XXI ENMC
ENCONTRO NACIONAL DE MODELAGEM COMPUTACIONAL
IX ECTM
ENCONTRO DE CIÊNCIA E TECNOLOGIA DE MATERIAIS

08 a 11 de Outubro de 2018
Instituto Federal Fluminense
Búzios - RJ

COMPARAÇÃO DE MÉTODOS DE APRENDIZADO DE MÁQUINA PARA A PREVISÃO DE CURTO PRAZO DE VAZÃO DO BAIXO CURSO DO RIO PARAÍBA DO SUL

Yulia Gorodestskaya[1] - yu.gorodetskaya@gmail.com
Gisele Goulart Tavares[1] - giselegoulart@ice.ufjf.br
Leonardo Goliatt da Fonseca[2] - leonardo.goliatt@ufjf.edu.br
Celso Bandeira de Melo Ribeiro[3] - celso.bandeira@ufjf.edu.br
[1]Pós-Graduação em Modelagem Computacional - UFJF, Minas Gerais, Brasil
[2]Departamento de Mecânica Aplicada e Computacional - UFJF, Minas Gerais, Brasil
[3]Departamento de Engenharia Sanitária e Ambiental - UFJF, Minas Gerais, Brasil

Resumo. *A bacia hidrográfica do rio Paraíba do Sul é responsável pelo abastecimento de cidades do estado do Rio de Janeiro, fluindo através de uma importante região industrial do Brasil. Com isso, seus recursos hídricos são utilizados de diversas formas, aumentando a importância do estudo.Esta previsão pode assumir valor estratégico na gestão da quantidade e da qualidade da água nesta bacia. Nesse contexto, este estudo tem como objetivo prever a vazão natural diária de curto prazo, com um horizonte de até 7 dias a frente, da estação Campos-Ponte Municipal do baixo curso do rio Paraíba do Sul. Uma base de séries históricas de precipitação e vazão foi utilizada na modelagem da previsão da vazão e a capacidade dos métodos de aprendizagem de máquina, tais como RF e ANN, frente a um modelo linear foi investigada na modelagem das previsões. De acordo com os resultados, todos os métodos de aprendizagem de máquina obtiveram desempenhos satisfatórios em relação às medidas de erro utilizadas para o horizonte de previsão, de modo que estes métodos podem vir a auxiliar no acompanhamento e previsão do fluxo de bacias hidrográficas.*

Keywords: *Métodos de aprendizagem de máquina, Previsão de vazão, Paraíba do Sul*

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, abastecendo a cidade do Rio de Janeiro e a região do Grande Rio. De acordo com Comitê de Integração da Bacia Hidrográfica do Rio Paraíba do Sul – CEIVAP, 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 (Comitê de Integração da Bacia Hidrigráfica do Rio Paraíba do Sul-CEIVAP, 2014).

A região do baixo curso do rio Paraíba do Sul (RH-IX) possui menores índices de precipitação anual. Devido ao ciclo de corte de árvores e queimadas regulares, restaram apenas 10% de cobertura da área por florestas em um revelo em que predominam planícies e colinas. Adicionamente, a região teve suas planícies alteradas por obras em seus sistemas fluviais realizadas pelo Departamento Nacional de Obras de Saneamento (DNOS, a partir da década de 1950), para torná-las aptas à exploração agrícola canavieira. Essas terras possuem um perfil de exploração alterado com o declínio da atividade canavieira nas últimas décadas, sendo ou convertidas para pastagem ou utilizadas na ocupação urbana. A expansão da demanda regional pela construção civil ampliou a exploração de argila e a instalação de pedreiras. Os rios Pomba e Muriaé, principais afluentes do rio Paraíba do Sul na RH-IX, nascem na Zona da Mata mineira. A intensa exploração de recursos minerais e eventos hidrológicos críticos na parte mineira exercem grande impacto sobre a gestão de recursos hídricos na região fluminense (Fundação COPPETEC, 2014).

