Risk Analysis, Vol. 41, No. 12, 2021 DOI: 10.1111/risa.13742

A Two-Stage Data-Driven Spatiotemporal Analysis to Predict Failure Risk of Urban Sewer Systems Leveraging Machine Learning Algorithms

John E. Fontecha,[1] **Puneet Agarwal**,[1] **María N. Torres**,[2] **Sayanti Mukherjee**,[1,*] **Jose L. Walteros**,[1] **and Juan P. Rodríguez**[3]

Risk-informed asset management is key to maintaining optimal performance and efficiency of urban sewer systems. Although sewer system failures are spatiotemporal in nature, previous studies analyzed failure risk from a unidimensional aspect (either spatial or temporal), not accounting for bidimensional spatiotemporal complexities. This is owing to the insufficiency of good-quality data, which ultimately leads to under-/overestimation of failure risk. Here, we propose a generalized methodology/framework to facilitate a robust spatiotemporal analysis of urban sewer system failure risk, overcoming the intrinsic challenges of data imperfections—e.g., missing data, outliers, and imbalanced information. The framework includes a two-stage data-driven modeling technique that efficiently models the highly right-skewed sewer system failure data to predict the failure risk, leveraging a bidimensional space-time approach. We implemented our analysis for Bogotá, the capital city of Colombia. We train, test, and validate a battery of machine learning algorithms—logistic regression, decision trees, random forests, and XGBoost—and select the best model in terms of goodness-of-fit and predictive accuracy. Finally, we illustrate the applicability of the framework in planning/scheduling sewer system maintenance operations using state-of-the-art optimization techniques. Our proposed framework can help stakeholders to analyze the failure-risk models' performance under different discrimination thresholds, and provide managerial insights on the model's adequate spatial resolution and appropriateness of decentralized management for sewer system maintenance.

KEY WORDS: Infrastructure failure risk prediction; machine learning models; maintenance planning; predictive and prescriptive modeling; spatiotemporal analysis; urban sewer system

1. INTRODUCTION

Conventional sewer systems are vital components of the urban infrastructure that collect and dispose storm and waste water. Given their important role in sanitation and disease control, maintaining the sewer infrastructure in optimal working conditions is of paramount importance for the physical and economic well-being of any society (Duchesne, Beardsell, Villeneuve, Toumbou, & Bouchard, 2013). Among the different challenges that affect the operation of such infrastructure systems, aging and

[1]Department of Industrial and Systems Engineering, University at Buffalo, Buffalo, NY, USA.
[2]Department of Structural, Civil and Environmental Engineering, University at Buffalo, Buffalo, NY, USA.
[3]Department of Civil and Environmental Engineering, Universidad de los Andes, Bogotá, Colombia.
*Address correspondence to Sayanti Mukherjee, Department of Industrial and Systems Engineering, University at Buffalo, 411 Bell Hall, Buffalo, NY 14260; sayantim@buffalo.edu

deterioration processes of their components are the principal factors that lead to failures, for which the high cost is associated with service disruptions, adverse publicity, as well as health and safety problems (Rodríguez, McIntyre, Díaz-Granados, & Maksimović, 2012). To ensure the proper functionality of sewer systems, water authorities are responsible for allocating resources and schedule routine inspections for maintenance and rehabilitation, keeping in mind strict time schedules and budget constraints. While these tasks are already challenging given the size and complex nature of the underground infrastructure, they get further complicated when considering the uncertainties of the future environment, such as changes in population size, land use, and climate patterns (Kleidorfer et al., 2013; Shortridge & Camp, 2019).

One of the primary tasks that support the strategic management of sewer systems is the effective vulnerability assessment of the infrastructure through the prediction of failure risks (i.e., knowing *a-piori* the likelihood of where and when a failure may occur given the current conditions of the system). To conduct this complex task, most previous studies have typically used either physics-based or statistics-based models as their core predictive engines. Physics-based models are generally composed of computer simulations based on hydrodynamic and structural models that analyze the system's usage and estimate expected times between consecutive failures (Montes, Vanegas, Kapelan, Berardi, & Saldarriaga, 2020; Rodríguez, McIntyre, Díaz-Granados, & Maksimović, 2012; Santos, Amado, Coelho, & Leitão, 2017). On the other hand, statistics-based models are based on machine learning and statistical concepts and are used to predict future system failures by studying the probability of occurrences of such events in the past (Roehrdanz, Feraud, Lee, Means, Snyder, & Holden, 2017; Salman and Salem, 2012a). In general, because of the complexity of simulating the sewer system deterioration processes, incorporating a large list of influencing factors (e.g., the physical properties of pipelines, land and environmental characteristics, and interactions of the system with other urban infrastructure), the popularity of physics-based models has decreased over the past years (Fontecha et al., 2016; Torres, Rodríguez, & Leitao, 2017), while the statistical models are gaining significant attention (Ana et al., 2009).

As with most predictive tools, the accuracy of these statistics-based models relies heavily on the quality of input data (i.e., whether or not the datasets used to design such models are comprehensive, accurate, and complete). For the particular case of sewer systems, despite their importance, datasets with such desirable characteristics are rarely available. There are several reasons why the absence of comprehensive datasets of sewer systems is a prevalent problem. First, because of their underground nature, physical access to the sewer systems is generally limited, which renders the inspection activities to be capital-intensive and often disruptive (e.g., temporarily cutting service lines and stopping urban traffic to access the system through manholes) (Fontecha et al., 2020). Second, time and budget constraints limit the inspection capabilities of water utilities as they can only investigate a small section of the infrastructure at a time, rather than inspecting the entire sewer system serving a wide region, city, county, or state. Previous studies have recommended an overall turnover time for a full sewer system inspection, ideally, to be between 2 and 25 years depending on the size, condition/state, and age of the system (McDonald & Zhao, 2001); nevertheless, municipalities often report these times to be doubled or even tripled under real circumstances (Allouche & Freure, 2002; López-Kleine, Hernández, & Torres, 2016).

Despite these data limitations, statistics-based data-driven research is still a valuable resource to study the key factors contributing to higher deterioration rates (Ana et al., 2009), as well as identifying the infrastructure components that have higher risks of failure (Younis & Knight, 2010). With the advent of new low-cost sensing and data storage techniques (Duran, Althoefer, & Seneviratne, 2002) and the increasing use of interconnected platforms (Sirkiä et al., 2017), valuable data collected from multiple sources are rapidly becoming available (Tscheikner-Gratl et al., 2019). This major progress in data accessibility can empower new models with valuable information that can be used to mitigate some of the aforementioned limitations, leading to more accurate predictions, and therefore, to a better sewer system vulnerability assessment, facilitating risk-informed investment decisions, and optimal resource allocation/management (Fontecha et al., 2020).

From the data quality perspective, previous research studies focusing on the statistical analysis of sewer systems have handled issues of missing data, outliers, and imbalanced information in a vague or indirect manner. In the presence of such data anomalies, most authors practice the removal of outliers and observations with missing values and often disregard the fact that most real datasets are typically

populated with imbalanced data (Harvey & McBean, 2014). We argue that these atypical observations should not be removed since they may provide interesting information about extreme failure events and, if removed, their occurrence may not be properly captured by the statistical models. Moreover, as it is widely accepted by most researchers, imbalanced data must be handled meticulously because it tend to bias the prediction algorithms toward the frequently occurring trend (e.g., not failure) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Harvey and McBean, 2014; Krawczyk, 2016), resulting in models that may underestimate the critical but less frequently occurring failure events, and thus, significantly underestimating the failure risks.

Although data-driven techniques, such as *statistical learning* and *machine learning*, have gained significant traction to predict sewer system failures, another critical factor that may hinder the prediction quality of these methodologies is the lack of an adequate spatiotemporal analysis in the current state-of-practice. Despite the fact that failures in a sewer system infrastructure occur in a spatiotemporal context (i.e., a pipe is more prone to fail if it has failed in the past, has not been replaced, or if its neighboring pipes have also failed), most previous works have considered explanatory variables from an either temporal or spatial perspective, ignoring the high interdependence of these two key dimensions (Duchesne et al., 2013; López-Kleine et al., 2016; Soriano-Pulido, Valencia-Arboleda, & Rodríguez Sánchez, 2019; Younis & Knight, 2010). One of the main reasons why both the spatial and temporal dimensions are typically considered in an isolated way is the absence of comprehensive datasets that provide descriptive information directly regarding these two aspects. Therefore, to perform an adequate spatiotemporal study, it is important to first conduct a meticulous analysis of multiple data sources in order to *mine* and *extract* data that can reflect changes in both time and space dimensions. Following the initial exploration introduced in Korving, Van Noortwijk, Van Gelder, and Clemens (2009), to the best of our knowledge, this manuscript is one of the first attempts to incorporate explanatory variables that covary with space and time simultaneously for sewage systems failure risk analysis.

To address the above-mentioned limitations of the existing methodologies and approaches, we propose a novel two-stage data-driven methodology that predicts the risk of sewer system failures using spatiotemporal features and considering intrinsic data imperfections such as imbalanced data, missing values, and outliers. We test the performance of our methodology with a case study that analyzes the sewer system of the city of Bogotá, Colombia. Given the size and complexity of Bogotá's sewer system (a system serving close to 8 million people), we provide evidence that our framework can be used to efficiently predict the failure risk of large-scale systems with limited availability of spatiotemporal variables and comprehensive datasets (common characteristic observed in large cities' infrastructure systems). In doing so, we train, test, and validate a battery of statistical learning models such as logistic regression (LR), decision trees (DTs), random forests (RFs), and extreme gradient boosting, and select the best-performing methods using a bias-variance trade-off approach (Mukherjee & Nateghi, 2019). Finally, we provide an example of how the results from the proposed data-driven risk assessment predictive model can be used to inform planning and scheduling of maintenance operations of the urban sewer system infrastructure.

