COOPERAÇÃO EM TECNOLOGIAS PARA ANÁLISES HIDROLÓGICAS EM ESCALA NACIONAL

SUBPROJETO – REGIONALIZAÇÃO DE VAZÕES VIA MODELAGEM HIDROLÓGICA

ESTIMATIVA DE VAZÃO EM LOCAIS SEM DADOS USANDO MACHINE LEARNING

IPH-ANA-HGE-SR-R8

Porto Alegre - RS

Outubro 2021

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE PESQUISAS HIDRÁULICAS

ANA
AGÊNCIA NACIONAL DE ÁGUAS E SANEAMENTO BÁSICO

Estimativa de vazão em locais sem dados usando Machine Learning

ESTE MATERIAL FAZ PARTE DE UM CONJUNTO DE RELATÓRIOS CRIADOS NO CONTEXTO DO PROJETO DE COOPERAÇÃO EM TECNOLOGIAS PARA ANÁLISES HIDROLÓGICAS EM ESCALA NACIONAL, ENTRE O INSTITUTO DE PESQUISAS HIDRÁULICAS (IPH-UFRGS) E A AGÊNCIA NACIONAL DE ÁGUAS E SANEAMENTO BÁSICO (ANA).

AUTORES: Rafael Barbedo, Mino Sorribas, Walter Collischonn, Rodrigo Paiva.

COMO CITAR: Barbedo, R., Sorribas, M. V., Collischonn, W., Paiva, R. C. D., 2021. Cooperação em tecnologias para análises hidrológicas em escala nacional: Estimativa de vazão em locais sem dados usando machine learning: IPH-ANA-HGE-SR-R8. UFRGS: IPH, [Porto Alegre]. ANA, [Brasília].

Porto Alegre - RS

Outubro 2021

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE PESQUISAS HIDRÁULICAS

ANA
AGÊNCIA NACIONAL DE ÁGUAS E SANEAMENTO BÁSICO

Sumário

1 INTRODUÇÃO

O uso de modelos estatísticos para estimar vazões em locais sem dados não é novo, e vários métodos existem para isso, normalmente baseados em regressões lineares calibradas para a região de estudo relacionando algumas características físicas da bacia. Recentemente, com o advento de técnicas de aprendizado de máquina (Machine Learning), esses modelos estatísticos têm apresentado cada vez melhores resultados (Worland et al. 2018, Ferreira et al. 2021, Golian et al. 2021). Neste relatório, apresentamos alguns resultados obtidos testando alguns modelos de Machine Learning (ML) com o objetivo de estimar vazões em locais sem dados.

2 MATERIAIS E MÉTODOS

2.2 MODELOS ANALISADOS

2.2.1 Regressão linear (LR)

A regressão linear é um método estatístico que reconhece padrões que representam a resposta do modelo (variável dependente) a partir de uma função linear de todas as variáveis independentes (preditoras) (Rong and Bao-wen 2018). Ela pode ser simples ou multivariada. A regressão linear multivariada, que foi utilizada neste estudo, pode ser descrita pela fórmula:

$$Y = \beta_0 + \beta_i x_i + \cdots + \beta_n x_n, \quad (1)$$

onde Y é a variável de interesse, β são os coeficientes, e x são as variáveis preditoras.

2.2.2 K-Nearest Neighbors (KNN)

Modelos K-Nearest Neighbors utilizam a proximidade entre as variáveis no domínio amostral para realizar previsões (Altman 1992). Isto é, bacias com mais características similares estão mais "perto" uma da outra. Assim, para uma nova bacia, são selecionadas N bacias mais similares àquela de interesse, e o valor resposta é a média dos valores das bacias próximas.

2.2.3 Random Forest (RF)

O Random Forest é um algoritmo não linear e não paramétrico, que combina princípios de regressão e classificação baseados em árvores de decisão, associados com um grau de aleatoriedade nas decisões (amostragem randomizada ou bootstrap) (Breiman 2001, Tyralis *et al.* 2019). Isto é, em cada árvore de decisão são criados subsets de amostras que utilizarão um

número aleatório de preditores, que posteriormente serão agregados na árvore final, reduzindo a variância dos preditores e evitando “overfitting”.

