rede Neural Artificial

De acordo com [Haykin et al. 2009], Redes Neurais Artificiais (ANN) são modelos paralelos distribuídos, compostos por unidades de processamento simples, chamadas neurônios, interligadas entre si e com o ambiente. As ANNs aprendem através de exemplos e as conexões entre os neurônios estão associadas a pesos que armazenam o conhecimento da rede. As redes possuem grande habilidade para lidar com problemas complexos,

imprecisos e com grandes mudanças na informação de entrada tais como casos de inconsistência ou ausência de dados nas séries históricas de precipitação e vazão. Isso justifica seu uso em previsão de vazão. A arquitetura mais utilizada nesta metodologia, para resolver problemas não lineares, é a de redes de múltiplas camadas - Multilayer Perceptron (MLP). Essa arquitetura é organizada em camadas, com neurônios que podem estar conectados aos neurônios da camada posterior. O objetivo do treinamento da rede consiste em reduzir continuamente o erro até um determinado valor aceitável. Cada camada tem seus pesos ajustados de modo a minimizar o erro da rede. No presente estudo, foi treinada uma ANN do tipo MLP com uma camada oculta de 50 neurônios e a função ReLU (Rectified Linear Unit) [Nair and Hinton 2010] como função de ativação para a camada intermediária, minimizando o erro médio quadrático. A rede foi treinada com a técnica de Stochastic Gradient-based Optimizer [Kinga and Adam 2015]. Foi utilizada uma taxa de aprendizado igual a 0.001 e critério de parada do treinamento de 200 iterações.

2.4. Avaliação do desempenho dos modelos

Um dos problemas relacionados aos modelos de predição é o chamado sobreajuste ($overfitting$), que se dá quando não se tem acesso completo aos dados e o modelo fica condicionado aos dados de treino, falhando, assim, na validação quando dados diferentes são utilizados. Uma alternativa para esse tipo de problema é a aplicação da técnica de validação cruzada k-$fold$. O procedimento de validação cruzada fornece um mecanismo para avaliar o quão bem um modelo irá generalizar um conjunto de dados ainda não vistos evitando alguns problemas que podem aparecer com o uso de um único modelo em um único conjunto de dados. Essa técnica consiste na separação de k subconjuntos de igual tamanho, onde destes, k-1 subconjuntos são utilizados no treinamento do modelo e um serve como base para a validação do mesmo. Este processo é repetido k vezes, empregando um conjunto de teste distinto em cada iteração.

Em estudos semelhantes [Kohavi et al. 1995, Hastie et al. 2009] o valor de k geralmente varia entre 5 e 10. Nos experimentos computacionais conduzidos neste trabalho uma validação cruzada 5-$fold$ é aplicada na avaliação da performance dos modelos considerados neste trabalho para a redução de tempo computacional. O conjunto de dados de entrada (X, y) (demonstrado na Figura 1) foi dividido em k = 5 subconjuntos a cada iteração do processo da validação cruzada. O modelo é treinado considerando 4 subconjuntos (conjunto de treinamento) e ajustado para estimar a vazão em 1 subconjunto (conjunto de teste).

A qualidade do ajuste do modelo nos dados de entrada é normalmente determinada por comparações entre o modelo previsto e as observações. As habilidades do modelo preditivo são mensuradas através de taxas quantitativas de acerto encontradas durante das simulações. Porém, para uma avaliação completa do desempenho do modelo hidrológico, deve ser incluída pelo menos uma medida de erro relativo e uma medida de erro absoluto [Legates and McCabe 1999, Chai and Draxler 2014].

Entre vários critérios que são comumente usados para avaliação de desempenho do modelo, nesta pesquisa, foram utilizadas as seguintes métricas:

- Erro percentual absoluto médio (MAPE) representa a média percentual da divisão entre erro de previsão e o valor real e descreve a diferença entre as simulações do modelo e as observações nas unidades da variável. Seu retorno é dado em

porcentagem ($\%$) e quanto mais próximo de zero melhor é o modelo. O MAPE é calculado de acordo com a seguinte equação:

$$\text{MAPE} = 100 \times \frac{1}{n}\sum_{i=1}^{n-1}\frac{|O_i - P_i|}{|O_i|}, \tag{1}$$

onde O representa a vazão observada, P a vazão predita pelo modelo e n é o tamanho da série histórica.

