Assessment of random forest method to classify suspended solid and nutrient first flush in urban watersheds

C. Russo, B.Sc.[1], A. Gorgoglione, Ph.D.[2*], A. Castro, Ph.D.[2]

[1]*Politecnico di Milano, Milan, Italy*
[2]*Universidad de la República, Montevideo, Uruguay*
**Corresponding author email: agorgoglione@fing.edu.uy*

Highlights

- Random forest technique can be used to predict whether a rainfall event can generate pollutant first flush in urban watersheds.
- A ranking of rainfall characteristics in terms of their degree of importance in predicting pollutant first flush is generated with SHAP analysis.
- Average rainfall intensity is the most important variable in predicting TSS, TN, TP first flush.

Introduction

Worldwide, urbanization has led to an intensification of anthropogenic activities accompanied by an increase in impervious surfaces. During a precipitation event of particular duration and intensity, the first portion of the runoff contribution washes away pollutants accumulated on impervious surfaces during dry weather, generating runoff water that is more concentrated in pollutants (Di Modugno et al., 2015). The so-called first flush has been recognized and investigated as a typical phenomenon of urban areas since it represents one of the most critical non-point source pollutions and, therefore, can produce detrimental impacts on the quality of receiving water bodies (Gorgoglione et al. 2021).

The dynamic and random nature of urban runoff quality is demonstrated to be influenced by different variables. Numerous studies have been undertaken to quantify such relationships. However, a few works focused on ranking the influencing variables in terms of their degree of importance in generating first flush (Jeung et al., 2019; Perera et al., 2019).

Based on these considerations, the objective of this study is twofold: 1) developing a machine-learning algorithm, based on RF technique, able to predict whether a rainfall event can generate first flush, taking into account variables that characterize the precipitation event (dry and wet period), 2) ranking such variables in terms of their level of importance in predicting first flush.

The outcomes of this study are expected to contribute to the development of accurate and reliable stormwater-quality models and, consequently, effective stormwater treatment design.

Methodology

Study approach

A flowchart that summarizes the approach adopted in this study to accomplish the twofold objective is presented in Figure 1. Four main steps can be identified. The first one is represented by the dataset creation, including a monitoring campaign, a precipitation-generation model (IRP), and a hydrologic/hydraulic/water quality model (SWMM) (see “Study area and data gathering”). The second step includes the exploratory data analysis, carried out to discard from further analysis possible correlated variables. The third step presents the classification modeling (see “Feature-importance analysis and classification model”). Once the best model is obtained, we performed a feature (variable) importance analysis (see “Feature-importance analysis and classification model”).

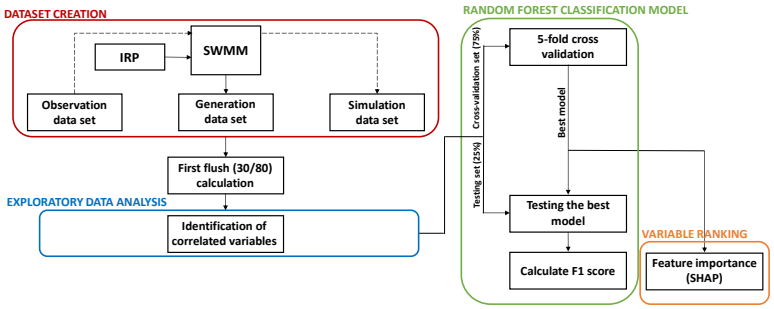

Figure 1. Overall procedure of the RF model classification and variable ranking.

	TSS	TN	TP
n estimators	2000	1000	1000
Min samples leaf	5	11	10
Max features	2	2	4
F1 score	**0.80**	**0.92**	**0.77**
Accuracy	**0.87**	**0.89**	**0.88**
F1 score $_{stratified}$	0.24	0.63	0.32
accuracy $_{stratified}$	0.60	0.50	0.67
F1 score $_{uniform}$	0.33	0.65	0.32
accuracy $_{uniform}$	0.48	0.57	0.48

Table 1. Best hyperparameters, F1 score, and accuracy (and baselines).

