Temporal cross-validation in forecasting: A case study of COVID-19 incidence using wastewater data

Mallory Lai (Department of Mathematics and Statistics, University of Wyoming, Laramie, Wyoming, USA)

Shaun S. Wulff (Department of Mathematics and Statistics, University of Wyoming, Laramie, Wyoming, USA)

Yongtao Cao (Department of Mathematical and Computer Sciences, Indiana University of Pennsylvania, Indiana, USA)

Timothy J. Robinson (Department of Mathematics and Statistics, University of Wyoming, Laramie, Wyoming, USA)

Rasika Rajapaksha (Department of Computer Systems Engineering, University of Kelaniya, University Drive, Kelaniya, Colombo, Sri Lanka)

Journal – Quality and Reliability Engineering International

Online ISSN: 1099-1638

Print ISSN: 0748-8017

Article publication date: 11 November 2024

Abstract

Two predominant methodologies in forecasting temporal processes include traditional time series models and machine learning methods. This paper investigates the impact of time series cross-validation (TSCV) on both approaches in the context of a case study predicting the incidence of COVID-19 based on wastewater data. The TSCV framework outlined in the paper begins by engineering interpretable features hypothesized as potential predictors of COVID-19 incidence. Feature selection and hyperparameter tuning are then utilized with TSCV to identify the best features and hyperparameters for optimal model performance given a specific forecast horizon. While evidence supporting the utility of TSCV for auto-regressive integrated moving average model with exogenous variables (TS-ARIMAX) forecasts is lacking in this study, such an approach proves advantageous for gradient boosting machine forecasts (TS-GBM). In Wyoming, for instance, TS-GBM had a 34.9% improvement compared to naïve predictions, whereas GBM without TSCV only had a 15.6% improvement. However, TSCV also enhances interpretability for both TS-ARIMAX and TS-GBM models as this approach selects specific features, such as lagged values of COVID-19 cases, based on forecast performance and forecast length. Future research should work to

explore the influence of stationarity and model averaging on the performance of TSCV in forecasting applications.

Citation

Lai, Mallory, et al. "Temporal cross-validation in forecasting: A case study of COVID-19 incidence using wastewater data." Quality and Reliability Engineering International. https://doi.org/10.1002/qre.3686.

Publisher

Wiley