Em função da importância da RH-IX, 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. Os modelos de precipitação-vazão são amplamente utilizados na previsão de vazão e no apoio à concepção de gestão de recursos hídricos e inundações. Uma abordagem para a modelagem da previsão de fluxo 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). Os modelos não-lineares baseados em métodos de aprendizagem de máquina conseguem identificar a relação direta entre as entradas e saídas sem consideração detalhada da estrutura interna do processo físico. As principais vantagens da máquina de aprendizagem em relação às técnicas de regressão estatística, são que os modelos resultantes se mostram mais resistentes à multicolinearidade e valores extremos, incluem métodos que reduzem $overfitting$ do modelo, melhoram a identificação das variáveis preditoras importantes, das relações não-lineares e das interações complexas entre preditores, e não são afetados por transformações de dados (Povak et al., 2013). Os métodos de aprendizado de máquina têm sido aplicados com sucesso para resolver problemas não-lineares em hidrologia (Rasouli et al., 2012; Povak et al., 2013; Khair et al., 2017; Bhagwat and Maity, 2012).

Neste trabalho, os métodos de aprendizagem de máquina Random Forest (RF) e Redes Neurais Artificiais (ANN), e o modelo linear Multi-task ElasticNet (LM) (Pedregosa et al., 2011) são aplicados ao problema de modelagem de vazão diária de curto prazo. A modelagem se baseia em valores previamente medidos de vazão e precipitação, com um horizonte de até 7 dias à frente, da estação Campos-Ponte Municipal no baixo curso do rio Paraíba do Sul. Além disso, o horizonte temporal sobre o qual isto pode se apoiar é investigado.

2. MATERIAIS E MÉTODOS

2.1 Área de Estudo e Conjunto de Dados

Todos os modelos, RF, ANN e LM, foram implementados utilizando-se séries históricas de dados diários de vazão e precipitação (Tabela 1), provenientes da Agência Nacional de Águas, entre os anos de 1990 a 2016, registrados na estação Campos-Ponte Municipal que está locali-

zada no Região Hidrográfica do Baixo Paraíba do Sul.

Tabela 1- Postos fluviométricos e pluviométricos utilizados para aquisição dos dados.

Tipo do posto	Código	Período de observação
fluviométrico	58974000	01/01/1990 a 31/05/2016
pluviométrico	2141002	01/01/1990 a 31/05/2016

No processo de modelagem foi necessário lidar com problemas nas bases de dados disponibilizadas. 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 as 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 Procedimento de previsão

Para realizar as previsões, foram selecionados, na série histórica de vazão, períodos com 21 dias, e, na série histórica de precipitação, períodos com 14 dias de informações contínuas, como mostrado na Figura 1, ou seja, onde não havia dados faltantes. O modelo de previsão recebe como entrada 28 valores, 14(P) de precipitação e 14(V) de vazão, e retorna 7(V) valores correspondentes às vazões estimadas. Consideramos, portanto, como uma amostra de dados, um conjunto com informações ininterruptas (X, y), onde X = (14(V), 14(P)), composto por 14 dias antecedentes de vazão, 14 dias antecedentes de precipitação e y = (7(V)), composto por 7 dias seguintes de vazão. De uma amostra de dados da série histórica de vazão, os 14 primeiros dias são usados para treinar o modelo e os 7 últimos dias usados para validar o treinamento realizado e fornecer uma estimativa do desempenho do modelo previsor. Ressaltamos que o mesmo conjunto de parâmetros é usado para treinar o modelo que estima a vazão dos 7 dias subsequentes. Desta forma, chamaremos por conjunto de dados de entrada completo todas as amostras $(X_i, y_i),\ i = 1, \ldots, N$, onde N é o número de todas as amostras selecionadas na série histórica, como mostrado na Figura 1.