- O coeficiente de determinação (R^2) descreve a proporção da variância total nos dados observados que pode ser explicada pelo modelo. R^2 varia de ∞ a 1.0, com valores mais altos indicando melhor ajuste, e é dado por

$$R^2 = 1 - \frac{\sum_{i=0}^{N-1}(O_i - P_i)^2}{\sum_{i=0}^{N-1}(O_i - \overline{O_i})^2}, \tag{2}$$

onde O é a vazão observada, P é a vazão prevista, a barra indica a média para todo o período de tempo da avaliação.
Notamos que, que o coeficiente de determinação é limitado na medida em que padroniza as diferenças entre as médias e as variâncias observadas e previstas, uma vez que apenas avalia relações lineares entre as variáveis [Legates and McCabe 1999, Chai and Draxler 2014].

A preparação dos dados foi realizada utilizando-se OpenOffice Calc 4.1 e a linguagem de programação Python 3.5. Os modelos foram alimentados por séries históricas armazenados em banco de dados MySQL e utilizam algorítmos de aprendizagem disponíveis na biblioteca scikit-learn para Python [Pedregosa et al. 2011], seaborn [Waskom et al. 2017] e pandas [McKinney 2010].

3. Resultados e Discussão

Com objetivo de alcançar uma melhor generalização, e assim melhores resultados, cada um dos modelos RF e ANN teve o processo de treinamento repetido 30 vezes onde, para cada iteração, foi utilizada a técnica de validação cruzada k-fold. Os dados referentes à 13 estações fluviométricas e pluviométricas foram utilizados como dados de entrada nos modelos para a previsão de vazão em um horizonte de 1-7 dias. Vários modelos foram implementados com intuito de investigar a influência das informações pluviométricas e fluviométricas das estações à montante na qualidade da previsão.

Neste caso, diferentes modelos foram desenvolvidos para a previsão da vazão diária: 5 modelos RF e 5 modelos ANN. A Tabela 2 mostra o resultado médio do desempenho dos modelos RF e ANN para cada tipo de dados de entrada com seu respectivo desvio padrão. Deve-se notar que o mesmo conjunto de dados foi usado para treinamento e teste dos modelos utilizando validação cruzada k-fold. Observou-se o aumento gradual do MAPE e o decréscimo do coeficiente R^2 com o aumento do alcance da previsão para todos os modelos.

De acordo com os resultados obtidos (Tabela 2), para todos os horizontes de previsão e combinações de entrada, os modelos RF apresentaram um desempenho superior em comparação aos modelos ANN em termos de MAPE e R^2. No melhor modelo RF para os dias 1, 3 e 7 obteve-se coeficiente R^2 e MAPE de (0.925, 0.84 e 0.645) e (7.68%,

Tabela 2. Resultados médias do desempenho dos modelos RF e ANN para a CAMPOS PONTE-MUNICIPAL (58974000).