Study area and data gathering

The urban watershed considered for this study is located in the Puglia region (Southern Italy), in Sannicandro di Bari (SB). A thorough description of the study area, including catchment characteristics, drainage network, and climatic region, can be found in Gorgoglione et al. (2021).

The dataset adopted in this study includes three data sources: *i) Observations*: 5 events during which precipitation, streamflow, total suspended solids (TSS), total nitrogen (TN), and total phosphorus (TP) concentration were monitored (Di Modugno et al., 2015). *ii) Simulations*: 5 events characterized by observed precipitation and simulated flow rate and water-quality variables (TSS, TP, and TN). The simulations were obtained using the Storm Water Management Model (SWMM). *iii) Generations*: 567 events characterized by synthetic precipitation, produced by the Iterated Random Pulse (IRP) model (Veneziano and Iacobellis, 2002), and simulated flow rate and water-quality variables (TSS, TP, and TN) obtained using the SWMM model. In this case, the synthetic precipitation events were used as input of the SWMM model for generating hydrographs and pollutographs for each event at SB.

An in-depth description of the dataset collected can be found in Gorgoglione et al. (2021).

It is essential to highlight that SWMM and IRP are purely used here as data generators.

Feature-importance analysis and classification model

The RF classifier is the supervised learning algorithm adopted in this study to predict whether a rainfall event can generate first flush, taking into account rainfall characteristics (dry and wet period). The original data was normalized (*sklearn.preprocessing* package). The dataset was split into two subsets, 75% and 25% of the data, respectively for cross-validation and testing processes. The Optuna framework was adopted for hyperparameter tuning and optimization (Akiba et al., 2019). To prevent overfitting, 5-fold cross-validation was performed. F1 score was the loss function used to calculate model performance and select the best model (with the best hyperparameters) from the cross-validation process that then will be validated with the testing data subset. With the purpose of comparing our model performance with others, we computed the accuracy (ratio between true positives and total observations).

The Python class *sklearn.ensemble.RandomForestClassifier* was used.

SHapley Addictive exPlanation (SHAP) was adopted to carry out the feature-importance analysis (Lundberg and Lee, 2017). It is based on the game theoretically optimal Shapley Values. The SHAP objective is to explain the prediction of an instance by computing the contribution of each feature to the prediction. SHAP is therefore a technique for estimating the expected marginal contribution of a factor among all possible contributions. To run this analysis, the *SHAP* python package was used.

Results and discussion

First flush classification

The variables considered for each event were: antecedent dry period (ADP) [days]; total rainfall (TR) [mm]; runoff volume (RV) [l]; event duration (D) [minutes], average rainfall intensity (I_{AVG}) [mm/h]; maximum rainfall intensity (I_{MAX}) [mm/h]; event mean concentration (EMC_{TSS}, EMC_{TN}, EMC_{TP}) [kg/l]; event mean load (EML_{TSS}, EML_{TN}, EML_{TP}) [kg]. From the exploratory data analysis, we found a strong correlation between TR and RV (r=0.99); therefore, RV was excluded from further analysis.

First flush occurrence was computed for all the rainfall events for the three pollutants, based on the 30/80 definition (Bertrand-Krajewski et al., 1998). This information represents the ground truth of the three different RF models developed to predict whether an event generates TSS, TN, and TP first flush. A data matrix (5 × 577) was the input for each of the three cross-validation processes, where 5 are the rainfall variables (ADP, TR, D, I_{AVG}, I_{MAX}) and 577 are the precipitation events. Optuna was configured to search for the best set of hyperparameters: number of estimators (trees) (*n estimators*), minimum number of samples in a leaf (*min samples leaf*), and maximum number of features per estimator (*max features*). The objective function was to maximize the F1 score. For this purpose, we performed 500 experiments with early stopping of 100 runs. The three best hyperparameter configurations (one per pollutant) with the corresponding F1 score and accuracy are shown in Table 1. For the three pollutants, F1 score and accuracy are higher than those obtained if the stratified and uniform predictors were used (baselines). It is worth remarking that our TSS model accuracy is higher than Perera et al. (2019) one by 16% (0.872 and 0.714 respectively).