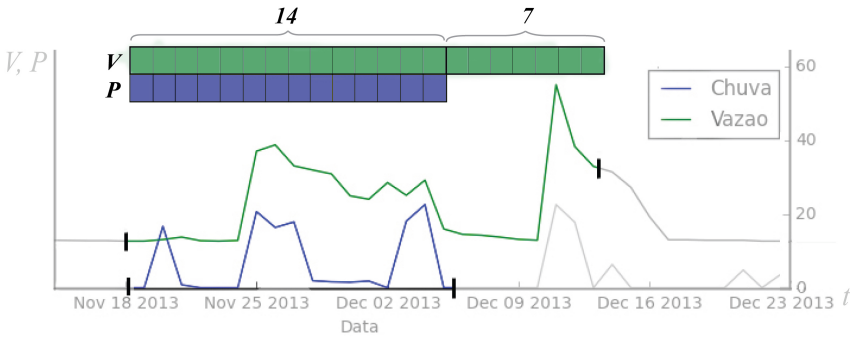

Figura 1- Amostra de dados com um conjunto de informações ininterruptas de 21 dias de vazão e 14 dias de precipitação.

Consequentemente, o modelo previsor resultante tem a seguinte forma:

$$(V_{t+1}, V_{t+2}, ..., V_{t+7}) = F(V_t, V_{t-1}, ..., V_{t-13}, P_t, P_{t-1}, ..., P_{t-13}),$$

onde a função F representa a relação entre as vazões e precipitações dos 14 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-13} denotam diferentes defasagens temporais da vazão natural do rio observada na estação fluviométrica nos dias t, $t-1$, $t-2$, ... , $t-13$ que antecedem a previsão; P_t, P_{t-1}, P_{t-2}, ... , P_{t-13} denotam diferentes defasagens temporais de precipitação natural observada na estação pluviométrica nos dias t, $t-1$, $t-2$, ... , $t-13$, respectivamente.

Em geral, em trabalhos correlatos (Shafaei and Kisi, 2016; Ribeiro et al., 2014; Bhagwat and Maity, 2012), os autores optam por construir e treinar um modelo para cada dia subsequente. Alternativamente, neste artigo um mesmo modelo foi usado para, a partir dos dados colhidos para os 14 dias anteriores, realizar a previsão de todos os 7 dias subsequentes num ponto da bacia. Esse critério buscou facilitar o treinamento e a interpretabilidade do modelo visando obter no futuro, a um custo computacional menor, sua aplicabilidade para toda bacia hidrográfica.

2.3 Métodos de Aprendizado de Máquina

Random Forest é um algoritmo de aprendizado de máquina proposto por Breiman (2001). É um método de combinação de classificação baseado na teoria de aprendizado estatístico. 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 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). Nos modelos RF, três parâmetros devem ser definidos: (1) o número de árvores ($ntree$); (2) o número de variáveis preditoras randomicamente selecionadas para cada nó ($mtry$); e (3) o número mínimo de observações nos nós terminais das árvores ($nodesize$) (Li et al., 2016). A partir de testes preliminares optou-se por utilizar 250 árvores neste estudo.

Rede Neural Artificial é um modelo não paramétrico composto de vários neurônios artificiais organizados em camadas, que são usualmente classificadas em três grupos: camada de entrada, camada de saída e camada intermediária ou oculta. ANNs com uma camada oculta são comumente utilizadas em modelagem hidrológica. O mecanismo básico consiste em treinar uma rede paralelamente interconectada de nós com dados reais, ou seja, dados empíricos, com amostras ou exemplos de padrões de entrada-saída para construir uma relação não-linear entre os exemplos e, a partir disso, criar um modelo que permita realizar previsões (Schaefli and Gupta, 2007). As informações teóricas detalhadas sobre ANN podem ser encontradas em Haykin (2009) e Nissen (2005).

No presente estudo, foi treinada uma ANN do tipo MLP com uma camada oculta de 500 neurônios e função de ativação ReLU (Rectified Linear Unit) (Nair and Hinton, 2010) minimizando o erro médio quadrático. A rede foi treinada usando 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 dos Modelos e Medidas de Desempenho

Um dos problemas relacionados aos modelos de predição é o chamado $overfitting$, 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.

Nos estudos semelhantes 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 completo (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).

