Leveraging Machine Learning to Understand Green Stormwater Infrastructure Performance Risks

Sizhe Zhang, Achira Amur, Emma Olson, Peleg Kremer, Xun Jiao, Virginia Smith, Bridget Wadzuk
Villanova University

***Abstract*—This investigation examines green stormwater infrastructure (GSI) performance data and environmental features through machine learning to identify drivers in the built environment that impact the performance and sustainability of GSI.**

I. Introduction

The effects of climate change are exacerbating environmental challenges caused by urban growth. As changing rainfall patterns collide with urban growth there is an increased need for new stormwater management solutions. Green stormwater infrastructure (GSI) is a strategy used to manage stormwater and offset the impacts of urban development on stormwater. However, municipal stormwater programs are scientifically limited by a lack of integrated data and associated analytics to quantify GSI performance and expected life-cycle dynamics. Understanding GSI maintenance requirements and how they differ by GSI type is not well studied, although maintenance is costly. Machine learning is a possible solution as there are many observations across disciplines impacting the science, engineering, and policy of GSI [1], [2], [3]. Using data across disciplines, an in-depth analysis can be conducted to understand the performance of GSI holistically and the spatial factors that impact performance over time.

II. Background

A. Current State of Data Availibility and Use

Data used for GSI performance analysis is a sparse high variance dataset of point measurements [4]. The hydrologic, hydraulic, and environmental complexities of GSI are not well understood. Due to the relatively new nature of GSI, which is often implemented and regulated by municipalities with limited budgets, there is minimal research on the intricate relationship among GSI function, impact, and sustainability, particularly under varying conditions, spatial areas (e.g., beyond single watersheds), and time scales (i.e., years to decades). This lack of complete research is impeding the scientific understanding of GSI dynamics and a move towards data-driven policy [5].

B. Machine Learning

Machine learning (ML) has achieved recent success across various fields, including flood prediction and susceptibility mapping. This knowledge, coupled with advancements in the empirical and theoretical understanding of hydrology and hydraulic modeling and design provides an opportunity for advancement in urban stormwater [6], [7].

III. Methodology

The Philadelphia Water Department (PWD) has established multiple GSI within the city to manage stormwater and reduce combined sewer overflows [8]. PWD conducts regular GSI inspections to ensure functionality and inform maintenance. The current analysis uses PWD maintenance data for three types of GSIs: Basins, Bioinfiltration systems, and Porous Pavements. Our approach is to analyze and predict the overall rating of GSI systems using various aspects of inspection data. First, we perform data preprocessing and cleaning. Subsequently, we analyze the correlation between each inspection data type and the final GSI performance rating. Finally, we employ various ML methods to assess the predictability of using assessment components to ascertain the overall GSI rating.

A. Data Preparation and Processing

Initially, we process the collected data by filling in missing or abnormal values with NULL and homogenized the data formats. The inspection data has different formats of data, including discrete data representing the contextual description of the inspection (e.g., weather and environmental conditions) and individually rated parameters pertaining to the components of the GSI (e.g., inlet, outlet, etc.). Each of the parameters contain a rating from 1-4, where a GSI functional rating of 1 or 2 indicates the GSI passed the inspection while a GSI functional rating or 3 or 4 indicates the GSI did not pass the inspection. To facilitate processing of these diverse data types, we label encode all data and treat them as discrete data. This choice stems from: (1) some data labels represent degrees, but the descriptions are not precise measurements, only estimations and (2) the complexity of data collection results in numerous missing values that are challenging to handle as continuous data. Consequently, we apply label encoding to all data, allowing missing values to be encoded as individual labels representing value absence.

B. Correlation Analysis of Inspection Data

To identify inspection data types with the greatest impact on the GSI rating, we use Entropy Correlation, Mutual Information, and Linear Discriminant Analysis correlation to analyze the discrete data [9]. These methods can analyze the correlation between two discrete variables. We conducted a correlation analysis for each class of inspection data and the final GSI rating in each dataset, obtaining a preliminary understanding of which data classes have a more substantial impact on the GSI rating.

C. Employed Machine Learning Techniques

We also experimented with training and predicting using multiple ML algorithms. Given that there is low data volume and the data is not continuous signals or images, we selected eight traditional ML algorithms. We use Decision Tree, Random Forest, Naive Bayes, SVM, CatBoost, LGBM, XGBoost, and Gradient Boosting for training and testing the datasets. We utilize the k-fold cross-validation technique on the datasets, dividing them into k subsets, training models on k-1 subsets, and validating models on the remaining subset [10]. K-fold cross-validation more efficiently uses dataset information, enhancing model generalization ability, mitigating overfitting risk, and providing a more accurate evaluation of model performance.