Dia	Entrada	MAPE		R^2	
		ANN	RF	ANN	RF
1	mod 1	11.1251(0.7625)	9.7139(0.0398)	0.8932(0.0058)	0.9(0.0024)
	mod 2	12.5988(0.3664)	9.3847(0.0434)	0.8879(0.0025)	0.8902(0.0011)
	mod 3	11.2234(0.5532)	8.9157(0.0364)	0.8977(0.0038)	0.9176(0.0013)
	mod 4	10.6633(0.5423)	7.6809(0.0336)	0.9126(0.0043)	0.9255(0.0013)
	mod 5	10.7933(0.4084)	7.8889(0.0285)	0.9176(0.0024)	0.9239(0.0011)
2	mod 1	15.6719(0.6328)	15.3686(0.058)	0.7437(0.0058)	0.7829(0.0028)
	mod 2	16.899(0.3962)	13.7959(0.0438)	0.7706(0.0042)	0.7981(0.0018)
	mod 3	15.7683(0.3743)	13.2014(0.0537)	0.7696(0.0048)	0.8278(0.0024)
	mod 4	15.3325(0.6526)	11.7856(0.0473)	0.7903(0.0059)	0.8399(0.0029)
	mod 5	15.1773(0.3415)	11.868(0.0367)	0.8067(0.0043)	0.8369(0.0017)
3	mod 1	19.056(0.7585)	19.5499(0.062)	0.6273(0.004)	0.7061(0.0029)
	mod 2	20.4309(0.5373)	17.8893(0.0645)	0.6581(0.0055)	0.7036(0.0029)
	mod 3	19.2243(0.5904)	16.3657(0.0701)	0.6653(0.0065)	0.7729(0.0036)
	mod 4	19.2448(0.8453)	15.1712(0.0708)	0.6669(0.0075)	0.7671(0.0036)
	mod 5	19.1543(0.4376)	15.5635(0.056)	0.6896(0.0067)	0.7518(0.0018)
4	mod 1	21.4561(0.6895)	22.4552(0.0686)	0.5539(0.0048)	0.6599(0.0029)
	mod 2	23.4006(0.6231)	21.5715(0.0757)	0.57(0.006)	0.6259(0.0033)
	mod 3	21.5366(0.6538)	18.6768(0.0628)	0.592(0.0066)	0.7373(0.0031)
	mod 4	21.8522(0.6776)	17.5969(0.1037)	0.58(0.0088)	0.7247(0.0038)
	mod 5	22.2832(0.6458)	18.64(0.0712)	0.6007(0.0069)	0.6898(0.0018)
5	mod 1	23.0559(0.7969)	24.4356(0.0673)	0.5007(0.0057)	0.6301(0.0044)
	mod 2	25.4636(0.5704)	24.2886(0.0835)	0.5011(0.0055)	0.5707(0.0028)
	mod 3	23.2459(0.7024)	20.3424(0.0628)	0.5344(0.0056)	0.7059(0.0036)
	mod 4	23.7982(0.6891)	19.3961(0.1085)	0.5069(0.0078)	0.6884(0.0036)
	mod 5	24.3732(0.7267)	20.871(0.0887)	0.5282(0.006)	0.6396(0.0022)
6	mod 1	24.1152(0.6975)	26.0084(0.0807)	0.4597(0.0053)	0.6065(0.0051)
	mod 2	26.6863(0.6015)	26.3677(0.0815)	0.4539(0.0066)	0.5295(0.003)
	mod 3	24.7014(0.5699)	21.6447(0.0821)	0.4808(0.0059)	0.6784(0.0043)
	mod 4	25.1158(0.6286)	20.7471(0.1351)	0.4515(0.0085)	0.6641(0.0044)
	mod 5	25.964(0.5302)	22.6055(0.1064)	0.4712(0.0066)	0.6034(0.0027)
7	mod 1	24.9873(1.0205)	27.4112(0.0807)	0.4261(0.0069)	0.5777(0.0051)
	mod 2	27.6624(0.6907)	27.8863(0.1146)	0.4197(0.0073)	0.5004(0.0039)
	mod 3	25.4248(0.6189)	22.7408(0.0792)	0.446(0.0059)	0.6611(0.0056)
	mod 4	26.0463(0.6088)	21.8186(0.1301)	0.4119(0.0064)	0.6455(0.0054)
	mod 5	26.9305(0.6308)	23.8629(0.1051)	0.4352(0.0057)	0.579(0.0035)

15.17% e 21.8%), respectivamente. Esse resultado foi superior em relação à modelagem utilizando ANN, que apresentou um coeficiente R^2 e MAPE, para os mesmas dias, de (0.917, 0.68 e 0.43) e (10.66%, 19.15% e 25.42%), respectivamente.