2.2.4 Support Vector Machines (SVM)

Support Vector Machines encaixam linhas de regressão usando apenas pontos de dados (vetores de suporte) que caem fora de um limiar definido pelo usuário (Awad and Khanna 2015). Os resíduos fora do limiar contribuem para a adequação do modelo. O efeito dos resíduos são controlados por um parâmetro de custo que tem um efeito de regularização. Uma função kernel é utilizada para encontrar relações não lineares entre as variáveis. Diferentes funções têm diferentes efeitos no resultado do modelo. Neste estudo, foi utilizada uma função polinomial de grau 3, por ter apresentado os melhores resultados.

2.2 VARIÁVEIS DE RESPOSTA

Duas variáveis de interesse foram analisadas nesse estudo, que refletem o comportamento hidrológico de uma bacia no escopo do gerenciamento dos recursos hídricos. São elas: a vazão média de longo termo (*Qm*) e a vazão com permanência de 95% do tempo (*Q95*), calculados a partir de séries diárias de postos fluviométricos obtidos da base de dados Hidroweb da ANA. Foram considerados somente postos com séries de dados ao longo do período entre jan/1980 e dez/2014 com pelo menos 20 anos de dados. Além disso, considerou-se somente postos sem efeito significativo de regularização artificial e/ou erros grosseiros nas séries de dados, ou de interesse especial para a ANA. Uma análise de erros de estimativa na vazão de longo termo associados a variabilidade amostral (Collischonn et al. 2021) demonstraram que as estimativas a partir de 20 anos de dados apresentam erros inferiores a 5% na *Qm* e inferiores a 15% na *Q95* (quando comparados a séries de 30 a 35 anos) na maioria dos casos avaliados. Considerando esses critérios, foram identificados 1069 postos fluviométricos. Para o treinamento e avaliação dos modelos, as variáveis utilizadas foram utilizadas em unidades específicas (*qm* e *q95*), dividindos os valores de *Qm* e *Q95* pela área da bacia a montante de cada posto.

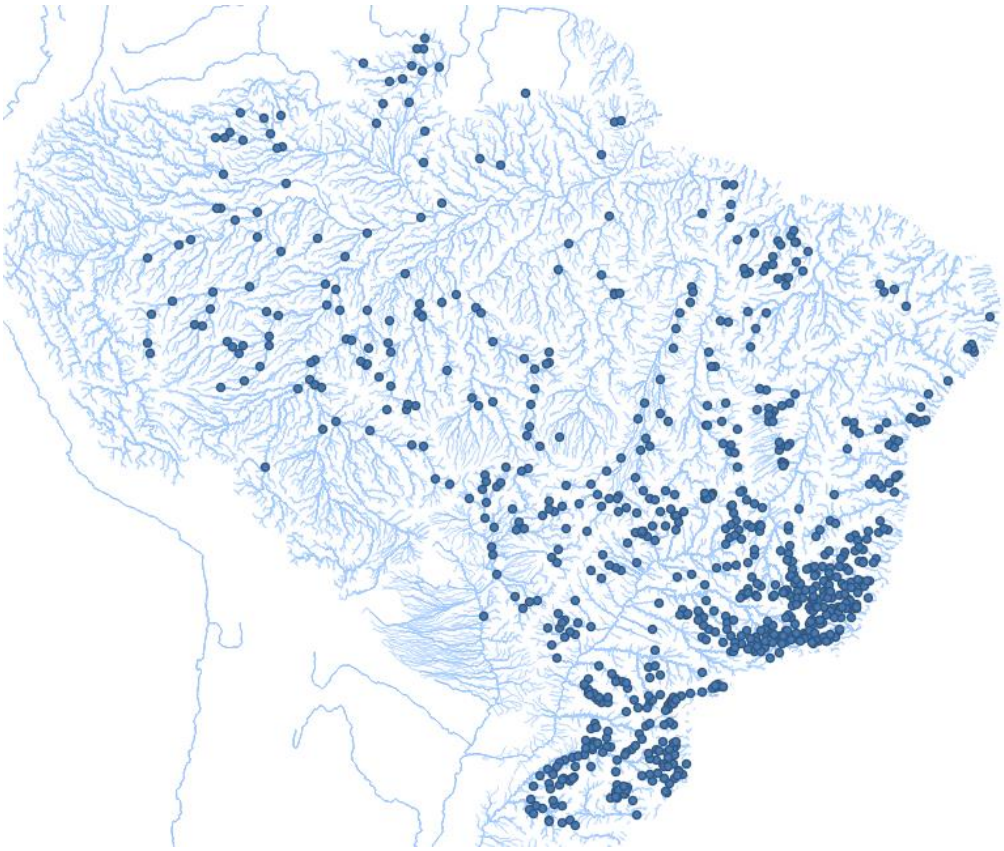

Figura 1. Postos fluviométricos utilizados para avaliação dos modelos de ML.

