REGIONALIZAÇÃO DE VAZÕES DE REFERÊNCIA EM SANTA CATARINA

Vinícius Bogo Portal Chagas [1] & Pedro Luiz Borges Chaffe [2]

Palavras-Chave – Estimativa de vazão, Santa Catarina, Aprendizado de máquina.

INTRODUÇÃO

Estimar a vazão de rios em locais sem medições é um desafio tradicional da hidrologia (Blöschl et al., 2013) apesar de ainda não ter sido inteiramente solucionado (Chagas et al., 2024), já que a maior parte dos rios não possui estações de monitoramento. Essas estimativas são essenciais para planejar o uso dos recursos hídricos ao embasar a concessão de outorgas, identificar locais com pouca água disponível e apontar as regiões mais suscetíveis às secas. Adicionalmente, a regionalização da vazão permite investigar quais variáveis melhor explicam a variabilidade espacial da vazão. Isto é especialmente útil para entender o contraste entre a variabilidade imposta pelo clima e a atenuação pelas dinâmicas de armazenamento de água no solo e no subsolo (e.g., Chagas et al., 2024), regulando o quanto da variação climática de fato é observada nos rios.

Apesar da regionalização de vazões ser um tema amplamente estudado, novas técnicas utilizando aprendizado de máquina ainda não foram comparadas afundo com as técnicas tradicionais no Brasil, levando ao potencial para explorar novas metodologias de estimativas. No Brasil, o Estado de Santa Catarina oferece excelentes condições para a construção e avaliação de técnicas de regionalização de vazão devido à grande densidade de postos fluviométricos e à baixa interferência humana no uso da água (particularmente reservatórios) comparado aos demais estados brasileiros (Chagas et al., 2020).

Os objetivos deste trabalho são: (*i*) construir modelos de estimativa de vazões de referência em bacias hidrográficas sem dados observados no Estado de Santa Catarina, incluindo ambos modelos tradicionais e modelos de aprendizado de máquina; (*ii*) comparar a performance dos modelos para identificar o modelo de melhor ajuste.

METODOLOGIA

A variável resposta é a vazão de referência Q98, isto é, a vazão igualada ou superada em 98% do tempo. A Q98 foi calculada a partir das séries diárias de postos fluviométricos (Chagas et al., 2020) de Santa Catarina e bacias vizinhas no PR e RS. As vazões foram naturalizadas a partir da sua soma com as demandas hídricas consultivas mensais estimadas pela ANA (ANA, 2019), removendo o efeito da retirada de água. As variáveis explicativas incluem índices climáticos (chuva e evapotranspiração médias e sazonais) e atributos fisiográficos da bacia baseados em Chagas et al. (2024): geologia (tipo de formação, condutividade hidráulica), solo (textura, profundidade e classe) e relevo (altitude, declividade). As vazões e a Q98 foram usadas em unidade de mm d^{-1}.

Foram construídos sete modelos de regionalização, dividido em quatro tipos. Modelos espaciais, que estimam a Q98 apenas a partir da posição das bacias (uniformidade espacial e krigagem). Modelos tradicionais, que combinam a identificação de regiões homogêneas com regressão linear. Modelos de

1) Departamento de Engenharia Sanitária e Ambiental, Universidade Federal de Santa Catarina. Email: vbchagas@gmail.com
2) Departamento de Engenharia Sanitária e Ambiental, Universidade Federal de Santa Catarina. Email: pedro.chaffe@ufsc.br

árvore com *ensemble,* que agregam várias árvores de decisão (random forest e *XGBoost*). Por fim, um modelo de kernel e margem (SVM).

A performance dos modelos foi estimada com validação cruzada aninhada (Filzmoser et al., 2009) de cinco dobras. Os hiperparâmetros foram ajustados nas dobras internas e a performance foi estimada nas dobras externas, evitando vazamento de dados. A otimização dos hiperparâmetros foi feita com otimização Bayesiana (Snoek et al., 2012), minimizando o RMSE. Tanto a modelagem quanto as métricas usaram a vazão em mm d^{-1}.

Os modelos espaciais estimam a Q98 a partir apenas da localização das bacias. O modelo de uniformidade espacial atribui a cada bacia a Q98 igual à estação mais próxima (modelo uniformidade espacial A) e à média da Q98 dos *n* postos mais próximos (modelo uniformidade espacial B), com *n* ajustado entre 1 e 10. A krigagem ordinária interpola a Q98 a partir de um variograma ajustado às coordenadas projetadas, sem outras variáveis explicativas (Pebesma, 2004).

Os modelos tradicionais dividem as bacias em regiões homogêneas e ajustam uma regressão linear em cada região (Blöschl et al., 2013). As regiões foram definidas por agrupamento das variáveis previamente selecionadas por regressão *stepwise*. Dentro de cada região, a Q98 foi modelada por regressão linear múltipla com seleção *stepwise*. Foram usadas duas variantes, que diferem nas variáveis usadas para formar as regiões, uma variante utilizando apenas variáveis fisiográficas (modelo regiões homogêneas A) e outra usando índices de vazão (modelo regiões homogêneas B).

Os modelos de árvore preveem a Q98 com *random forest* e *XGBoost*. O *random forest* agrega árvores treinadas em amostras aleatórias (*bagging*; Breiman, 2001), e o *XGBoost* ajusta árvores de forma sequencial (*boosting*; Chen & Guestrin, 2016). O SVM faz a regressão buscando a margem que melhor ajusta os dados, usando um kernel para captar relações não lineares (Cortes & Vapnik, 1995). Foi utilizado SVM do tipo radial.

RESULTADOS

A performance de teste de todos os modelos, estimada por validação cruzada aninhada, é apresentada na Fig. 1. Os maiores R^2 (coeficiente de determinação) foi dos modelos de árvore, entre 0,76 e 0,79, sendo a maior performance o *random forest*. Os modelos espaciais (uniformidade espacial e krigagem) tiveram performance intermediária, com R^2 variando entre 0,64 e 0,70. Os modelos tradicionais baseados em regiões homogêneas e regressão tiveram as menores performances, com R^2 abaixo de 0,60, além de terem as maiores instabilidades da performance (i.e., maior incerteza da performance).

As estimativas espaciais do modelo de melhor performance (*random forest*) são apresentadas na Fig. 2. O modelo identificou corretamente o principal padrão espacial da Q98, com valores maiores próximo à costa ambos nos dados observados como nos valores estimados (Fig. 2). A relação entre os valores previstos e observados mostra que não houve padrão espacial claro nos resíduos do modelo, revelando que não houve significativo viés espacial nas estimativas. O gráfico de dispersão entre o Q98 observado e previsto (Fig. 3) revela que o modelo superestima as Q98 mais baixas e subestima as Q98 mais altas.

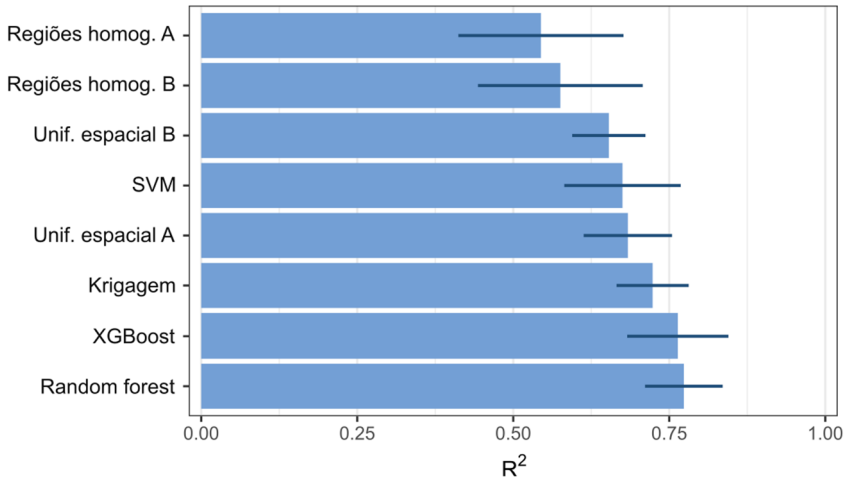


Figura 1. Performance dos modelos de estimativa da vazão de referência (Q98). As barras representam a performance média de teste das dobras da validação cruzada. As linhas indicam a variabilidade das performances (± 0,5 desvio padrão) entre as dobras.

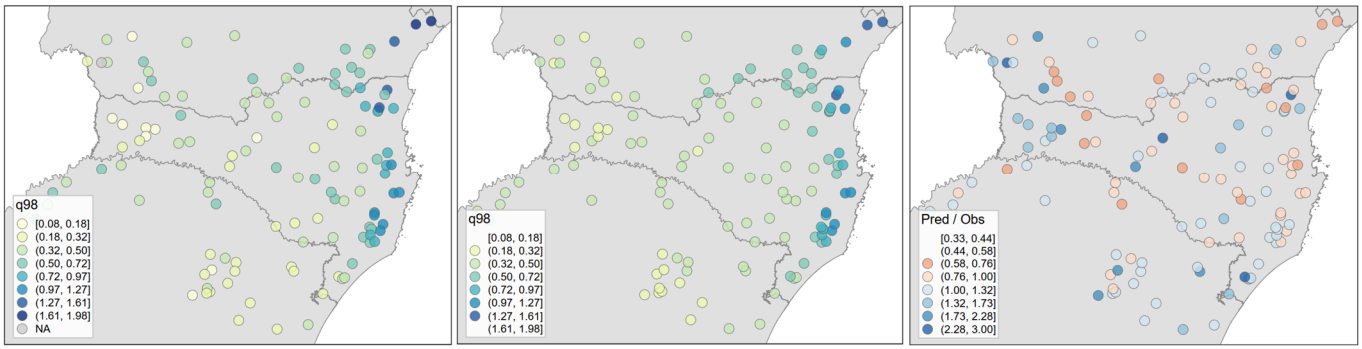


Figura 2. Vazão Q98 observada (a) e Q98 estimada (b) pelo random forest, ambos mm/d. Relação entre Q98 estimada e observada (c). As vazões estimadas são as médias de todas as dobras da validação cruzada (i.e., vazões estimadas de teste, não de treino). Cada círculo representa o exutório de uma bacia hidrográfica.

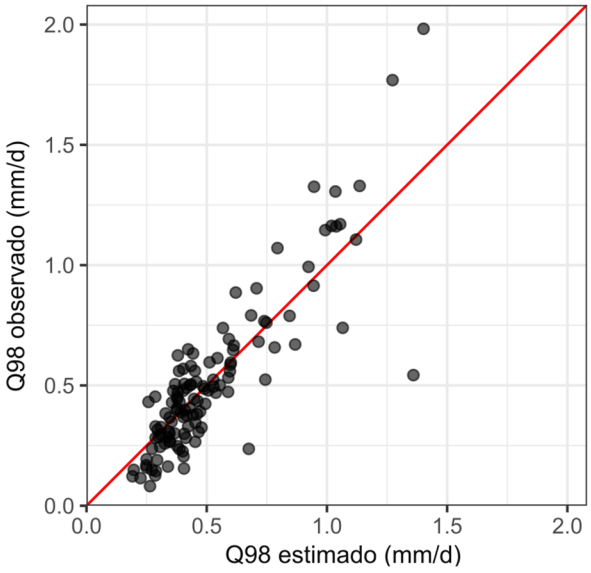


Figura 3. Vazão Q98 observada e estimada pelo random forest. As vazões estimadas são as médias de todas as dobras da validação cruzada (i.e., vazões estimadas de teste, não de treino). Cada círculo representa uma bacia hidrográfica.

CONCLUSÕES

Este trabalho construiu e comparou sete modelos de estimativa da vazão de referência Q98 em bacias hidrográficas de Santa Catarina, incluindo modelos tradicionais e de aprendizado de máquina. Os modelos de árvore com *ensemble*, em especial o *random forest*, tiveram as melhores performances de teste (R^2 até 0,79), superando os modelos espaciais e tradicionais. O *random forest* reproduziu bem o padrão espacial da Q98, sem viés espacial relevante, embora superestime as vazões mais baixas e subestime as mais altas. Os resultados mostram o potencial do aprendizado de máquina para a regionalização de vazões de referência no Brasil, em detrimento das técnicas tradicionais de regionalização.

REFERÊNCIAS

ANA (2019). *Manual De Usos Consuntivos Da Água No Brasil* (p. 75). Brasília.

Blöschl, G., Sivapalan, M., Savenije, H., Wagener, T., & Viglione, A. (2013). *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge University Press.

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

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., & Siqueira, V. A. (2020). CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil. *Earth System Science Data*, *12*(3), 2075–2096. https://doi.org/10.5194/essd-12-2075-2020

Chagas, V. B. P., Chaffe, P. L. B., & Blöschl, G. (2024). Regional Low Flow Hydrology: Model Development and Evaluation. *Water Resources Research*, *60*(2), e2023WR035063. https://doi.org/10.1029/2023WR035063

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Filzmoser, P., Liebmann, B., & Varmuza, K. (2009). Repeated double cross validation. *Journal of Chemometrics*, *23*(4), 160–171. https://doi.org/10.1002/cem.1225

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, *30*(7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012

Snoek, J., Larochelle, H., & Adams, R. P. (2012, August 29). Practical Bayesian Optimization of Machine Learning Algorithms. arXiv. https://doi.org/10.48550/arXiv.1206.2944

AGRADECIMENTOS

Os autores agradecem à Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC) pelo financiamento e apoio no Edital FAPESC n° 32/2024 (Termo de Outorga 2024TR002697 e 2024TR2698) e ao CNPq pelo apoio financeiro.