The rest of the article is organized as follows: Section 2 describes the collection and analysis of previous studies serving three purposes: (1) To highlight the importance of the machine/statistical learning models in the area of sewer systems failure risk prediction, (2) inform the gap in knowledge regarding spatiotemporal explanatory variables, and (3) present the gap related to the use of methods to handle data anomalies. Section 3 describes in detail the two-stage methodology, from the management of the spatial data to the application of the predictive models. Section 4 presents the case study, details the type of variables used in our study, explains the preprocessing steps to create the final database, and describes the specific settings for the prediction models. Section 5 summarizes and analyzes the results of the application of our methodology to the specific case study presented in this article, discussing the predictive performance of the models and the top influential variables identified by our selected model. Section 6 presents a discussion on the managerial insights relevant to our methodology; more specifically, a simple optimization exercise that leverages the outcomes of our failure risk prediction model as its inputs is presented. The goal of the optimization exercises is to perform the optimal planning and scheduling of the maintenance operations based on the predicted risks. Finally, in Section 7, we

summarize our research findings and present the implications of our work.

2. LITERATURE REVIEW

We begin this section by providing an overview of the existing methods in failure risk assessment of sewer system infrastructure, focusing our discussion on descriptive and predictive models. We continue the section by presenting a general discussion of some of the challenges that emerge from the management of data anomalies, identifying various gaps and potential opportunities for improvement. We then conclude the section by summarizing different ways in which some of these issues can be addressed leveraging our proposed methodology.

2.1. Statistics-Based Models for Predicting Sewer System Failures

2.1.1. Descriptive Models

The methodologies used to describe the failure mechanisms of sewer systems can be classified into three major areas: (1) traditional risk-matrix-based, (2) data-driven, and (3) probabilistic approaches.

Under the category of risk-matrix-based approaches, we reviewed a series of studies that used quantitative methods to analyze the sewer deterioration process and estimate potential failures. For example, Korving et al. (2009) proposed a risk-based model for the economic optimization of in-sewer storage. In this article, the authors consider several properties of the system, such as its dimensions, storage and pumping capacities, as well as other uncertain exogenous components like spatial and temporal variations in rainfall. The authors used several cost functions to capture the environmental and economic impacts with respect to the in-sewer storage design and rehabilitation decisions. Salman and Salem (2012b) studied the risk of failure of sewer pipes by modeling the probability and consequences-of-failure values using three different methods: simple multiplication, risk matrices, and fuzzy inference models. They used the resulting assessments to generate sewer risk maps to assist water management agencies to identify sewer-pipe sections that require immediate attention. Similarly, Kuliczkowska (2016) analyzed the risk of structural failures of sewer pipes due to internal corrosion using a risk-matrix generated by categorizing the structural failure probabilities and the associated consequences caused by sewer failures.

As for the area of data-driven approaches, several studies have used this type of models to analyze the different factors responsible for sewer system failures. For example, Ana et al. (2009), Ugarelli, Kristensen, Røstum, Sægrov, and Di Federico (2009), and Younis and Knight (2010) used regression analysis techniques for the selection of important system features that have a strong impact on sewer deterioration. In particular, Ana et al. (2009) proposed a backward stepwise regression approach to systematically drop insignificant variables in an iterative manner to improve the predicting capabilities of their model. Ugarelli et al. (2009) used an evolutionary polynomial regression (EPR) model to identify the critical attributes of sewer pipelines that have a significant influence on the number of blockages that may accumulate in a given time period. Younis and Knight (2010) proposed a generalized linear model (GLM) to estimate the deterioration behavior of reinforced concrete and vitrified clay pipes. To this end, the authors developed an ordinal regression model based on cumulative logits that considers the interaction effect between the different explanatory variables. The model estimates the probabilities for wastewater pipelines of being into one of five internal condition grades.

In addition to regression analysis techniques, other data-driven methods have also been used to identify relevant variables to model the system's deterioration. For example, López-Kleine et al. (2016) used principal components analysis (PCA) coupled with k-means clustering to determine the relationship between structural characteristics of sewer pipes and their deterioration states. The relationships that were detected in this study helped the authors to identify the variables with a strong influence on the state of pipelines. Carvalho, Amado, Brito, Coelho, and Leitão (2018) applied three different variable selection algorithms, namely: the mutual information indicator, the out-of-bag samples concept (based on RF algorithms), and the stepwise search method, for identifying the variables that most significantly influence the quality of sewer failure predictions.

In the area of probabilistic approaches, previous papers focused on studying several random factors that contribute to the sewer system deterioration process. Studies such as Micevski, Kuczera, and Coombes (2002), Korving and Van Noortwijk (2008), Jin and Mukherjee (2010), and Rodríguez et al. (2012) used a combination of stochastic

methods and statistical tests to model different performance metrics like the time between consecutive system failures. Micevski et al. (2002) used a homogeneous Markov chain to analyze the structural deterioration of stormwater pipelines. The authors applied a Bayesian approach to calibrate the parameters of the model, and chi-square tests to identify the significant factors responsible for sewer pipe deterioration. Korving and Van Noortwijk (2008) used a statistical model for assessing the sewer conditions based on a combination of expert knowledge and sewer inspections. The model was updated with data from the inspections using Bayesian statistics. Additionally, a Dirichlet distribution was used to model prior knowledge on condition-state probabilities. Jin and Mukherjee (2010) analyzed the interarrival time between sewer blockages using different statistical tests such as the Kolmogorov–Smirnov test to investigate any statistically significant differences between the available subsets of data, and the Anderson–Darling test to compare and select the best distribution for modeling blockage interarrival times. Rodríguez et al. (2012) used homogeneous and nonhomogeneous Poisson processes to model the interarrival time between sewer blockages as a function of several system properties, and Post, Pothof, ten Veldhuis, Langeveld, and Clemens (2016) used statistical tests to further investigate whether sewer failures can be modeled as a homogeneous Poisson process.

In addition to the aforementioned traditional stochastic models, numerous studies such as Duchesne et al. (2013) and Egger et al. (2013) used mathematical techniques to solve problems inherent to typical failures in sewer systems. For example, Duchesne et al. (2013) developed a model based on survival analysis principles to assess the overall structural state of a sewer network. In contrast to most Markov chain models, this approach allows for the possibility of modeling transitions from multiple deterioration stages at each time step. Egger et al. (2013) proposed a combined sewer deterioration and rehabilitation model that is able to mitigate the challenges behind the lack of historical records of sewer conditions.

A careful analysis of the extant literature indicates that the methodologies used to conduct descriptive analytics on sewer systems have some disadvantages. In particular, risk-matrix-based approaches may exhibit an inherent subjectivity introduced by the decision-maker during the assignment of the weights and scores of the input factors responsible for sewer failures. Such subjective assessments can potentially result in loss of information, thus hindering the overall predictive power of the models. Similarly, most approaches that use probabilistic models are often calibrated for the specific conditions of the systems being considered in their respective studies and are often difficult to adapt for analyzing other systems. Without introducing a major refactoring, such models may be invalid for other systems featuring entirely different failure mechanisms and risk conditions. As for the case of existing data-driven models, despite being effective at identifying failure risk factors, these models are not necessarily designed for forecasting sewer failures as a function of the identified risk factors. In this article, we provide evidence that some of these limitations can be mitigated by utilizing a combination of statistical and machine learning methods; which, extracting complex patterns from historical data, provide reliable future estimates of sewer failures without relying on subjective assessments by experts.

2.1.2. Predictive Models

In this section, we provide a discussion on the existing predictive models, both parametric and nonparametric, which have been used to estimate the risk of sewer system failures.

Parametric approaches are model-based methods that assume a predefined functional form of the response defined by a set of parameters (Hastie, Tibshirani, & Friedman, 2009; James, Witten, Hastie, & Tibshirani, 2013). This category of models includes regression analysis and stochastic modeling. Studies that have utilized regression analysis to study sewer system failures include Chughtai and Zayed (2008), Younis and Knight (2010), Salman and Salem (2012a), and Roehrdanz et al. (2017). In particular, Chughtai and Zayed (2008) used multiple linear regression to predict the structural deterioration of sewer systems. These models were created independently for different materials (concrete, asbestos, cement, and polyvinyl chloride [PVC] pipes), while a generalized operational condition model was developed considering all the materials together. The data used in this study included general pipeline inventory records, AutoCAD drawings, and closed-circuit television (CCTV) inspection reports. Salman and Salem (2012a) estimated the probability of failure for sewer system sections using ordinal regression, multinomial LR, and binary LR. The authors also compared the estimated condition ratings with the observed data and found that the binary LR model provided the

most accurate results. Similarly, Roehrdanz et al. (2017) computed the probability of an exfiltrating defect occurrence within each pipeline section using the multifactor binary LR introduced in Salman and Salem (2012a).

In the domain of parametric models, past studies have also utilized stochastic modeling to predict failure conditions in sewer systems. For example, Le Gat (2008) modeled the deterioration of urban drainage infrastructure using a mixed multistate stochastic process represented by a nonhomogeneous Markov Chain. In this representation, the authors considered the transition probabilities between states not only to be time-dependent but also conditioned on a set of covariates and pipeline-specific random frailty factors. Soriano-Pulido et al. (2019) proposed a framework using a log-Gaussian Cox process to assess the impact of the spatiotemporal correlation between sediment-related blockages. This framework was based on factors associated with the spatiotemporal clustering of failures, which are characterized by physical properties of the sewer system or characteristics of the location where the failure event occurs, and temporal properties such as the precipitation data. The principal advantage of using these frameworks is to provide a flexible and relatively tractable class of empirical models for effectively describing spatially and temporally correlated phenomena.

Unlike parametric models, nonparametric models neither make strong assumptions about the input parameters nor the form of the mapping function. These flexibilities allow these models to freely learn any functional form from the training data. Nonparametric models include mainly tree and network-based models and have been widely used over the past few years. In the context of tree-based models, Harvey and McBean (2014) used the information from CCTV inspections to predict individual pipeline conditions of the sewer system. In this study, the training data were fed into a RF model designed to inform proactive maintenance policies. Baah, Dubey, Harvey, and McBean (2015) computed the risk of pipeline failure by estimating the condition grade of sewer pipes using probability values provided by Harvey and McBean (2014), and then, determining the consequence of failure (CoF) as the weighted average of the performance scores in terms of several impact factors (e.g. pipe size and roadway type). The condition grade of the pipelines was calculated by training an RF algorithm using a dataset consisting of 138 bad pipes and 1,117 good pipes, reaching an overall accuracy of 72%. In a subsequent study, Baah et al. (2015) provided a map incorporating the risk of sewer pipe failures and the conditions of failures using ArcGIS. In addition to the studies that used tree-based models, Mashford, Marlow, Tran, and May (2011) used a support vector machine model to predict the sewer condition over a sample dataset of 1,441 observations. The observations were randomly divided into a training set (75%) and a test set (25%), and the performance of the model was evaluated using several metrics such as the overall success rate, misclassification rate, goodness of fit, and agreement test.

In the area of network-based models, Bayesian belief networks (BBNs) and artificial neural networks (ANNs) are some of the most commonly used methods to make predictions about the condition of underground infrastructure (such as sewer pipelines). These models can be used to handle the relationships between the parameters and factors that affect the deterioration process while considering uncertainties for risk and consequence analysis. The use of BBN is appropriate when the data available for analysis are scarce and incomplete, while ANN can be used when a considerable amount of historical data is available. In case of missing data, BBN models benefit from experts' knowledge and technical literature, while ANN models can provide insights into cause–effect relationships and uncertainties through learning from the data (Kabir, Balek, & Tesfamariam, 2018).

Furthermore, Hahn, Palmer, Merrill, and Lukas (2002), Anbari, Tabesh, and Roozbahani (2017), and Kabir et al. (2018) developed a Bayesian network to predict CoF, likelihood of failure, and the need for inspection. Hahn et al. (2002) used six parameters to obtain the likelihood of failure (i.e., structural defects, interior corrosion, exterior corrosion, erosion, infiltration, and operational defects) and employed two mechanisms to predict the CoF (i.e., socioeconomic impacts and reconstruction impacts). The knowledge base was evaluated with a series of case studies and the proposed methodology was found to be effective at mimicking the knowledge of experts. Anbari et al. (2017) proposed a risk assessment model based on BBN to prioritize sewer pipe inspections. In this model, the risk of a sewer pipe failure was obtained from the integration of probability and CoF values using a fuzzy inference system (FIS). On the other hand, Tran, Ng, and Perera (2007) provided an example of the use of ANN for the prediction of serviceability condition deterioration of the sewer

system. The serviceability condition of the pipes was associated with the reductions in pipe diameter.

In the search for efficient models and tools to predict the physical condition of underground sewer infrastructure, studies such as Sousa, Matos, and Matias (2014), Jiang, Keller, Bond, and Yuan (2016), Santos et al. (2017), Caradot et al. (2018), and Hernández, Caradot, Sonnenberg, Rouault, and Torres (2018) were aimed to compare a collection of different models, and identifying the ones that produced the best results under several conditions. Additionally, Laakso, Kokkonen, Mellin, and Vahala (2018) and Elmasry, Hawari, and Zayed (2017) coupled different models as a part of a single framework with the idea of combining the predictive capabilities of such models in a single tool. Sousa et al. (2014) compared ANN, support vector machines (SVM), and LR to classify the sewer system into sections requiring immediate intervention and sections that are not expected to fail in the near future. Jiang et al. (2016) compared linear regression and ANN to identify the initiation time for corrosion and the corrosion rate after its initial detection in the concrete sewer pipes. Santos et al. (2017) used five different stochastic prediction models with the objective of identifying a tool that performed well with respect to prediction accuracy and robustness. Their findings showed that the nonhomogeneous Poisson process provided the best prediction results, while the performance of DTs based on RF had a better performance for cases with short-term prediction window. Caradot et al. (2018) developed a set of metrics to assess and compare the performance of RF and Markov chains. These models were used to predict three levels of sewer condition: good, medium, and bad. Hernández et al. (2018) evaluated two different models' predictive outcomes, namely, LR and RF, for two different case studies, a city in Europe and a city in South America. The models were used to predict the critical structural condition of sewer pipes in both cities on a four-level scale.

With respect to the integration of models, Laakso et al. (2018) proposed an LR approach and compared its performance with the corresponding RF model. This framework was also coupled with the Boruta method that is a variable selection algorithm for detecting relevant explanatory variables. Elmasry et al. (2017) developed a BBN and then integrated it to multinomial LR in order to transform the static BBN into a dynamic Bayesian network.

Table 1 summarizes the classification of the methodologies found in this literature review.

2.2. Challenges Related to Data-Driven Research

We identified several challenges and issues related to the data used in data-driven research, and categorized them into three fronts: (1) the type of information used for the explanatory variables; (2) the potential issues identified by authors regarding missing data, outliers, and imbalanced datasets; and (3) the dimension (temporal, spatial, and spatiotemporal) of the input data used to explain the response variables. Table 2 provides a detailed summary of the literature from the perspective of the various types of features or explanatory variables used and the data issues identified in the previous research studies.

2.2.1. Type of Explanatory Variables Used

While identifying the explanatory variables used to assess and predict the sewer system condition, we found four major sets of explanatory variables generally used in previous research studies: (1) physical-condition-related factors, 2) environmental factors, 3) demographic factors, and 4) variables related to other infrastructure or elements of the urban landscape. In the case of physical-condition-related variables, the most common variables are found to be related to the pipelines' characteristics (e.g., the diameter, length, slope, age, material, depth, type of effluent, and shape). Besides the physical characteristics of the infrastructure, the second most important variables used in previous studies correspond to interaction-based features. In this regard, we refer to the variables that are usually related to the urban landscape such as roads or trees (e.g., road types, road proximity, traffic volume, tree density, tree types, and tree proximity). The studies that considered these types of variables account for more than half of the papers reviewed, and it is summarized in Table 2. Weather and environment-related variables are not as common as the previously mentioned variables. This is an important shortcoming of the previous studies as it is well established that the management and assessment of infrastructure risk usually rely on climate-related historical data (Shortridge & Camp, 2019). The most common variables in this category are soil type, humidity, temperature, and precipitation. The type of variables that are most rarely considered in the previous studies includes the demographics and land use type of the region where the sewer system infrastructure is located. In this category, we identified variables related to land use, population density, district, zone, location, and

Table 1. Summary of Literature Review: Methodologies

Reference	Descriptive				Predictive					
	Risk-Matrix	Data-Driven		Probabilistic and Stochastic	Nonparametric		Parametric			Other
		Regression	Other		Tree	Network	Regression	Stochastic Process		
Korving et al. (2009)	✓									
Salman and Salem (2012b)	✓									
Kuliczkowska (2016)	✓									
Ana et al. (2009)		✓								
Ugarelli et al. (2009)		✓								
Younis and Knight (2010)		✓								
López-Kleine et al. (2016)			✓							
Carvalho et al. (2018)			✓							
Micevski et al. (2002)				✓						
Korving and Van Noortwijk (2008)				✓						
Jin and Mukherjee (2010)				✓						
Rodríguez et al. (2012)				✓						
Duchesne et al. (2013)				✓						
Egger et al. (2013)				✓						
Post et al. (2016)				✓						
Harvey and McBean (2014)					✓					
Baah et al. (2015)					✓					
Hahn et al. (2002)						✓				
Anbari et al. (2017)						✓				
Kabir et al. (2018)						✓				
Tran et al. (2007)						✓				
Chughtai and Zayed (2008)							✓			
Salman and Salem (2012a)							✓			
Roehrdanz et al. (2017)							✓			
Le Gat (2008)								✓		
Soriano-Pulido et al. (2019)								✓		
Mashford et al. (2011)										✓
Hernández et al. (2018)					✓		✓			
Sousa et al. (2014)						✓	✓			✓
Jiang et al. (2016)						✓	✓			
Santos et al. (2017)								✓		
Caradot et al. (2018)					✓			✓		
Laakso et al. (2018)					✓		✓			
Elmasry et al. (2017)						✓	✓			
Proposed approach	✓		✓		✓		✓			✓

The methodologies can be classified as descriptive or predictive. Each category can be split into subcategories based on the modeling tools used.