Feature importance

For TSS, the two most important variables are I_{AVG} and I_{MAX}. This is in accordance with Perera et al. (2019) and can be justified by the fact that the highest the rainfall intensity is, the greater the TSS runoff concentration. For TN, D and I_{AVG} are the two most important variables, according to Jeung et al. (2019). As well as for TP, I_{AVG} and D are the features with the highest importance for predicting the occurrence or non-occurrence of first flush, in accordance with Jeung et al. (2019). It is worth highlighting that TP and TN are characterized by the same most important features and that I_{AVG} represents the variable that we always have to consider for predicting the existence of first flush in urban areas.

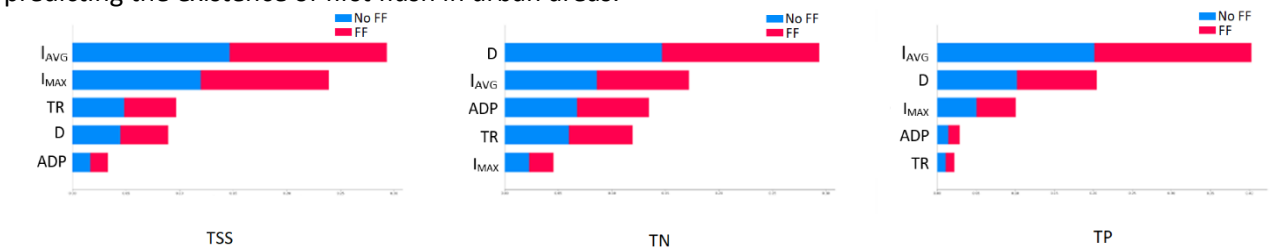

Figure 2. SHAP values to explain the prediction of a first flash event for the three pollutants.

Conclusions

This study provides an RF model able to successfully predict the occurrence of first flush for TSS, TN, and TP in urban watersheds (F1 score average=0.83). The proposed model improved the state of the art (accuracy improvement=16%). I_{AVG} represents the variable that always has to be considered for predicting the existence of first flush in urban areas.

References

Bertrand-Krajewski, J., Chebbo, G. and Saget, A. (1998). Distribution of pollutant mass vs volume in stormwater discharges and the first flush phenomenon. Water Res. 32 (8), 2341–2356.

Di Modugno, M., Gioia, A., Gorgoglione, A., Iacobellis, V., la Forgia, G., Piccinni, A.F. and Ranieri, E. (2015) Build-up/wash-off monitoring and assessment for sustainable management of first flush in an urban area. Sustainability, 7, 5050–5070.

Gorgoglione, A., Castro, A., Iacobellis, V. and Gioia, A. (2021). A Comparison of Linear and Non-Linear Machine Learning Techniques (PCA and SOM) for Characterizing Urban Nutrient Runoff. Sustainability, 13, 2054.

Jeung, M., Beak, S., Boem, J., Cho, K.H., Her, Y. and Yoon, K. (2019). Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments. J. Hydrology, 575, 1099-1110.

Lundberg, S.M. and Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in neural information processing systems. pp. 4765–4774.

Perera, T., McGree, J., Egodawatta, P., Jinadasa, K. and Goonetilleke, A. (2019). Taxonomy of influential factors for predicting pollutant first flush in urban stormwater runoff. Water Research, 199, 115075.

Veneziano, D. and Iacobellis, V. (2002). Multiscaling pulse representation of temporal rainfall. Water Resour. Res., 38, 131–1313.

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD.