25th International Congress on Modelling and Simulation, Darwin, NT, Australia, 9 to 14 July 2023
mssanz.org.au/modsim2023

Resampling techniques for rare events prediction using data-driven and hydrological models

M. Zeinolabedini Rezaabad [a,b], **L. Marshall** [a,b,c] **and F. Johnson** [a,b]

[a] *Water Research Centre, School of Civil and Environmental Engineering, University of NSW, Australia*
[b] *ARC Training Centre DARE, Data Analytics for Resources and the Environment, Sydney, Australia*
[c] *Faculty of Science and Engineering, Macquarie University, North Ryde, NSW, Australia*
Email: m.zeinolabedini_rezaabad@unsw.edu.au

Abstract: Understanding rare or extreme events is an important part of water engineering because of their substantial impacts on natural and human systems. Examples of such rare/extreme events include floods, droughts and harmful algal blooms. It is crucial to detect and predict rare events with as much skill as possible to ensure that they are properly designed for and managed. When framed as a data science problem, rare events are associated with imbalanced datasets where the numbers of samples in each class of data are substantially different. The class with the smallest number of rare events (i.e., high impact) is called the minority class. Problems arise because the minority class may not have sufficient samples compared to the majority class for training a predictive model and as a result the model will not be able to detect rare events.

To increase the rare events importance for predictive model calibration, a group of methods called resampling techniques have been developed. Resampling techniques are data-level methods that balance the training dataset by either removing some samples from the majority class (under-sampling), adding some samples to the minority class (over-sampling) (Chawla et al. 2002), or combining under-sampling and over-sampling (hybrid-sampling). In this study, the influence of resampling methods was investigated for two case studies – a Bayesian Network (BN) used to predict increased algal activity and GR4J, with a focus on flood estimation. The two models are interesting case studies for resampling methods because of the differences in their model structures. For the BN, both model structure and the model parameters can be modified through the resampling. In the case of the hydrological model, only the model parameters are changed with the resampling implemented by modifying the objective function used for model calibration.

For the BN case study, the resampling methods were used in a model to predict chlorophyll-a (chl-a) in Lake Burragorang in New South Wales, Australia. The original data and balanced dataset were used for BN structures and parameter learning. The results showed that resampling techniques increased the ability of the BN in terms of both structure and parameter learning to detect events with higher probability of increased algal activity.

In the second case study, the resampling methods were applied to GR4J models for 168 of the Hydrologic Reference Stations across Australia with a focus on flows exceeding 99.8th percentile. Differences in the skill of the GR4J models calibrated using the original and balanced datasets were partially explained through catchment characteristics (i.e., area, base flow index, runoff-ratio, and runoff skewness). The results showed that resampling techniques increased extreme flow estimation in most catchments, especially those with lower baseflow index and higher runoff skewness.

Compared to original data, resampling techniques increased the ability of BN and GR4J models to detect rare/extreme events. Resampling techniques can be used for better understanding of the most effective parameters on rare/extreme events.

REFERENCES

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 16, 321–357.

Keywords: *Rare events, imbalanced data, over-sampling, under-sampling, Bayesian network, GR4J*