Table 2. Summary of Literature Review: Data Dimension

Reference	Features				Data Issues		
	Physical	Environmental	Demographic	Interaction	Missing	Outliers	Imbalanced
Kuliczkowska (2016)	✓	✓	✓	✓			
Laakso et al. (2018)	✓	✓		✓	✓	✓	
Tran et al. (2007)	✓	✓		✓	✓		
Mashford et al. (2011)	✓	✓		✓	✓		
Soriano-Pulido et al. (2019)	✓	✓		✓			
Micevski et al. (2002)	✓	✓					
Le Gat (2008)	✓						
Jiang et al. (2016)	✓	✓					
Harvey and McBean (2014)	✓		✓	✓	✓		✓
Baah et al. (2015)	✓		✓	✓			✓
Kabir et al. (2018)	✓		✓	✓			
Caradot et al. (2018)	✓			✓	✓		
Chughtai and Zayed (2008)	✓			✓		✓	
Salman and Salem (2012a)	✓			✓			✓
Elmasry et al. (2017)	✓			✓			
Anbari et al. (2017)	✓			✓			
Roehrdanz et al. (2017)	✓			✓			
Hernández et al. (2018)	✓			✓			
Santos et al. (2017)	✓				✓	✓	
Duchesne et al. (2013)	✓				✓		✓
Sousa et al. (2014)	✓				✓		
Ugarelli et al. (2009)	✓				✓		
Carvalho et al. (2018)	✓					✓	
López-Kleine et al. (2016)	✓					✓	
Rodríguez et al. (2012)	✓						
Younis and Knight (2010)	✓						
Ana et al. (2009)	✓						
Hahn et al. (2002)	✓						
Salman and Salem (2012b)	✓		✓	✓			
Post et al. (2016)					✓	✓	
Korving and Van Noortwijk (2008)							
Jin and Mukherjee (2010)							
Korving et al. (2009)	✓	✓					
Egger et al. (2013)							
Proposed approach	✓	✓	✓	✓	✓	✓	✓

the proximity to some specific city landmarks such as hospitals and schools. Finally, it can be observed from Table 2 that most of the studies do not consider more than two categories as their explanatory variables, the only exception being the study conducted by Kuliczkowska (2016). A summary of the distribution of these types of variables as observed from the literature can be found in Table 2.

2.2.2. Data Issues

Three common data issues that need to be considered include: (1) missing data, (2) outliers, and (3) imbalanced datasets. Note that missing information is generally related to the explanatory variables (i.e., the **X** variables), while outliers and imbalanced data are usually related to the target variable (i.e., the **y** variable).

Perhaps, the most common problem that previous research studies have faced while working with the infrastructure condition and maintenance data is the missing data. Table 2 shows a summary of the reviewed studies that deal with missing data. It is surprising that most of the studies that were reviewed do not mention the presence of missing data in their corresponding datasets or even if it is mentioned, most of the researchers simply opted for removing the corresponding observations from the original dataset without any further considerations, perhaps leading to a loss of valuable information (e.g., Caradot et al., 2018; Harvey & McBean, 2014; Post et al., 2016; Santos et al., 2017; Sousa et al., 2014; Tran et al., 2007;

Ugarelli et al., 2009). Only a few of the aforementioned studies explained in detail the data imputation process. For example, Duchesne et al. (2013) discussed that the installation date of pipes, which was not available in the dataset, was estimated based on the age of surrounding infrastructure; Laakso et al. (2018) used RF imputation to face the problem of missing data; and Egger et al. (2013) introduced a sewer system deterioration model to deal with the missing historical records.

Similar to the missing data, in the case of the outliers, only a few research studies mentioned their importance and presence in the dataset (see Table 2). In most of the studies, the anomalous observations were simply removed from the datasets (e.g., Carvalho et al., 2018; Laakso et al., 2018; Santos et al., 2017). In some of the studies, outliers were identified and removed after a further analysis. For example, in studies by Chughtai and Zayed (2008) and López-Kleine et al. (2016), atypical observations were marked based on normal probability plots and box plots, respectively, and then were removed from the final dataset.

In addition to missing data and outliers, the distribution of failure counts associated with the sewer systems is generally imbalanced (Harvey & McBean, 2014). However, in most of the literature reviewed in this article, authors mostly fail to account for the issues regarding the imbalanced data and even if they consider it, they do not discuss any specific procedures used to handle this common problem. For example, Baah et al. (2015) only mentioned that the data used were imbalanced, but provide no further information on how they handle such an issue, leading the reader to assume that they followed a similar strategy used by Harvey and McBean (2014). A few studies, such as Harvey and McBean (2014) and Duchesne et al. (2013), addressed this issue by combining multiple risk classes into a small number of classes.

2.2.3. Dimensionality of the Data

Our comprehensive literature review indicated that most of the research studies considered variables in the dimension of space and time in an isolated way (i.e., they did not consider variables defined over more than one dimension). Only the study conducted by Soriano-Pulido et al. (2019) described a spatiotemporal analysis of the failures in the sewer system infrastructure. Several authors, for example, Ugarelli et al. (2009), Rodríguez et al. (2012), and Soriano-Pulido et al. (2019) presented the target variable (**y**) as a spatiotemporal variable, while Korving et al. (2009) described the explanatory variable such as the rainfall as a spatiotemporal variable, in their studies, respectively. However, the unique work that claims to analyze the risk of failures from a spatiotemporal dimension is by Soriano-Pulido et al. (2019). Despite this effort, the authors fail to acknowledge that the only spatiotemporal variable in their model is the response variable (i.e., the occurrence of failures), while the other explanatory variables (**X**) are either spatial or temporal.

2.3. Insights and Gaps Identified from the Current Body of Knowledge

We reviewed an extensive pool of the literature to present the various types of state-of-the-art methodologies used to evaluate the risk of sewer system failures (see Table 1) and summarized the various types of data issues identified in the previous studies (see Table 2). As evident from the reviewed literature, prediction algorithms based on statistical and machine learning techniques have gained wide attention from researchers for developing tools and frameworks that support proactive maintenance of urban sewer systems. In addition, to accurately identify trends and patterns prevalent in historical data, these algorithms prove to be quite effective in providing critical insights regarding the risks and inherent uncertainties associated with the sewer system deterioration process. Despite the advances, there are several challenges related to the quality of data that need to be efficiently tackled and addressed. Thus, we develop a failure risk model for an urban sewer system infrastructure by leveraging several machine learning algorithms based on different mathematical approaches (both parametric such as regression-based and nonparametric such as tree-based methods). By offering a better prediction performance as well as facilitating the identification and evaluation of the various risk factors, our proposed model can help the water utilities in the efficient and proactive maintenance of urban sewer system infrastructure. In addition, we also present some generalized strategies to handle the various data issues as mentioned in previous sections. Finally, unlike previous studies, we propose a framework that can efficiently account for the multidimensionality of the (spatiotemporal) explanatory variables (**X**).

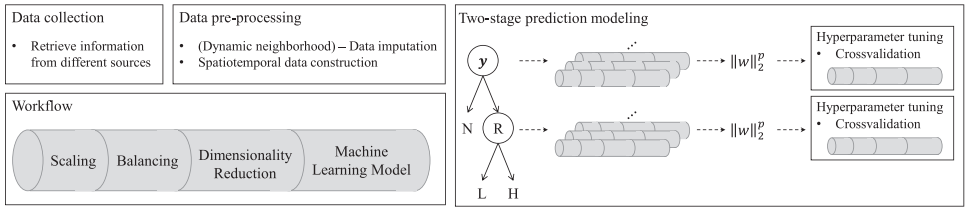


Fig 1. Overview of the two-stage methodology. First, collect data. Then, process the data correcting data issues and constructing spatiotemporal information. Finally, run the workflows in the different stages to define the hyperparameters for the machine learning models. Evaluate the performance of specific workflows through the use of norm 2 of the seek performance metrics.

3. METHODOLOGY

This section describes our proposed two-stage data-driven risk-prediction framework leveraging state-of-the-art machine learning algorithms and considering spatiotemporal information. We discuss the data preprocessing step focusing on resolving the various data issues as described before (e.g., presence of outliers, missing data), and then we describe the research methodology. We also present ideas behind the process of exploiting dimensionality transformation of data features (i.e., transforming single-dimensional raw data, mainly defined individually over the time or space dimensions) into two-dimensional spatiotemporal data to match the dimension of explanatory and response variables, and therefore, performing a robust analysis. Fig. 1 presents an overview of the proposed methodology.

3.1. Data Preprocessing

The data preprocessing part can be divided into two steps, namely: data issues management and the spatiotemporal transformation of explanatory variables. As mentioned in the literature review, some of the typical issues when working with sewer systems' databases include the presence of missing data, outliers, and imbalanced observations. Generally, the former is related to the explanatory variables $\mathbf{X}$, while the latter two are related to the target variable $\mathbf{y}$.

3.1.1. Data Issues Management

To address the missing data issue, we employ a methodology that uses information from the neighborhood of the missing observation to impute its corresponding value. Such a neighborhood can be defined and selected dynamically depending on the information available within the neighborhood that will make the missing data imputation possible. For example, if a neighborhood of a specific size (initially selected) fails to provide enough information to impute the missing data, the proposed algorithm will iteratively increase the size of the neighborhood, based on certain problem-dependent rules, until the targeted data imputation is achieved. This procedure can be applied to impute missing data on any of the variables in the dataset $\mathbf{X}$, by defining the neighborhood from a spatial and/or temporal closeness perspective.

In this study, our rationale for utilizing a neighborhood data imputation approach is based on the fact that there is a higher likelihood that similar types of infrastructure are installed around the same timeline (Duchesne et al., 2013) (i.e., construction timeline of the sewer systems) in adjacent locations, thus sharing similar structural characteristics between them. Furthermore, because of the underground nature of sewer systems infrastructure, renovations and new installations are both costly and disruptive. Considering the economies of scale behind the construction process, major updates and renovations to sewer systems are typically conducted when there are enough adjacent components to be added/replaced (e.g., pipelines, manholes, and gullypots) to justify the costs. Therefore, unless several adjacent data points are missing, it is unlikely for individual components to differ greatly in physical condition from their neighbors.