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. Seu retorno é dado em porcentagem ($\%$) e, quanto mais próximo de zero, melhor é o modelo. Neste trabalho é considerado aceitável para este índice valores abaixo de 30%. O MAPE é calculado de acordo com a seguinte equação:

$$\text{MAPE} = 100 \times \frac{1}{N} \sum_{i=1}^{N-1} \frac{|Q_{ob} - Q_{pr}|}{|Q_{ob}|}, \tag{1}$$

  onde Q_{ob} representa a vazão observada, Q_{pr} a vazão predita pelo modelo e N é o tamanho da série histórica.

- O Coeficiente de eficiência de Nash-Sutcliffe (NS) descreve a proporção da variância total nos dados observados que pode ser explicada pelo modelo, que varia de $-\infty$ a 1, sendo 1 a representação do modelo perfeito e zero significando que o modelo é tão preciso quanto os meios dos dados observados. Valores negativos significam que a média dos dados observados poderia ser um estimador melhor que o próprio modelo. O cálculo deste coeficiente é realizado de acordo com a seguinte equação (Shafaei and Kisi, 2016):

$$NS = 1 - \frac{\sum_{i=0}^{N-1} (Q_{ob} - Q_{pr})^2}{\sum_{i=0}^{N-1} (Q_{ob} - \overline{Q_{ob}})^2}, \tag{2}$$

  onde N e t são o número de dados e o tempo, respectivamente. Q_{ob} é a vazão observado, Q_{pr} é a vazão predita, $\overline{Q_{ob}}$ é a média da vazão observada.

  Baltokoski et al. (2010) notam que, quando o valor de NS resultar maior que 0.75, o desempenho do modelo é considerado bom. Para valores de NS entre 0.36 e 0.75, o desempenho é considerado aceitável, enquanto valores de NS inferiores a 0.36 fazem com que o modelo seja julgado como inaceitável.

A preparação dos dados foi realizada utilizando-se OpenOffice Calc 4.1 e a linguagem de programação Python 3.5. Especificamente, foi utilizada a implementação dos modelos disponíveis na biblioteca scikit-learn (Pedregosa et al., 2011).

3. RESULTADOS E DISCUSSÃO

Com objetivo de alcançar uma melhor generalização, e assim melhores resultados, cada um dos modelos RF, ANN e LM teve o processo de treinamento repetido 25 vezes, onde para cada iteração foi utilizado a técnica de validação cruzada. Os dados referentes a precipitação e vazão de 26 anos foram utilizados como dados de entrada nos modelos para a previsão de vazão em um período de 1-7 dias.

Tabela 2- Variação de medidas de desempenho MAPE e NS referentes a cada um dos modelos de previsão de vazão para os dias de 1 a 7.

Dia	MAPE(%)			Coeficiente NS		
	LM	RF	ANN	LM	RF	ANN
1	9.8499	**9.4040**	11.0435	**0.9161**	0.9056	0.9078
2	16.3924	**14.895**	15.5579	0.7709	**0.7822**	0.7667
3	21.7504	19.2707	**18.5970**	0.6471	**0.6883**	0.6496
4	25.8712	22.6593	**20.7972**	0.5700	**0.6269**	0.5721
5	28.8671	25.1277	**22.5975**	0.5176	**0.5899**	0.5141
6	31.1120	26.9867	**23.8648**	0.4778	**0.5616**	0.4696
7	32.7309	28.5439	**24.6646**	0.4481	**0.5325**	0.4350

Observou-se o aumento gradual do MAPE e o decréscimo do coeficiente NS com o aumento do alcance da previsão para todos os modelos.