2.3 VARIÁVEIS PREDITORAS

Baseado em estudos anteriores (Worland *et al.* 2018, Ferreira *et al.* 2021) e também em conhecimento dos autores, 32 variáveis com potencial preditivo foram consideradas (Tabela 1). As variáveis refletem características das bacias em termos de clima, topografia, morfometria, e uso e cobertura do solo. As variáveis foram primeiramente coletadas para todas as bacias BHO5k e posteriormente agregadas para a área a montante de cada posto fluviométrico.

Todos os dados climáticos (CL) foram coletados a partir de médias mensais do período 2001-2020, sendo Xavg a média mensal, Xmin a mínima média mensal, e Xmax a máxima média mensal da variável X. Os dados de precipitação (P) foram obidos a partir do produto GPM (Hou *et al.* 2014). Os dados de temperatura (T) e de evapotranspiração potencial (E) foram obtidos a partir do produto ERA5 (Muñoz Sabater 2019).

Os dados de topografia (TP) foram obtidos a partir do MERIT-DEM (Yamazaki *et al.* 2017), sendo estes a média e o desvio padrão da elevação (elv_avg e elv_std) e da declividade (s_avg e s_std). Também nesta categoria estão os dados de altura à rede de drenagem mais próxima (hnd_avg e hnd_std), obtida a partir do algoritmo HAND (Rennó *et al.* 2008) com a drenagem definida pelo método TPS (Barbedo et al. 2021). A drenagem também foi utilizada para computar a densidade de drenagem da bacia (dd). A área foi obtida pelo atributo *"nuareamont"* do trecho correspondente ao posto na rede BHO5k.

As classes de solo (ST) foram obtidas com base no HAND e em classes de declividades, metologia utilizada em alguns estudos hidrológicos (Savenije 2010, Gharari *et al.* 2011, Gao *et al.* 2014). Onde wetland correponde aos pixels onde $HAND < 5\ m$, e no restante da bacia são utilizados limiares de declividades (S), sendo flat onde $S < 3\ \%$, gentle onde $S < 8\ \%$, moderate onde $S < 20\ \%$, steep onde $S < 45\ \%$, e extreme onde $S > 45\ \%$.

As classes de cobertura do solo (LC) foram obtidas a partir do Mapbiomas (Souza *et al.* 2020) no território brasileiro e MOD12 (Friedl, Mark and Sulla-Menashe, Damien 2019) na área restante. Estas foram reclassificadas, para fins de simplificação, em florestas (forest), campos naturais (grassland), agricultura e pecuária (agriculture), regiões urbanas ou rochas expostas (semi-permeable) e águas abertas (water).

Tabela 1. Variáveis com potencial poder de predição.

CL	TP	MF	LC	ST
p_avg	elv_avg	dd	forest	wetland
p_min	elv_std	A	grassland	flat
p_max	s_avg		agriculture	gentle
t_avg	s_std		semi-permeable	moderate
t_min	hnd_avg		water	steep
t_max	hnd_std			extreme
e_avg				
e_min				
e_max				

2.4 APLICAÇÃO DOS MODELOS

A aplicação dos modelos foi dividida em duas etapas. Primeiro, de todas as variáveis preditoras coletadas, somente algumas foram selecionadas, de acordo com suas importâncias. Depois os modelos foram submetidos à etapa de treinamento e validação. Mais detalhes sobre as etapas serão cobridos em seguida.

2.4.1 Seleção das variáveis preditoras

A seleção das variáveis foi feita através da técnica de Recursive Feature Elimination (RFE) (Chen and Jeong 2007). Isto é, o modelo selecionado é treinado em todos os pontos e com todas as variáveis, e são computadas as importâncias de cada variável admitos pelo próprio modelo. A variável com menor importância é eliminada e o processo se repete até que o número de variáveis (parâmetro da RFE) seja atingido. Foram testados diferentes números de variáveis (indo de 5 a 25), sendo selecionado o número a partir do qual mais variáveis não apresentavam melhora significativa nos resultados. A técnica RFE só pode ser usada em modelos que computam coeficientes ou importâncias das variáveis preditoras, que neste caso seriam LR e RF, respectivamente. O modelo RF, então, foi utilizado para selecionar as variáveis dos modelos RF, SVM e KNN e então as mesmas foram utilizadas na etapa posterior, o que é uma pratica comum em aplicações de ML. As variáveis do modelo LR foram selecionadas por LR.