Inspired by k-nearest neighbors technique (Faisal & Tutz, 2017), as well as by scattered data interpolation methods (Franke, 1982; Shepard, 1968), we propose a new neighborhood-based approach by defining such a neighborhood through spatial variables, where the neighborhood is dynamically increased in size at each iteration. Here, such variables correspond to the physical location of the infrastructure distributed over space, but in other contexts, it can be defined using other specific sets of variables, for example, the neighborhood can be

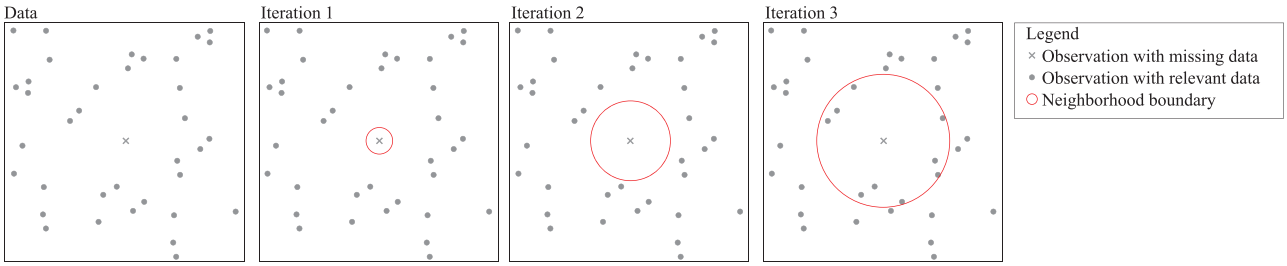


Fig 2. Three iterations defining the neighborhood for the data imputation. From iteration 1 to 3, the neighborhood is increasing. Iterations 1 and 2 do not have observations in the neighborhood; therefore, no data can be imputed. Iteration 3 provides observations within the neighborhood and specific imputation techniques can be applied.

defined over time dimension rather than the space dimension. Fig. 2 depicts the iteration process for the data imputation. It is noteworthy that within the boundary of a neighborhood, any specific method for data imputation as those described in studies by Allison (2002); Yuan (2010) and Soley-Bori (2013) can be applied (e.g., mean, median, linear regression, and maximum likelihood).

Importantly, we note that the inference quality of this method depends on several factors like the stopping criteria for growing the boundary size of the neighborhood. Clearly, no general framework exists to identify a one-size-fits-all optimal value for these parameters, as those are strongly dependent on the specific type of application being modeled, the integrity of the datasets, and the expected model accuracy as needed for informed decision-making. For example, while imputing missing data about a sewer system in a densely populated urban area, it is likely that the amount of infrastructure that exists in a relatively small neighborhood conveys sufficient information for a proper estimation of missing data, whereas in a less populated rural area, a much larger neighborhood may be required to produce similar results. We will provide the step-by-step process of implementation of this method, with an application to our case study, in Section 4.

From the data quality perspective, outliers represent a key challenge that needs to be addressed during the data-processing step. In contrast to the missing data, outliers are usually related to the target variable **y**. Although in several previous studies researchers removed outliers from the dataset, we argue that outliers (i.e., atypical observations) should not be removed, especially in the context of our study, as they present important information about extreme failure events and associated higher risks; thus, we retained the atypical observations in our final dataset. Similarly, imbalanced datasets are also common in the context of infrastructure failure data. To address this issue, we propose to develop a two-stage predicting modeling approach to model the sewer system failure risk (Wang, Lan, & Wu, 2017), which is explained in detail in Subsection 3.2.

3.1.2. Spatiotemporal Transformation

In this section, we discuss the strategies that can be used to address some of the discrepancies that often exist with the spatiotemporal aspect of the explanatory variables **X** and the target variable **y**. In general, when designing predictive frameworks, sufficient historical information exists to define the target variable over space and time dimensions, but there is not enough information to do the same for some of the explanatory variables. For instance, in the context of predicting sewer system failures, where the target variable **y** represents the number of failures observed at a given location during the course of a predefined time frame (e.g., a month), water utilities tend to record the detailed information of the failure events including both the time of occurrence and the location of the event. On the other hand, as discussed in Section 2.2, despite the fact that some of the explanatory variables **X** also span in both space and time dimensions, datasets consisting of historical information of those variables often lack sufficient information associated with both the dimensions.

For example, in the context of a weather-related explanatory variable like the precipitation level, most of the databases provide thorough information of the precipitation levels across time, but not in the spatial dimension (Soriano-Pulido et al., 2019). This is because the weather agencies collect their data by placing only a few sensors scattered over the area of study. Therefore, weather databases often contain sparse information on the spatial variations of such

variables, which, in turn, is significantly less granular than the target variable.

Most studies that attempt to model failure risks of sewer systems often resort to using one-dimensional variables as they are directly accessible from the data, without attempting to interpolate them into their bidimensional natural form by means of additional available information. We propose a systematic approach to produce spatiotemporal data, matching the scale of the target variable.

Let $x(s,t)$ be the *transformed* explanatory variable that one expects to generate, defined over both space s and time t, and let $x(s) = \text{proj}_s x(s,t)$ and $x(t) = \text{proj}_t x(s,t)$ represent the corresponding projections of $x(s,t)$ into the space and time dimensions, respectively. For a given context, suppose that the available database contains sufficient temporal information for a good sampling of $x(t)$. Ideally, we would like to obtain the inverse of the temporal projection, so as to generate the spatiotemporal variable $x(s,t)$ directly as $x(s,t) = \text{proj}_t^{-1} x(t)$. However, since dimension-reduction projections like the ones described before are not bijective (i.e., multiple values of $x(s,t)$ may project into the same $x(t)$), such an inverse is not well defined.

The key idea of our proposed approach is to take advantage of the partial information that is often available for the other dimension to produce a two-dimensional estimation of the given explanatory variable. To this end, there are different types of other functions f that can be used in place of such inverse functions to estimate the two-dimensional spatiotemporal variables. For example, consider the precipitation-level variable described before. To account for spatial variability, hydroscientists have traditionally used interpolation techniques to create heterogeneous precipitation surfaces over study areas. One of the most well-known methods used is "Kriging interpolation," which uses geostatistics to generate an estimated surface from scattered points (Karnieli, 1990; Yang, Xie, Liu, Ji, & Wang, 2015). More specifically, the Kriging interpolation algorithm is used to produce a mean precipitation surface for each time step (e.g., mean precipitation per year/month/day). Note that in this case, the function f is the Kriging interpolation function that takes the temporal variable $x(t)$ as an input and transforms it to a spatiotemporal variable $x(s,t)$. Similarly, using other functions, spatial variables can be transformed into spatiotemporal variables. For example, consider the population density as a variable distributed over space (i.e., $x(s)$); using forecasting techniques, such a variable can be transformed to its corresponding spatiotemporal form $x(s,t)$. Thus, identifying the accurate function f to transform one-dimensional data to the spatiotemporal dimension (s,t) is instrumental in the data preprocessing phase.

3.2. Prediction Modeling

3.2.1. Two-Stage Model

After the collected data are preprocessed using the various techniques mentioned in Section 3.1, our prediction model is developed using advanced machine learning algorithms. To predict the failure risk in an urban sewer system infrastructure, we propose a two-stage classification procedure. In the first stage, we consider the response variable **y** that consists of two classes: (1) No Risk (**N**) and (2) Risk (**R**). Here, the risk can be assessed based on different measures, for example, the number of failures in a specific component/place in a specific slot of time. Therefore, for a given input data, if the number of failures is zero, the response variable is classified as **N**; otherwise, it is categorized as **R**. In the second stage, given that there is a risk of failure (**R**), the level of risk is further categorized into various levels based on the frequency of failures. To define the risk-level categories, we propose to use a quantile-based classification approach that has been found to efficiently model a response with heavy-tailed distribution (Mukherjee, Nateghi, & Hastak, 2018; Mukherjee, Vineeth, & Nateghi, 2019). In the context of the case study presented in this article, we categorized the failures (**R**) into high-risk failures (**H**) (all observations above the third quartile) and low-risk failures (**L**) (all observations less than the third quartile). Based on this design of our response variable, we implement the proposed two-stage risk prediction modeling framework. In stage 1, failure risk class (**R**) is predicted, and the set of observations correctly predicted by the model (true positives) is used as an input for the prediction of risk levels (severity) in stage 2.

The two-stage modeling approach offers several advantages. As evident from the literature, the distribution of failure risks in a sewer system is highly skewed since the class representing the no risk (**N**) category appears to be much more prevalent in the dataset (Harvey & McBean, 2014). This type of distribution poses a challenge for training prediction algorithms, which tend to be biased toward the majority class. However, as mentioned before, the extreme observations (rarely occurring extreme failures) cannot be just removed, treating them as outliers. Especially, in the context of characterizing infrastructure

failure risk, it is critical to characterize the heavy-tail distribution of the data (Krawczyk, 2016). By design, our proposed two-stage risk prediction model can address these challenges by categorizing and analyzing the dataset into distinct classes (e.g., **N**, **R**, **H**, **L**), and thus, eliminating the bias toward any class.

3.2.2. *Prediction Workflow*

The aforementioned two-stage risk estimation model can be implemented using our proposed *workflows* that automate the machine learning procedures. These workflows execute several steps of data transformation and model execution in an iterative manner to improve the prediction performance of machine learning algorithms. The specific steps in our proposed workflow are as follows: (1) use methods of data transformation by applying feature scaling to the input features (e.g., normalization, standardization) and avoid bias toward higher-order-of-magnitude values; (2) apply algorithms for balancing the class distribution on training data (e.g., SMOTE [Chawla et al., 2002], SMOTE with ENN [Batista, Prati, & Monard, 2004]) to eliminate bias toward the majority class; (3) use dimensionality reduction methods (e.g., PCA, linear discriminant analysis) to reduce the dimension of the input feature space under consideration and obtain a set of principal variables (Roweis & Saul, 2000); and finally, (4) train and test different machine/statistical learning algorithms (e.g., DTs, RF, neural networks, support vector machines) to predict the sewer system failure risk.