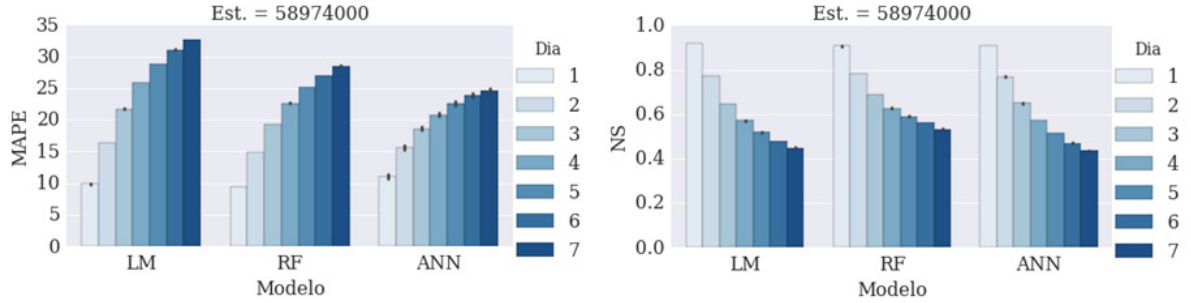

Figura 2- Variação do MAPE e do NS para cada modelo de previsão.

Comparando o desempenho dos modelos para os dias 1 e 2, observou-se que o modelo RF produziu o menor MAPE, com valores abaixo de 10% e 15%, respectivamente. Porém, conforme a Tabela 2, notou-se que a rede neural apresentou os menores erros percentuais para os dias 3-7 com valores médios menores que 25%. Ao considerar os resultados do coeficiente NS, conforme a Tabela 2, é possível perceber que o desempenho para o primeiro dia é semelhante para todos os modelos. Entretanto, com a evolução do período de previsão, o RF produziu um desempenho ligeiramente superior que os modelos de regressão linear e de redes neurais.

A Figura 3 mostra a relação entre vazão observada e predita por RF, ANN e LM ao longo de todo o período de observação para o primeiro, terceiro e sétimo dia. Observou-se a piora