Para horizontes mais distantes, baixos valores de coeficiente R^2 podem ser explicados pela dificuldade dos modelos em responder às vazões extremas. Isto pode ser

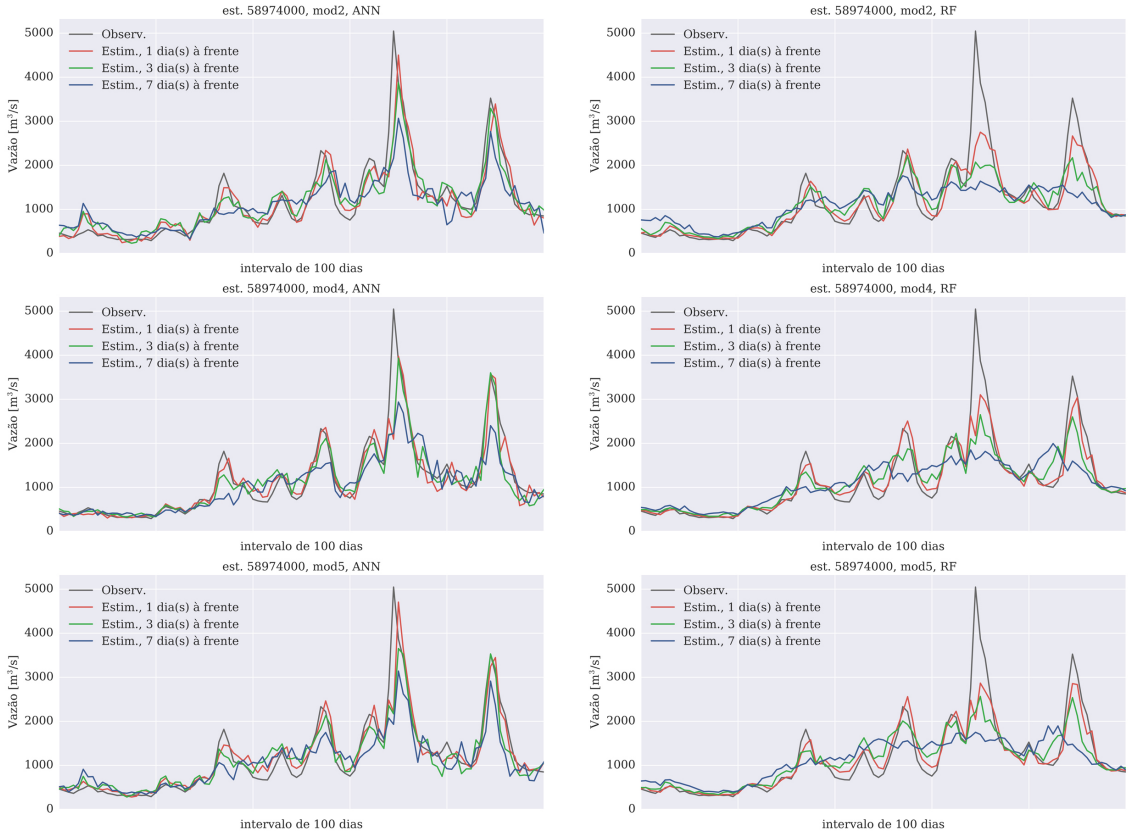

Figura 2. Vazão observada e estimada por modelos RF e ANN com as entradas "mod 2", "mod 4" e "mod 5" para o horizonte de previsão 1, 3 e 7 dias.

observado na Figura 2, que apresenta a comparação entre a série histórica de vazão observada e vazão prevista por diferentes modelos ANN e RF com 1, 3 e 7 dias de antecedência. De fato, observa-se que maiores erros acontecem nos períodos de picos da vazão. Durante períodos com vazões médias e baixas, a previsão parece ser satisfatória até para um horizonte de previsão mais distante. Notou-se que os modelos de redes neurais mostraram melhor desempenho na previsão dos picos de vazão em comparação aos modelos RF.

Observou-se que os modelos que possuíam além de vazão também a precipitação como dado de entrada mostraram-se inferiores em comparação aos modelos baseados somente na vazão. Além disso, notou-se que modelo "mod 4", que possui informação de vazão na entrada de 7 estações, produziu melhor desempenho em comparação ao modelo "mod 3", que possui informações de vazão somente de 3 estações fluviométricas.

Pode-se verificar que o modelo "mod 4", no qual a previsão das vazões foi realizada utilizando apenas dados de vazão no próprio local e à montante, obteve o menor MAPE e maior R^2 para horizontes de previsão de 1 a 7 dias. Não há exatidão na afirmação de que a utilização da precipitação na previsão de vazões seja dispensável, pois é relativamente pequena a diferença entre os erros obtidos com as outras situações. Entretanto, isso pode apontar que o uso de outras variáveis de entrada relacionadas às vazões não exerça tanto impacto na inferência de valores diários de vazões.

Os resultados para os períodos mais avançados de previsão reforçam a necessidade do desenvolvimento de métodos preditivos com maior acurácia e robustez na previsão das

vazões. Uma estratégia para aumentar a capacidade preditiva é realizar o ajuste de um modelo para um determinado dia subsequente. Neste contexto, no cenário proposto neste trabalho, seriam necessários 7 modelos, onde cada modelo é responsável pela previsão de um dia específico dentro do intervalo de previsão.

Entretanto, a precisão preditiva por si só não é uma justificativa suficiente para aplicar um modelo a um determinado problema. Os modelos devem não apenas ser precisos, mas também adequados a um propósito. Por exemplo, uma representação precisa de fluxos de baixo período de retorno é mais importante em um modelo de previsão de inundações do que uma representação que visa prever quantidades médias de água disponível para retirada e consumo humano. Da mesma forma, a capacidade de fornecer informações sobre a função física das bacias hidrográficas pode ser mais importante em bacias onde a mudança do uso da terra pode alterar o regime hidrológico, em comparação com uma bacia fortemente urbanizada. Entretanto, procedimentos de treinamento mais refinados não necessariamente abordarão outros aspectos do desempenho do modelo que o tornem adequado para fins de planejamento, como a interpretabilidade. Uma consideração mais abrangente das forças e limitações do modelo deve ser prática padrão no desenvolvimento e na seleção do modelo, em vez de simplesmente avaliar métricas de erro globais. Uma limitação das abordagens orientadas a dados para a estimativa de vazão é que os relacionamentos que elas modelam geram previsões confiáveis somente para condições comparáveis àquelas observadas historicamente. O uso destes modelos para gerar previsões em condições que excedem a variabilidade histórica pode inserir um considerável incerteza em suas previsões de vazão [Shortridge et al. 2016].

4. Conclusões

Neste trabalho, a previsão da vazão média diária do rio Paraíba do Sul é realizada para 7 dias subsequentes com base nas informações de precipitação e de vazão das estações do próprio local e à montante. A aplicabilidade dos métodos de aprendizagem de máquina, tais como RF e ANN, foi investigada na modelagem de vazões e a influência dos dados pluviométricos e fluviométricos de outros locais da bacia na qualidade da previsão foi analisada.

De acordo com os resultados, os modelos RF e ANN obtiveram desempenhos satisfatórios em relação às medidas de erro usadas e demostraram a capacidade de acompanhar a tendência dos dados de vazão observados, sendo considerada eficaz para modelagem de vazão diária com um horizonte de 1 a 7 dias. Observou-se que os modelos Random Forest apresentaram um desempenho superior em relação aos modelos ANN. Entretanto, os picos de vazão do rio não foram capturados com razoável precisão para os dias de previsão mais distantes.

Verificou-se que os conjuntos de dados que possibilitaram a obtenção de modelos com melhores desempenhos foram os conjuntos de dados históricos de vazões à montante do rio e no próprio local. Os resultados para os períodos mais avançados de previsão reforçam a necessidade, em trabalhos futuros, do desenvolvimento de métodos preditivos com maior acurácia e robustez na previsão das vazões. Para isso, uma estratégia para ampliar essa capacidade preditiva é realizar o ajuste de um modelo para um determinado dia subsequente. Objetiva-se ampliar a aplicação da mesma metodologia para todas as estações fluviométricas ao longo do rio.

Referências

[Asadi et al. 2013] Asadi, S., Shahrabi, J., Abbaszadeh, P., and Tabanmehr, S. (2013). A new hybrid artificial neural networks for rainfall–runoff process modeling. *Neurocomputing*, 121:470–480.