IV. Experimental Results

A. Correlation Analysis

We assessed the correlation between each class and the overall GSI rating (Table 1 is an example of the "Porous Pavement" dataset). While different calculation methods yield varied correlation values, the top three features consistently exhibit high correlations. In contrast, the bottom three features display relatively low correlations across all methods. These findings indicate that the top three features have a more substantial impact on the porous pavement GSI ratings. We analyzed the highest correlating features in the Basin and Bioinfiltration datasets. In both datasets, features related to the debris within the drainage area, the condition of the drainage area and the condition of the outflow structure show high correlations.

Table 1. Feature Ranking

Features	Entropy Correlation	Mutual Information	Linear Discriminant
Trash Debris Sediments	0.36	0.52	0.63
Structural Defects	0.32	0.47	0.60
Surface Clogged	0.44	0.65	0.66
Ponding	0.07	0.10	0.17
Permeable Pavement	0.08	0.12	0.19
Clogging Rating	0.07	0.10	0.16

B. Machine Learning Model Performance

We trained and tested the performance of eight ML algorithms on the three datasets. Across the datasets, Random Forest and various boosting methods (i.e., CatBoost, LGBM, XGBoost, and Gradient Boosting) had superior performance (Table 2). Although these algorithms have some differences, they all achieve an accuracy rate over 80% or close to it for the datasets. Based on these results, we believe that these traditional ML techniques can effectively predict GSI ratings.

Table 2. Performance of machine learning algorithms

Method	Dataset		
	Basin	***Bioinfiltration***	***Porous***
Decision Tree	79.27%	83.61%	74.93%
Random Forest	82.84%	84.59%	77.04%
Naive Bayes	61.52%	74.75%	40.96%
SVM	77.45%	80.16%	75.20%
CatBoost	83.07%	85.08%	76.78%
LGBM	82.91%	85.25%	75.99%
XGBoost	82.28%	85.41%	77.57%
Gradient Boosting	82.99%	84.92%	78.10%

V. Conclusion

This research uses ML to understand features impacting GSI performance, which can inform GSI inspection to ultimate reduce maintenance and improve GSI sustainability. The results from the ML analysis show that the models are able to represent the inspection ratings, highlighting the parameters that have a substantial impact on the overall rating of the system, allowing for preventative maintenance to ensure long-term functionality. The outcome of this project will be rapid gains in GSI sustainability knowledge that will strengthen core scientific understanding of GSI processes and urban environments, which will be widely applicable across academia and private and public industry. Building on this analysis, we propose to develop ML-based models that can forecast the GSI performance ratings for a given set of GSI data including GSI design parameters, environmental parameters, and social-economic data.

Acknowledgment

This research was funded by the National Science Foundation (Award Number 2228035).

References

[1] D. G. Tarboton, K. a. T. Schreuders, D. W. Watson, and M. E. Baker, “Generalized terrain-based flow analysis of digital elevation models,” p. 2000, 2009.

[2] K. Lehnert, S. Carbotte, V. Ferrini, W. Ryan, R. Arko, and S.-L. Chan, “IEDA: Integrated Earth Data Applications to Support Access, Attribution, Analysis, and Preservation of Observational Data from the Ocean, Earth, and Polar Sciences,” vol. 13, no. 13439, p. 1, 2011.

[3] I. Zaslavsky, T. Whitenack, M. Williams, D. Tarboton, K. Schreuders, and A. Aufdenkampe, “The Initial Design of Data Sharing Infrastructure for the Critical Zone Observatory,” p. 6, 2011.

[4] B. Wadzuk, B. Gile, V. Smith, A. Ebrahimian, M. Strauss, and R. Traver, “Moving Toward Dynamic and Data-Driven GSI Maintenance,” *J. Sustain. Water Built Environ.*, vol. 7, no. 4, p. 02521003, Nov. 2021, doi: 10.1061/JSWBAY.0000958.

[5] B. Wadzuk, B. Gile, V. Smith, A. Ebrahimian, and R. Traver, “Call for a Dynamic Approach to GSI Maintenance,” *J. Sustain. Water Built Environ.*, vol. 7, no. 2, p. 02521001, 2021, doi: 10.1061/JSWBAY.0000945.

[6] M. A. A. Mehedi, V. Smith, H. Hosseiny, and X. Jiao, “Unraveling The Complexities of Urban Flood Hydraulics Through AI,” In Review, preprint, May 2022. doi: 10.21203/rs.3.rs-1602023/v1.

[7] H. Hosseiny, F. Nazari, V. Smith, and C. Nataraj, “A Framework for Modeling Flood Depth Using a Hybrid of Hydraulics and Machine Learning,” *Sci. Rep.*,vol. 10, no. 1, May 2020, doi: 10.1038/s41598-020-65232-5.

[8] “Green City Clean Waters – Philadelphia Water Department.” https://water.phila.gov/green-city/ (accessed May 14, 2023).

[9] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto. and S. Ridella, “The ‘k’ in k-fold cross validation,” *ESANN*, pp441-446, Nov. 2022.