Note that for each of the workflow's steps, only one of several methods can be used. As a result, different workflow candidates (i.e., combinations of different workflow steps) can be considered to predict the sewer system failure risks. To evaluate the performance of the candidate workflows against each other and select the one that outperforms all the other candidates, we define an index, called *performance index*, based on the norm 2 of a set of relevant performance metrics (e.g., accuracy,[1] precision,[2] recall,[3] F1-score[4]). Let

- $\mathcal{M}$ be the set of desired performance metrics to be considered,
- $\mathcal{P}$ be the set of defined workflows,
- $\hat{w}_m$ be the specific and desired value of the performance metric $m \in \mathcal{M}$, and
- w_{mp} be the real value of performance metric obtained for workflow $p \in \mathcal{P}$ in measure $m \in \mathcal{M}$.

With this information at hand, we can define the performance index as:

$$\|w\|_2^p = \sqrt{\sum_{m \in \mathcal{M}} (w_{mp} - \hat{w}_m)^2} \quad \forall p \in \mathcal{P}. \tag{1}$$

Lesser this value, the better is the predictive performance of the workflow. With this index, we can select the best workflow, and consequently, the best statistical/machine learning model. In case of a tie, other criteria can be used. Note that the performance metrics should be normalized; therefore, the maximum and minimum values for each metric should be 1 and 0, respectively.

3.2.3. *Bias-Variance Trade-off*

One of the main tasks in the prediction workflow is to evaluate the generalization performance of machine learning algorithms using an appropriate resampling procedure. k-fold cross-validation is a widely used resampling technique that can be used to balance the bias and variance, and estimate the out-of-sample predictive performance of machine learning models (Agarwal, Tang, Narayanan & Zhuang, 2020; Alipour, Mukherjee, & Nateghi, 2019; Hastie et al., 2009; James et al., 2013; Jung, 2018; Mukherjee & Nateghi, 2019, 2017; Obringer, Mukherjee, & Nateghi, 2020). This approach involves randomly dividing the set of observations into k-folds of approximately equal sizes. In each iteration, the kth-fold is treated as the test set, and the remaining $k-1$-folds as the training set. The test set is used to calculate the models' predictive accuracy, while the training set is used to calculate the goodness-of-fit performance of the models (Hastie et al., 2009). Further, the splitting of data into folds can be done ensuring that each fold has the same proportion of observations for each class of the response variable. This strategy is called stratified cross-validation.

3.2.4. *Hyperparameter Tuning*

For most of the machine/statistical learning models, it is required to determine the optimal values

[1] Accuracy is the ratio of the total number of correct predictions to the total number of predictions made for a given dataset.
[2] Precision is the ratio of correctly predicted positive examples to the total number of positive examples that were predicted.
[3] Recall quantifies the number of correct positive predictions made out of all positive predictions that could have been made.
[4] F1-score provides a way to combine both precision and recall into a single measure that captures both properties.

of model hyperparameters. The workflow itself can contribute to defining such values. A hyperparameter is an external characteristic of a model whose value cannot be estimated from data during the learning process. The values of the hyperparameters are set before the training of a model begins. Several mechanisms can be used to determine the optimal hyperparameters that lead to the superior performance of the models. Examples include grid search (Lameski, Zdravevski, Mingov, & Kulakov, 2015) and random search (Bergstra & Bengio, 2012). In a grid search, a whole set of combinations of the different values for each hyperparameter is used, while in a random search, only a few such combinations are used. The reason to select random search over grid search is mainly to reduce the computational cost that the grid search implies. In our case study, we applied a stratified cross-validation and random search approach in each of the stages to tune the hyperparameters of our finally selected models.

4. CASE STUDY

This section presents how our proposed methodology can be applied to a real sewer system infrastructure system to predict the failure risks. For this purpose, we selected the sewer system for the city of Bogotá (Colombia). We present a general overview of the case study and explain how our methodology can be implemented.

4.1. General Description

The sewer system infrastructure of the capital city of Bogotá (Colombia) served as a testbed for our study. Bogotá comprises a total area of 350 km^2 and more than 9,500 km of sewer pipes. The water and sewer utility in the city, Empresa de Acueducto y Alcantarillado de Bogotá (EAAB), serves over 1.7 million residential customers (family units) and about 400,000 commercial customers, with a coverage of over 98% of the city (Empresa de Acueducto y Alcantarillado de Bogotá, 2015). The water utility divides the city into five operative zones, each one with its own manager and independent resources (equipment and personnel) (Empresa de Acueducto y Alcantarillado de Bogotá, 2015). We selected zones 2 (76.72 km^2) and 3 (76.75 km^2) to perform our analysis, as access to the required data was available only for these two zones. Fig. 3 highlights the location of these two zones within the city of Bogotá.

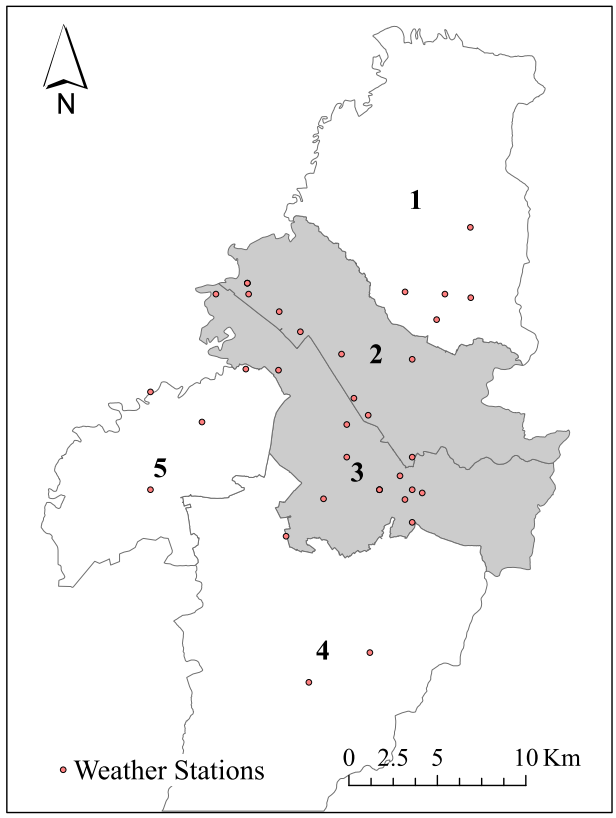


Fig 3. Operational zones 2 and 3, and distribution of weather stations used in our analyses.

4.2. Setup for Methodology Steps

In this section, we provide information about the specific settings for each step of the proposed methodology. We start by explaining the data collection process, and then, discuss the database design through an illustration of how the data issues were managed and how the spatiotemporal data were constructed. Subsequently, we present the specifics of the prediction modeling in the context of our case study. Finally, we provide additional details about the implementation of our methodology and discuss the practical implications of our results that can provide insights to the water utility to optimally allocate resources to manage the sewer system failure risks efficiently.

4.2.1. Data Sources and Description

The data for this research were collected from multiple sources, including several governmental and private agencies such as the Institute of Hydrology,

Table 3. Summary of Available Databases

Database	Source	Original Format	Dimensionality	
			Spatial	Temporal
Elevation	EAAB	Raster (Digital Elevation Model)	✓	
Failures	EAAB	Text file	✓	✓
Gullypots	EAAB	Shape file (points)	✓	
Intrusive trees	JBB	Shape file (points)	✓	
Land use	SDP	Shape file (polygons)	✓	
Manholes	EAAB	Shape file (points)	✓	
Pipes	EAAB	Shape file (polylines)	✓	
Population	SDP	Shape file (polygons)	✓	✓
Slope	EAAB	Raster	✓	
Streets	SDP	Shape file (polylines)	✓	
Weather	IDEAM	Text file	✓	✓

Table 4. List of Features in Each Database

Dabase	Features
Elevation	elevation, geographic coordinates
Failures	date, geographic coordinates
Gullypots	type (sanitary or stormwater), terrain elevation, material, installation date, geographic coordinates
Intrusive trees	total height, root exposition, physiology, geographic coordinates
Land use	type (residential, commercial, or industrial), geographic coordinates
Manholes	type (sanitary or stormwater), terrain elevation, depth elevation, material, installation date, geographic coordinates
Pipes	type (sanitary or stormwater), service (main or local), terrain elevation, crown elevation, invert elevation, length, material, installation date, diameter, geographic coordinates
Population	date (year), population, geographic coordinates
Slope	geographic coordinates, slope
Streets	type (based on hierarchy—primary and secondary), weather station geographic coordinates
Weather	date, total brightness, total evaporation, mean humidity, total precipitation, max temperature, mean temperature, min temperature, geographic coordinates

Meteorology, and Environmental Studies (IDEAM from its name in Spanish), EAAB, Bogotá's Botanical Garden (JBB from its name in Spanish), and the District Planning Agency (SDP from its name in Spanish). The information was obtained in different formats such as shapes, rasters, and text files. Table 3 summarizes the databases, while Table 4 provides the list of features present in each database.