2.4.2 Treinamento e validação dos modelos

As etapas de treinamento e validação de cada modelo foi feito com Leave-One-Out Cross-Validation (LOOCV). Neste procedimento, é retirado um ponto amostral em cada rodada do modelo, e neste ponto o modelo é testado, totalizando 1058 rodadas por modelo (uma para cada ponto amostral). Assim, o treinamento do modelo é realizado com a remoção da amostra de interesse para previsão, o que simularia uma situação real em estimativas de vazões em locais sem dados. Os modelos foram avaliados através das seguintes métricas: viés em porcentagem (BIAS, eq. 2), raiz do erro médio quadrático (RMSE, eq. 3), coeficiente de determinação (R^2, eq. 4) e percentil 75 da máxima razão entre as vazões (RQ75, eq. 5).

$$BIAS = \left[\frac{\sum(Qpred - Qobs)}{\sum(Qobs)}\right] \times 100 \quad (2)$$

$$RMSE = \sqrt{\frac{\sum(Qobs - Qpred)^2}{n}} \quad (3)$$

$$R^2 = 1 - \frac{\sum(Qobs - Qpred)^2}{\sum(Qobs - \overline{Qobs})^2} \quad (4)$$

$$RQ75 = P75\left[\max\left(\frac{Qobs}{Qpred}, \frac{Qpred}{Qobs}\right)\right] \quad (5)$$

3 RESULTADOS E DISCUSSÃO

3.2 Variáveis selecionadas e respectivas importâncias

Das 30 variáveis preditoras iniciais, foram selecionadas 12 pelo método RFE. As variáveis e seus respectivos coeficientes (para LR) e importâncias (para os outros modelos) podem ser visualizadas na Figura 2. O número 12 foi escolhido por apresentar melhores resultados em todos os modelos. Na regressão linear (LR), as variáveis com coeficientes de maior magnitude corresponderam às 6 classes de solo (wetland, flat, gentle, moderate, steep, extreme), tanto na *qm* quanto na *q95*. Esses coeficientes referem-se aos valores das variáveis normalizadas (etapa do pré-processamento), que vão sempre de 0 (valor mínimo da amostra) a 1 (valor máximo da amostra). As outras variáveis, apesar dos coeficientes aparentarem ser pequenos, tiveram bastante influência nos resultados. A precipitação média (p_avg) foi a única variável (além das classes de solo) que foi selecionada nas duas vazões de referência.

Para o restante dos modelos, como já descrito na metodologia, o método RFE foi utilizado a partir do RF, que é o único modelo entre os utilizados aqui que computa a importância das variáveis nas predições. Nesses casos, a importância da precipitação média (p_avg) foi consideravelmente maior que às demais variáveis, chegando a *0.68* na *qm* e *0.38* na *q95* (soma das importâncias = 1). Em segundo lugar, aparece a precipitação mínima (p_min), e em terceiro a elevação média (elv_avg), em ambas as bacias. Algumas diferenças notáveis são a presença de valores de HAND (hnd_avg e hnd_std) na *qm* e de evapotranspiração potencial (e_max e e_min) e densidade de drenagem (dd) na *q95*. Com relação às outras variáveis, temos a área (A) e temperatura média (t_avg) nas duas vazões, e algumas classes de solo e cobertura vegetal.

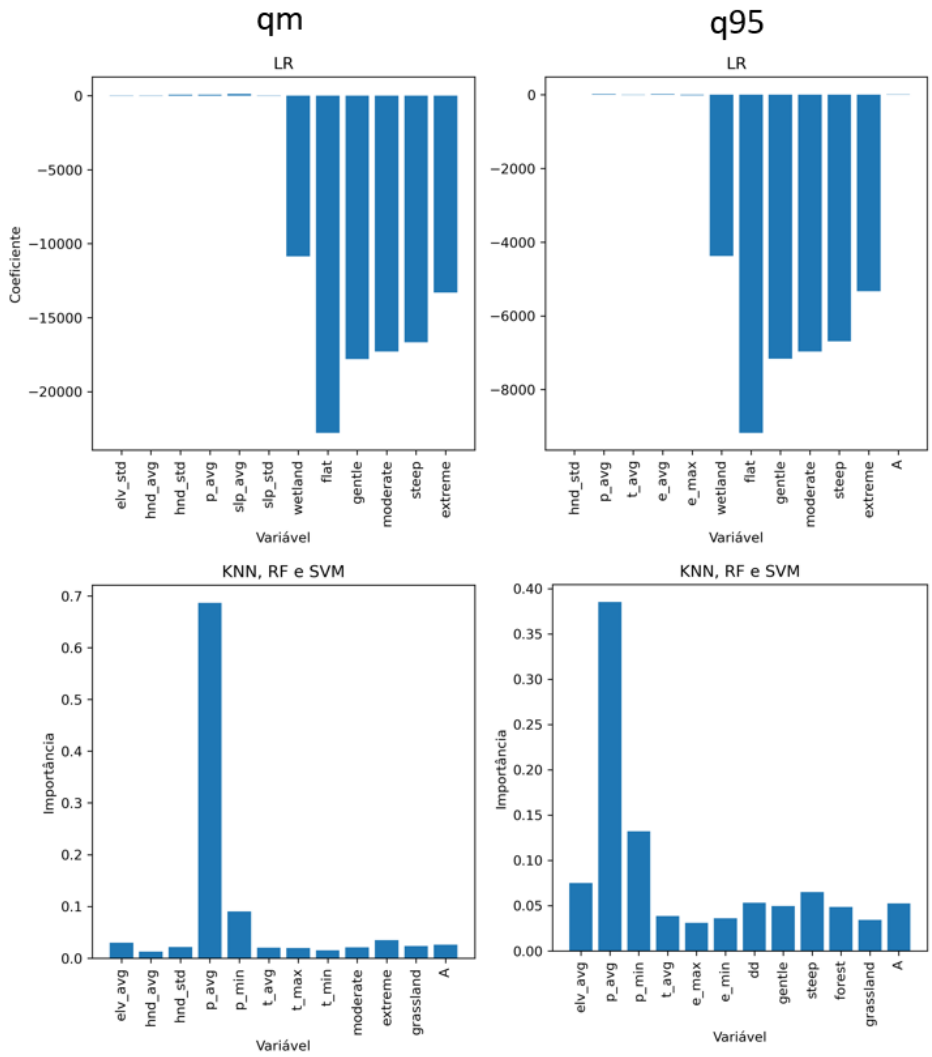

Figura 2. Coeficientes (LR) e importâncias (KNN, RF e SVM) das variáveis preditoras selecionadas para treinamento dos modelos de ML.

3.3 Performance dos modelos

Com relação a vazão média (Figura 3), todos os modelos apresentaram resultados satisfatórios. O modelo LR ficou com o resultado ligeiramente inferior aos outros (R^2 de *0.8* e RQ75 de *1.34*). O modelo RF apresentou os melhores resultados em termos de R^2 e RQ75 (*0.85* e *1.23*, respectivamente). Na *q95*, os modelos KNN e RF se mostraram ligeiramente superiores ao SVM e consideravelmente superiores ao LR, sem diferenças significativas entre os dois. Os valores de R^2 e RQ75 atingiram *0.765 ± 0.05* e *1.565 ± 0.05*, respectivamente, contra *0.71* e *1.7* no SVM e *0.51* e *2.11* no LR.

Os modelos KNN e RF, apesar de apresentarem os melhores resultados de modo geral, fazem previsões somente no domínio onde foram treinados, por isso não apresentaram resultados < 0, já os modelos LR e SVM permitem extrapolações, o que deve ser considerado se as características da bacia estiverem fora do domínio de treinamento. O modelo KNN ainda possui a vantagem computacional, com as rodadas tendo levado em torno de *1/20* do tempo que levou o RF (*30 s* contra *12 min*, aproximadamente). O modelo LR levou em torno de *5 s* e o SVM em torno de *2 min*). É também importante ressaltar que os hiperparâmetros dos modelos KNN, RF e SVM foram minimamente calibrados para esta aplicação, sendo usados valores em torno dos "valores padrão" do pacote scikit-learn (Python). O modelo LR não possui hiperparâmetros.

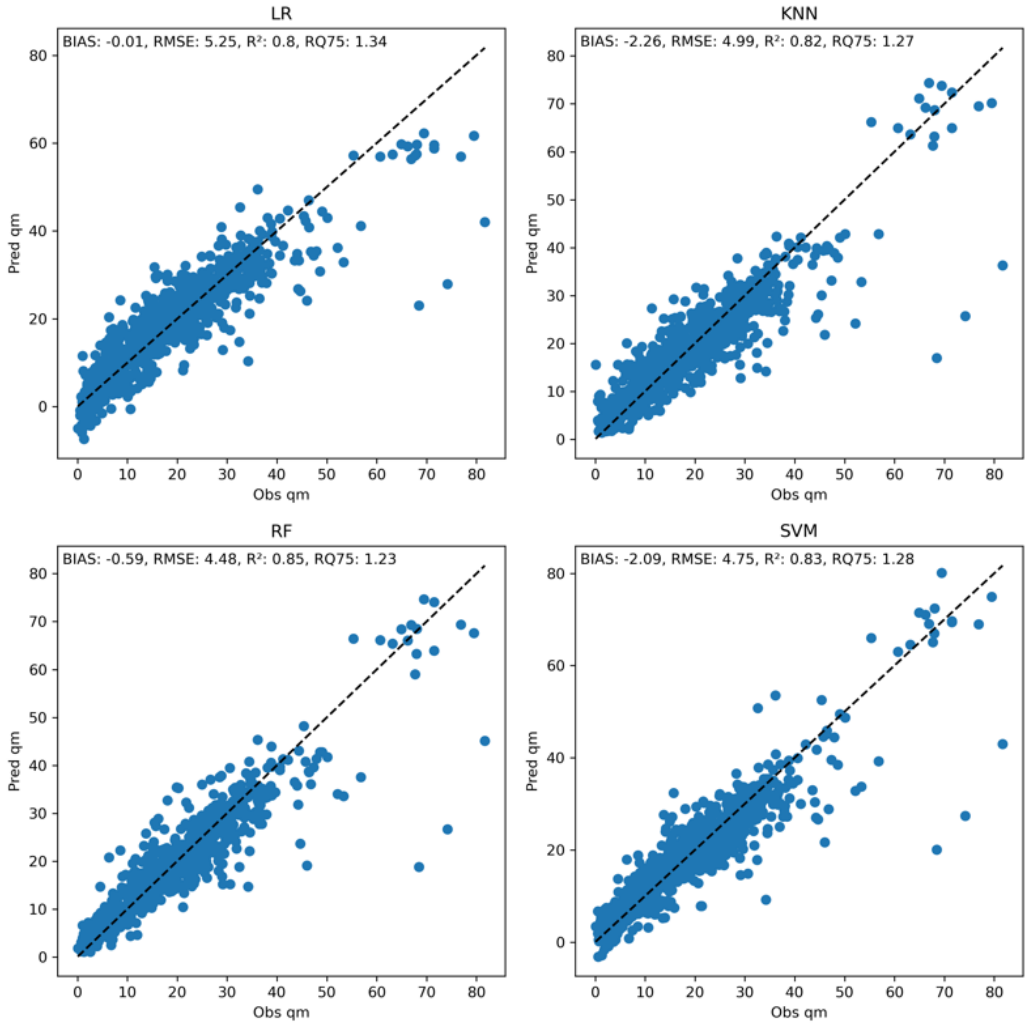

Figura 3. Resultados dos modelos de ML em relação a vazão média.

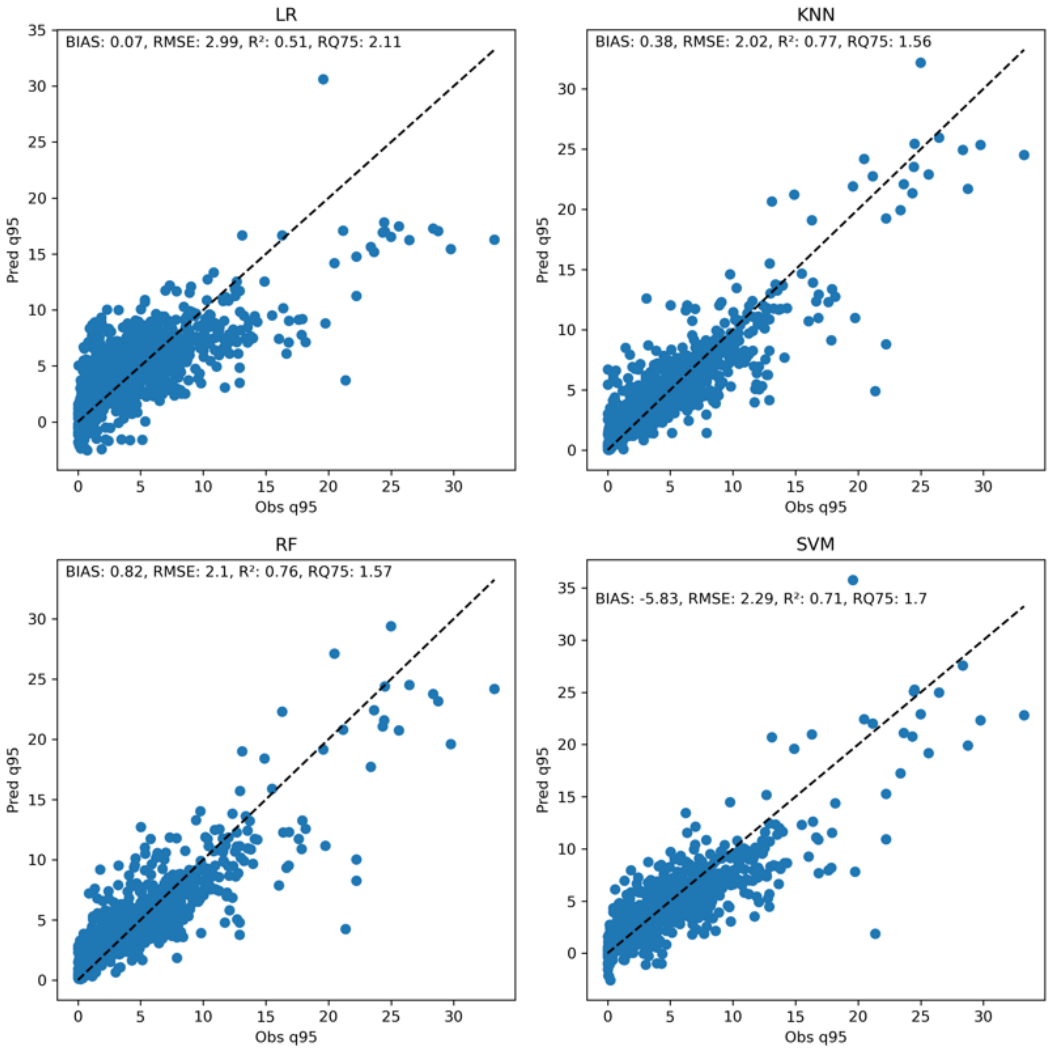

Figura 4. Resultados dos modelos de ML em relação à vazão com permanência de 95% do tempo.

4 CONCLUSÕES

Este relatório apresentou uma avaliação dos modelos de machine learning (ML) para a previsão de vazões de referência (*qm* e *q95*) em locais sem dados. O campo de estudo é relativamente novo, e possui muitas possibilidades de expansão nos próximos anos. Aqui foram exploradas apenas algumas aplicações diante de uma miríade de alternativas. Com relação a estudos anteriores, por exemplo, estes exploraram um número consideravelmente maior de variáveis preditoras (> 70) em regiõs consideravelmente menores (bacia hidrográfica, país pequeno). Não é de conhecimento dos autores alguma aplicação parecida na escala continental do território brasileiro, o que abre portas para este tipo de estudo.

As variáveis testadas foram cuidadosamente selecionadas com base no conhecimento conceitual dos autores, e a novidade desta aplicação foram as classes de solo baseadas em topografia, que se mostraram grandemente influentes em todos os modelos, principalmente na regressão linear. Dentre os modelos testados, a clássica regressão linear (LR) se mostrou competitiva para estimar as vazões médias, porém bem inferior aos outros modelos mais robustos na *q95*. De modo geral, os modelos KNN e RF apresentaram os melhores resultados em ambas vazões analisadas, com o KNN tendo vantagem computacional considerável. A desvantagem desses dois modelos é que eles não extrapolam os resultados para fora do domínio de aplicação. Dessa forma, nenhum modelo aqui testado deve ser descartado antes de uma avaliação cuidadosa da aplicação pretendida.

5 REFERÊNCIAS

Altman, N.S., 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, 46 (3), 175–185.

Awad, M. and Khanna, R., 2015. Support Vector Regression. *In*: M. Awad and R. Khanna, eds. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Berkeley, CA: Apress, 67–80.

Breiman, L., 2001. Random Forests. *Machine Learning*, 45 (1), 5–32.

Chen, X. and Jeong, J.C., 2007. Enhanced recursive feature elimination. *In*: *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*. Presented at the Sixth International Conference on Machine Learning and Applications (ICMLA 2007), 429–435.

Ferreira, R.G., Silva, D.D. da, Elesbon, A.A.A., Fernandes-Filho, E.I., Veloso, G.V., Fraga, M. de S., and Ferreira, L.B., 2021. Machine learning models for streamflow regionalization in a tropical watershed. *Journal of Environmental Management*, 280, 111713.

Friedl, Mark and Sulla-Menashe, Damien, 2019. MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006.

Gao, H., Hrachowitz, M., Fenicia, F., Gharari, S., and Savenije, H.H.G., 2014. Testing the realism of a topography-driven model (FLEX-Topo) in the nested catchments of the Upper Heihe, China. *Hydrology and Earth System Sciences*, 18 (5), 1895–1915.

Gharari, S., Hrachowitz, M., Fenicia, F., and Savenije, H.H.G., 2011. Hydrological landscape classification: investigating the performance of HAND based landscape classifications in a central European meso-scale catchment. *Hydrology and Earth System Sciences*, 15 (11), 3275–3291.

Golian, S., Murphy, C., and Meresa, H., 2021. Regionalization of hydrological models for flow estimation in ungauged catchments in Ireland. *Journal of Hydrology: Regional Studies*, 36, 100859.

Hou, A.Y., Kakar, R.K., Neeck, S., Azarbarzin, A.A., Kummerow, C.D., Kojima, M., Oki, R., Nakamura, K., and Iguchi, T., 2014. The Global Precipitation Measurement Mission. *Bulletin of the American Meteorological Society*, 95 (5), 701–722.

Muñoz Sabater, J., 2019. ERA5-Land monthly averaged data from 1981 to present [online]. Available from: https://cds.climate.copernicus.eu/cdsapp#!/dataset/10.24381/cds.68d2bb30?tab=overview [Accessed 19 Oct 2021].

Rennó, C.D., Nobre, A.D., Cuartas, L.A., Soares, J.V., Hodnett, M.G., Tomasella, J., and Waterloo, M.J., 2008. HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia. *Remote Sensing of Environment*, 112 (9), 3469–3481.

Rong, S. and Bao-wen, Z., 2018. The research of regression model in machine learning field. *MATEC Web of Conferences*, 176, 01033.

Savenije, H.H.G., 2010. *HESS Opinions* 'Topography driven conceptual modelling (FLEX-Topo)'. *Hydrology and Earth System Sciences*, 14 (12), 2681–2692.

Souza, C.M., Z. Shimbo, J., Rosa, M.R., Parente, L.L., A. Alencar, A., Rudorff, B.F.T., Hasenack, H., Matsumoto, M., G. Ferreira, L., Souza-Filho, P.W.M., de Oliveira, S.W., Rocha, W.F., Fonseca, A.V., Marques, C.B., Diniz, C.G., Costa, D., Monteiro, D., Rosa, E.R., Vélez-Martin, E., Weber, E.J., Lenti,

F.E.B., Paternost, F.F., Pareyn, F.G.C., Siqueira, J.V., Viera, J.L., Neto, L.C.F., Saraiva, M.M., Sales, M.H., Salgado, M.P.G., Vasconcelos, R., Galano, S., Mesquita, V.V., and Azevedo, T., 2020. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. *Remote Sensing*, 12 (17), 2735.

Tyralis, H., Papacharalampous, G., and Langousis, A., 2019. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. *Water*, 11 (5), 910.

Worland, S.C., Farmer, W.H., and Kiang, J.E., 2018. Improving predictions of hydrological low-flow indices in ungaged basins using machine learning. *Environmental Modelling & Software*, 101, 169–182.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J.C., Sampson, C.C., Kanae, S., and Bates, P.D., 2017. A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, 44 (11), 5844–5853.