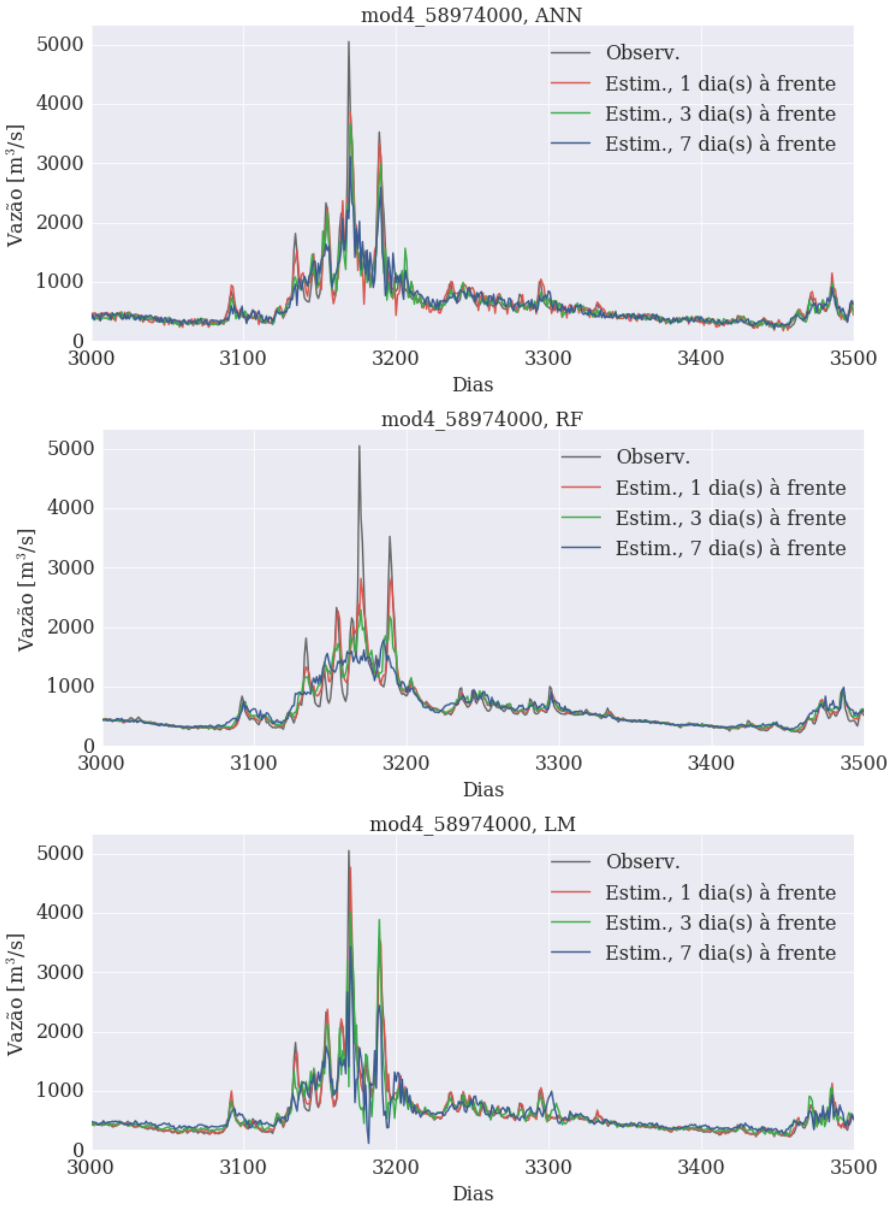

Figura 3- Comparação entre vazão observada e prevista por modelos ANN, RF e LM para primeiro, terceiro e sétimo dia.

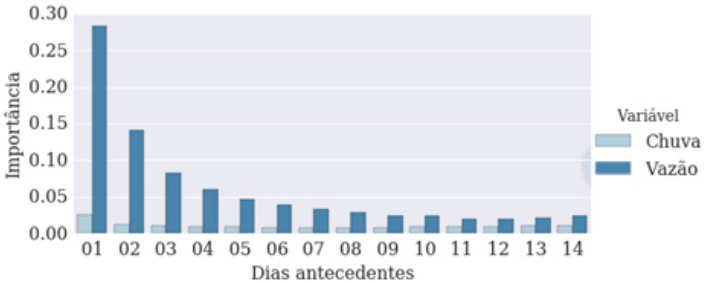

Figura 4- Importância relativa das informações de precipitação e de vazão para o modelo Random Forest em função dos dias antecedentes que compõem o modelo.

na qualidade da previsão com o avanço dos dias, principalmente na capacidade de previsão dos picos de vazão. Notou-se, também, que os modelos de redes neurais mostraram melhor desempenho na previsão dos picos de vazão.

Ressalta-se que a boa capacidade preditiva dos picos de vazão é essencial para melhorar o controle de inundações e a operação de obras hidráulicas, como reservatórios de energia. Em reservatórios, a previsão pode auxiliar a operação segura das comportas em obras, que condicionam a montante, a jusante e a produção de energia (da Silva et al., 2006). 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. 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. Uma estratégia para aumentar a capacidade preditiva é realizar o ajuste de um modelo para um determinado dia subsequente.

O processo de treinamento das árvores de decisão que compõem o modelo RF envolve o cálculo da importância relativa das variáveis para as estimativas das vazões. Os valores das importâncias relativas de cada variável podem ser recuperados e fornecem uma informação importante sobre a relação entre as variáveis utilizadas como entrada para o modelo e sua respectiva saída. A Figura 4 ilustra a média em 25 execuções independentes da importância relativa das informações de precipitação e de vazão para o modelo RF em função dos dias antecedentes que constituem as entradas do modelo. Observa-se que as informações fluviométricas (provenientes das vazões observadas), para o modelo RF, têm maior importância que as informações pluviométricas (proveniente da quantidade de precipitação medida). Nota-se também que as importâncias das vazões referentes aos dias antecedentes diminui com o aumento do período com o qual as informações são consideradas. Este resultado sugere que aumentar a quantidade de informações antecedentes pode não contribuir para a acurácia do modelo RF.

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 (Shortridge et al., 2016).

4. CONCLUSÕES

Neste trabalho, a previsão da vazão de curto prazo da região do baixo curso do rio Paraíba do Sul foi realizada para 7 dias subsequentes, com base nas informações de precipitação e de vazão coletadas nos 14 dias anteriores. A capacidade dos métodos de aprendizagem de máquina Random Forest e Redes Neurais Artificiais frente a um modelo linear foi investigada na modelagem das previsões. De acordo com os resultados obtidos os modelos RF e ANN obtiveram desempenhos satisfatórios em relação aos medidas de erros usadas. Na maioria dos casos, foi percebida uma equivalência entre ambas as técnicas de aprendizagem de máquina ainda que, em raros casos, os testes estatísticos tenham apontado superioridade de uma em

relação à outra. Por outro lado, conclui-se que árvores randômicas e redes neurais produziram um desempenho ligeiramente superior que o modelo de regressão linear. Entretanto, as vazões máximas não foram capturadas com razoável precisão para os dias de previsão mais distantes. Observou-se também que, para o modelo Random Forest, as informações de vazões passadas têm maior importância do que as informações de precipitação passada da região da estação modelada.

Os resultados deste estudo podem ser úteis para engenharia hidrológica, tomadores de decisão em previsão de vazão de rios, reação a inundações em cidades, possibilidade de colapso de obras hidráulicas e planejamento de disponibilidade hídrica em áreas urbanas, na irrigação, navegação fluvial e na distribuição sustentável de água. O horizonte de previsão adotado neste estudo, para até 7 dias de antecedência, permite que, em casos extremos, seja possível o fornecimento de informações estratégicas para a mobilização de recursos humanos na esfera pública. Busca-se estender a metodologia desenvolvida neste estudo para outras estações fluviométricas do rio Paraíba do Sul.

AGRADECIMENTOS

Os autores agradecem às agências CAPES e FAPEMIG pela concessão de bolsas de pesquisa, e ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de pelo auxílio financeiro e apoio na realização deste trabalho.

Referências

Baltokoski, V., Tavares, M. H. F., Machado, R. E., and Oliveira, M. d. (2010). Calibração de modelo para a simulação de vazão e de fósforo total nas sub-bacias dos rios conrado e pinheiro–pato branco (pr). *Revista Brasileira de Ciência do Solo*, 34(1):253–261.

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

Comitê de Integração da Bacia Hidrigráfica do Rio Paraíba do Sul-CEIVAP (2014). Dados geoambientais. http://www.ceivap.org.br/geoambientais.php. Acesso em: 14 set. 2018.

da Silva, B. C., Tucci, C. E., and Collischonn, W. (2006). Previsão de vazão com modelos hidroclimáticos.

Fundação COPPETEC (2014). Elaboração do plano estadual de recursos hídricos do estado do Rio de Janeiro. R2-F-caracterização ambiental. https://goo.gl/NNdpDQ. Acesso em: 14 set. 2018.

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

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

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.

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.

Nissen, S. (2005). Neural networks made simple. *Software*, 2(2):14–19.

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, 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, 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, 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*.

Schaefli, B. and Gupta, H. V. (2007). Do nash values have value? *Hydrological Processes*, 21(15):2075–2080.

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, 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.

APÊNDICE A

COMPARISON OF MACHINE LEARNING METHODS FOR THE SHORT-TERM FORECAST OF THE LOW LANDS OF PARAÍBA DO SUL RIVER

Abstract.

The Paraíba do Sul river basin is responsible for supplying cities in the state of Rio de Janeiro, flowing through an important industrial region of Brazil. With this, their water resources are used in several ways, increasing the importance of the study. This forecast can assume strategic value for managing the quantity and quality of water in this basin. The ability of machine learning methods, such as random forest and artificial neural network, against a linear model was investigated in modeling predictions. Each model is trained on historical stream flow and precipitation data to forecast stream flow with a lead time of up to seven days. Campos-Ponte Municipal station on the Paraíba do Sul river (RH-IX) are considered for experimentation. According to the results, it was found that all the machine learning methods have performed satisfactorily in relation to the error measures used for the forecast horizon, so that these methods can help to monitor and predict the flow of watersheds.

Keywords: *Machine learning methods, Stream flow forecast, Paraíba do Sul River*