Sewer system failure database. Bogotá's sewer system failure records are obtained from a customer complaints database provided by the utility. The failure database is generated by the water utility using the following steps: (1) first, the water and sewer utility center receives and records the customer complaints; (2) then, personnel is sent to the reported address to verify the failure; and finally (3) verified failures and their coordinates are reported back to the Customer Complaint Center including information such as failure type (we only considered sediment-related failures) and the required corrective actions. For a detailed description of the customer complaints database, the interested reader is referred to Rodríguez et al. (2012). In this study, we used the failure records data spanning over a period of six years (2005–2010). Monthly distributions of failures throughout the study period (see Fig. 4) are apparently bimodal in nature, depicting their similarity with the temporal distribution patterns of rainfall in the region during the same time period (see Fig. 6).

Information on physical characteristics of sewer system infrastructure. Similar to the failures, databases related to physical information of the sewer system infrastructure are provided by the utility. In this regard, we have access to data related to pipes, manholes, and gullypots. Our study area comprises two types of sewer pipes—sanitary (carries sewage from bathrooms, sinks, kitchens, and other plumbing components) and stormwater (designed

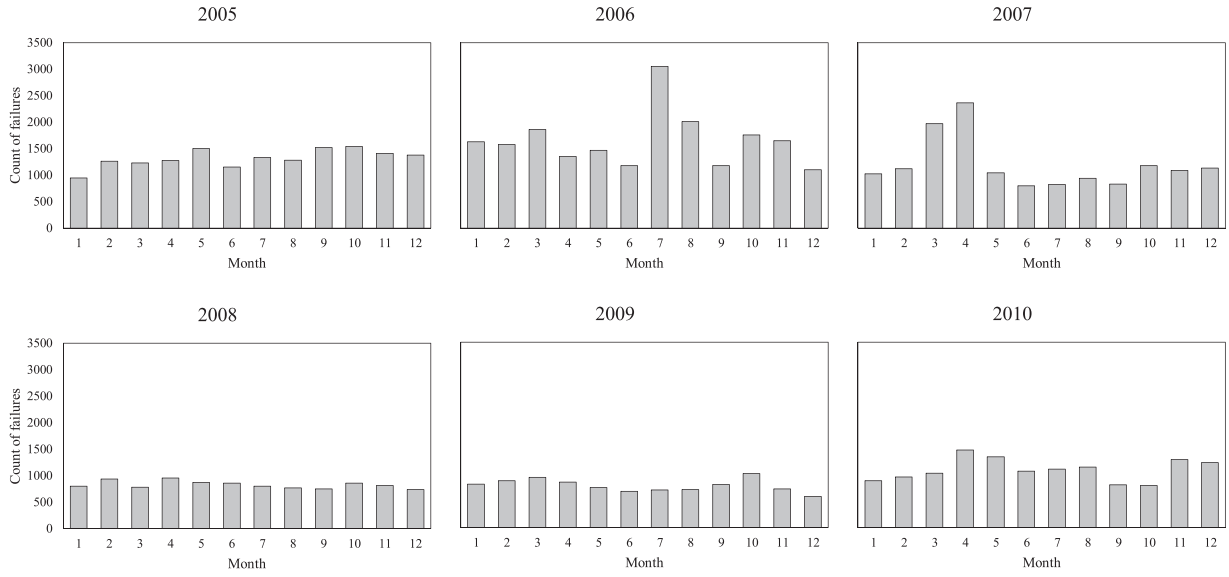


Fig 4. Distribution of the number of failures in the city for every month of the period of study.

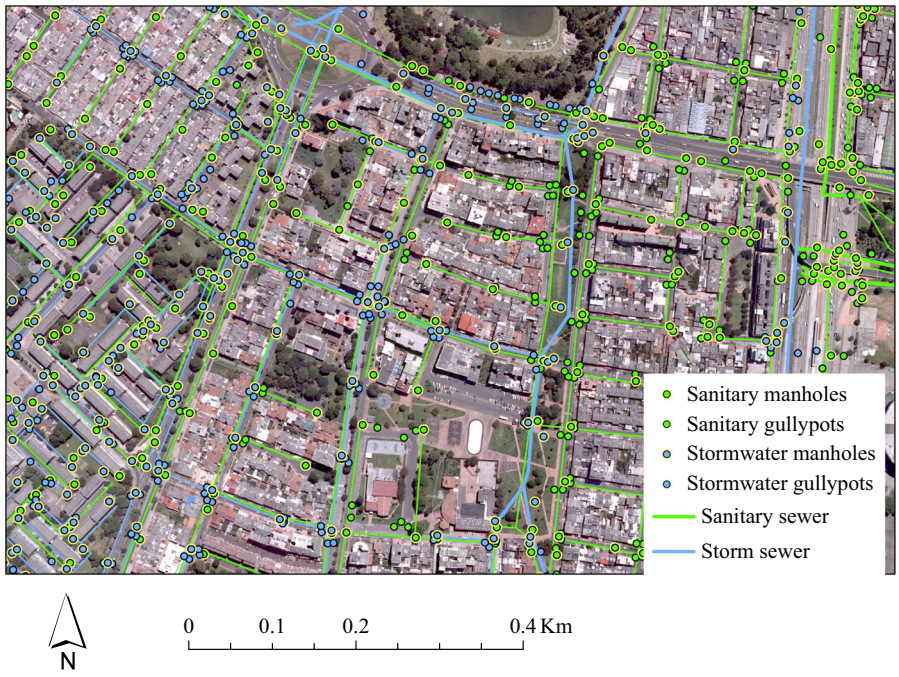


Fig 5. Sewer system infrastructure elements.

to drain excess rain from impervious surfaces such as paved streets, car parks, parking lots, roofs, etc.). In total, our study areas comprising zones 2 and 3 cover an area of 153.2 km^2 and include 3,758.5 km of sewer pipes, of which 2,786.1 km corresponds to the sanitary sewer and 972.4 km corresponds to the stormwater sewer. Similarly, the study area contains 69,888 gullypots and 66,377 manholes. Fig. 5 shows a zoomed snapshot of the case study area, illustrating the spatial distribution of the physical infrastructure elements in a few residential blocks (approximately 0.08 km^2).

Other physical variables considered in the study include elevation and slope. These variables, obtained from the EAAB, contained specific information for any geographical location of the city.

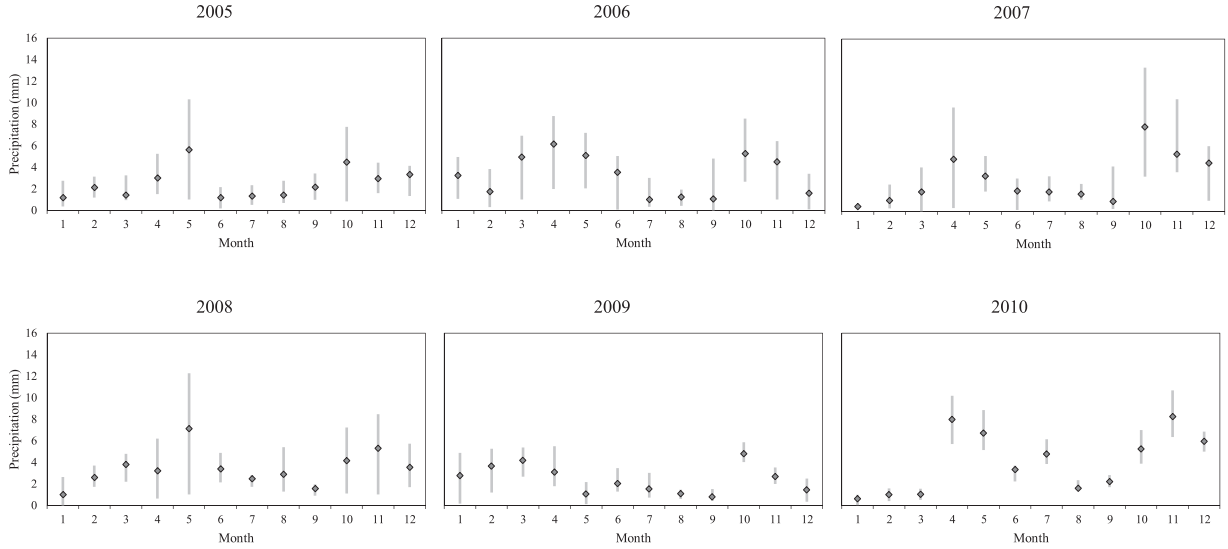


Fig 6. Distribution of the daily average rain across the city.

Environmental factors. It is known that environmental conditions can significantly influence the occurrence of failures in the sewer system. Therefore, we included several weather conditions in our study such as total brightness, rainfall, temperature, humidity, evaporation. The data were collected through the local environmental authorities (i.e., IDEAM). The specific data were obtained for 36 weather stations scattered along the study area (see Fig. 3). Fig. 6 shows the distribution of average rainfall in the city during the study period.