[Bhagwat and Maity 2012] Bhagwat, P. P. and Maity, R. (2012). Multistep-ahead river flow prediction using ls-svr at daily scale. *Journal of Water Resource and Protection*, 4(07):528.

[Breiman 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Carlisle et al. 2010] Carlisle, D. M., Falcone, J., Wolock, D. M., Meador, M. R., and Norris, R. H. (2010). Predicting the natural flow regime: models for assessing hydrological alteration in streams. *River Research and Applications*, 26(2):118–136.

[Chai and Draxler 2014] Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250.

[da Silva et al. 2011] da Silva, A. N., Chaves, M. B., Coelho, F. A., and de Oliveira Carvalho, F. (2011). Previsão de vazões diárias utilizando redes neurais na bacia do rio mundaú/al.

[Hastie et al. 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer, Verlag, New York, 2 edition.

[Haykin et al. 2009] Haykin, S. S., Haykin, S. S., Haykin, S. S., and Haykin, S. S. (2009). *Neural networks and learning machines*, volume 3. Pearson Upper Saddle River, NJ, USA:.

[Karimi et al. 2016] Karimi, S., Shiri, J., Kisi, O., and Shiri, A. A. (2016). Short-term and long-term streamflow prediction by using'wavelet–gene expression'programming approach. *ISH Journal of Hydraulic Engineering*, 22(2):148–162.

[Khair et al. 2017] Khair, A. F., Awang, M. K., Zakaraia, Z. A., and Mazlan, M. (2017). Daily streamflow prediction on time series forecasting. *Journal of Theoretical and Applied Information Technology*, 95(4):804.

[Kinga and Adam 2015] Kinga, D. and Adam, J. B. (2015). A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

[Kohavi et al. 1995] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

[Legates and McCabe 1999] Legates, D. R. and McCabe, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water resources research*, 35(1):233–241.

[Li et al. 2016] Li, B., Yang, G., Wan, R., Dai, X., and Zhang, Y. (2016). Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the poyang lake in china. *Hydrology Research*, 47(S1):69–83.

[McKinney 2010] McKinney, W. (2010). Data structures for statistical computing in python. In van der Walt, S. and Millman, J., editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56.

[Nair and Hinton 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

[Patel and Ramachandran 2015] Patel, S. S. and Ramachandran, P. (2015). A comparison of machine learning techniques for modeling river flow time series: the case of upper cauvery river basin. *Water resources management*, 29(2):589–602.

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

[Povak et al. 2013] Povak, N. A., Hessburg, P. F., Reynolds, K. M., Sullivan, T. J., McDonnell, T. C., and Salter, R. B. (2013). Machine learning and hurdle models for improving regional predictions of stream water acid neutralizing capacity. *Water Resources Research*, 49(6):3531–3546.

[Rasouli et al. 2012] Rasouli, K., Hsieh, W. W., and Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, 414:284–293.

[Ribeiro et al. 2014] Ribeiro, F. M., Mendes, E. M., and Lemos, A. P. (2014). Sistema de previsão de afluência utilizando árvore de regressão linear evolutiva nebulosa. *In Anais do XX Congresso Brasileiro de Automática*.

[Shafaei and Kisi 2016] Shafaei, M. and Kisi, O. (2016). Predicting river daily flow using wavelet-artificial neural networks based on regression analyses in comparison with artificial neural networks and support vector machine models. *Neural Computing and Applications*, 28(1):15–28.

[Shortridge et al. 2016] Shortridge, J. E., Guikema, S. D., and Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, 20(7):2611.

[Wang et al. 2015] Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., and Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527:1130–1141.

[Waskom et al. 2017] Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., Brian, Fonnesbeck, C., Lee, A., and Qalieh, A. (2017). mwaskom/seaborn: v0.8.1 (september 2017).

[Zhao et al. 2012] Zhao, T., Yang, D., Cai, X., and Cao, Y. (2012). Predict seasonal low flows in the upper yangtze river using random forests model. *Journal of Hydroelectric Engineering*, 3(005).