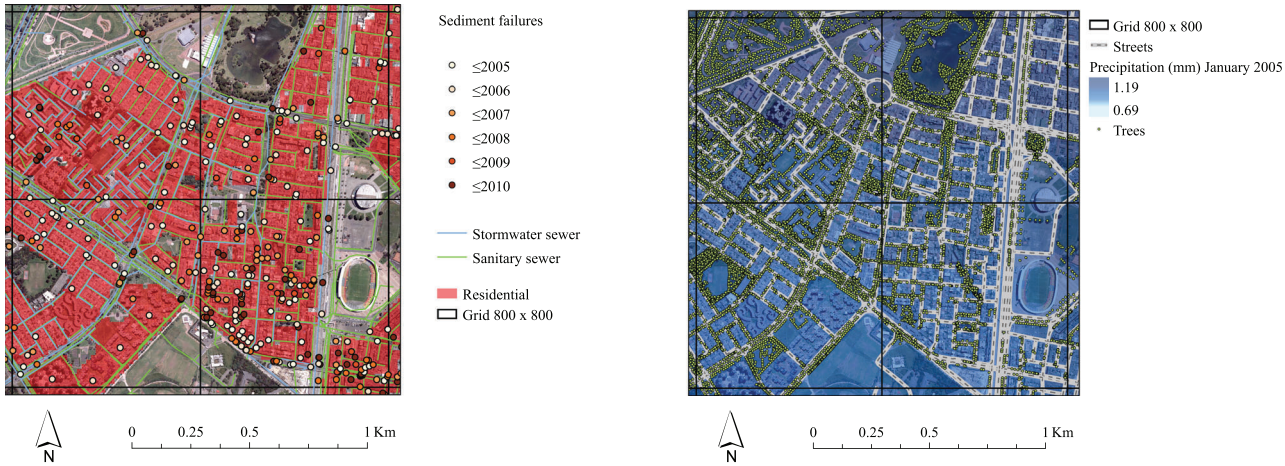
Demographic information. Demographic information on population was obtained from the publicly available most recent census data for Bogotá. The governmental agency in charge of Colombia's official statistics performs demographic projections using planning zone units (Unidad de Planeamiento Zonal [UPZ]). Using 2005 census data, the SDP estimated the total population for each UPZ for the period 2005–2015, which includes the analysis period of this study. In addition to population data, we collected land use data from the SDP that specifies if the land is used for residential, commercial, or industrial purposes.

Urban landscape information. Since previous studies established that surrounding urban infrastructure and their interactions with the sewer system can significantly influence the risk of failure, we investigated the influence of urban landscape on the risk of sewer system failure (see Table 2). In this article, the urban landscape is described by urban elements such as intrusive trees and streets. From the tree species identified in the urban area of Bogotá, the urban tree planting authority (JBB) has identified 54 tree species that are capable of causing root intrusions. We extracted the location and information of the 81,592 intrusive trees present in the case study area from the most-updated tree census—elaborated in 2007. Besides the (x,y) coordinates of each tree in the urban area, the census contains tree species, total height, physiology, and a brief description of the leaves and the site location, among other characteristics. For more information on Bogotá's tree census, the interested reader is referred to Torres et al. (2017).

We also included a geo-referenced database of highways and streets in our study. This database was gathered in 2013 and contained information on each street segment classified into one out of five categories: two for primary and three for secondary roads.

4.2.2. *Database Design: Data Preprocessing and Aggregation*

The data were collected from several sources; thus, the format and the dimensionality units varied significantly among the different databases. Therefore, the first step is to merge the data under the



(a) Failures per year, pipes, land use

(b) Streets, precipitation (daily average), trees

Fig 7. Intersection of data from different databases and the grid of equally-sized cells.

same metric system. It is important to note that the sewer system failures reported by the users are not associated with a particular component of the sewer system (e.g., pipe, manhole, or gullypot); but rather with an address that is converted into a point coordinate in the zone where the failure is reported. Despite this might pose a difficulty when identifying the precise element that failed, this information does provide insights on the areas of the city (i.e., block or group of blocks) that are more likely to experience sewer system failures. Given the format of available information, Rodríguez et al. (2012) proposed to partition the sewer system into a set of equally-sized cells; the authors used a 170×170 m^2 area such that it corresponds to approximately one residential block. Following this idea, in our study, we partitioned the zones into equally-sized cells, where the cell size is considered as the unit for space. The time unit considered is month, similar to that reported in the previous studies by Rodríguez et al. (2012) and Soriano-Pulido et al. (2019).

With this information at hand, we built a database with the target variable **y** as the number of failures reported in a cell in a specific month. Note that the explanatory variables were constructed based on the variables provided by the collected databases (see Table 4). In this case, the explanatory variables were computed as aggregated information of the variables in the original databases (e.g., counting of gullypots, counting of gullypots of a specific material, counting of trees, sum of the length of the streets). The complete list of constructed variables is provided in Table B1 in the Appendix. Fig. 7 helps to visualize this database construction by showing how the equally-sized cells intersect with the data collected from different databases. In the figure, we observe pipelines (stormwater and sanitary), failures in different periods of time, residential areas, precipitation for one period of time, trees, and streets for an example grid.

Managing data issues. As described in Section 3.1.1, addressing the data issues is instrumental for predictive modeling. In this section, we focus on addressing the issues for **X**, while the issues regarding **y** are addressed in Section 4.2.3. Given that the space unit is a cell, we constructed the explanatory variables as aggregations (sum, average, counting) of the original variables based on their intersection with the predefined grid of equally-sized cells (see Table B1). Nonetheless, to do that, we needed to impute the missing data in the original databases first.

To impute missing data, we applied the methodology described in Section 3.1.1. More specifically, we designed the dynamic neighborhoods based on the partition of the zones into equally-sized cells. Fig. 8 depicts the dynamic of the iteratively changing neighborhood. To impute missing values of numerical data, we averaged the observations falling within the neighborhood, referred to as the "mean" approach; while for categorical data, we used a "random approach," which assigns a random value to the missing data point based on the distribution of

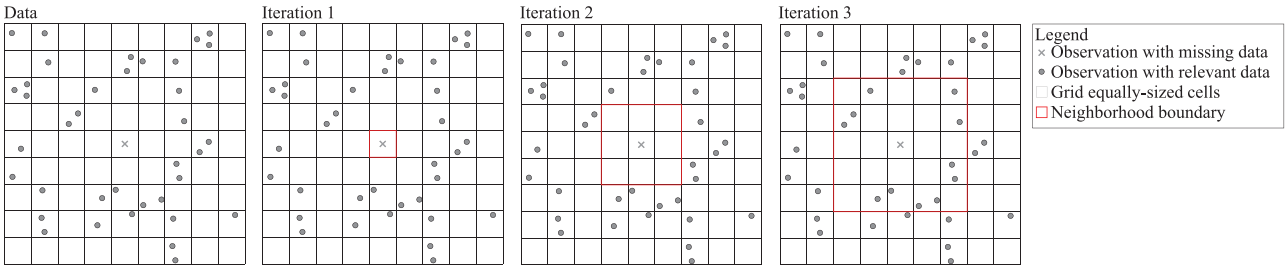


Fig 8. Three iterations defining the neighborhood used for the data imputation. From iteration 1 to 3, the neighborhood is increasing. Iterations 1 and 2 do not have observations in the neighborhood; therefore, no data can be imputed. Iteration 3 provides observations within the neighborhood and imputation can be performed.

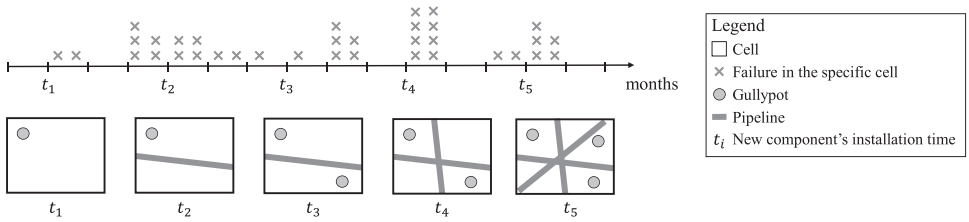


Fig 9. Failures per month in a specific cell. Installation of new infrastructure over the time. Failures are only related to the existent infrastructure at the time of the failure. Failures between time t_1 and t_2 are related to the infrastructure installed in t_1. Failures between time t_2 and t_3 are related to the infrastructure installed in t_1 and t_2, and so on.

the specific variable within the neighborhood. For example, to impute terrain elevation for manholes, we used average terrain elevation of other manholes within the neighborhood, while for pipeline diameter (which is not a continuous variable), we randomly selected a diameter from the distribution of pipelines diameters within the neighborhood. Note that additional constraints were added to the imputation process, for example, considering pipelines within the neighborhood that share the same material as the imputed observation.

Spatiotemporal data. Table 3 shows the dimensionality of the data, while Table 4 presents the specific variables found in each dataset. As mentioned in Section 3.1.2, to match the spatiotemporal dimensions of our response variable, we transformed the explanatory variables to account for the spatiotemporal perspective as well.

For those variables that are spatial, if there is any type of time-dimension information (e.g., installation date), their aggregated counterparts (as mentioned in section "Managing data issues") can be calculated as spatiotemporal variables. For example, if the installation date for gullypots is given, in the final variable "count of gullypots" used in the analysis, we only consider those gullypots that were already installed at the time of the failure. This process is depicted in Fig. 9, where the failures in a specific cell between time t_1 and t_2 are only related to the infrastructure installed at time t_1 and before (in the illustration, just one gullypot is shown) in such a cell. Similarly, failures between t_2 and t_3 are only related to the infrastructure installed in t_2 and before (in the illustration, a gullypot and a pipe). As new infrastructure is built over time, the failures are related to the corresponding infrastructure following this logic. This process was applied to gullypots, manholes, and pipes in our study. As an outcome of this method, we are removing the assumptions made by Rodríguez et al. (2012) and Soriano-Pulido et al. (2019), who considered that spatial variables related to the infrastructure remain the same throughout the study period.

Likewise, for temporal variables possessing spatial information (e.g., cartographic coordinates or an identifier that can be joined to spatial objects), it was possible to create their aggregated counterparts (e.g., mean, minimum, and maximum values) over the spatiotemporal dimension. For example, population data (given originally per UPZ) are joined with the corresponding polygons to observe population variations in space and time. Similarly, weather data from different weather stations were used to create

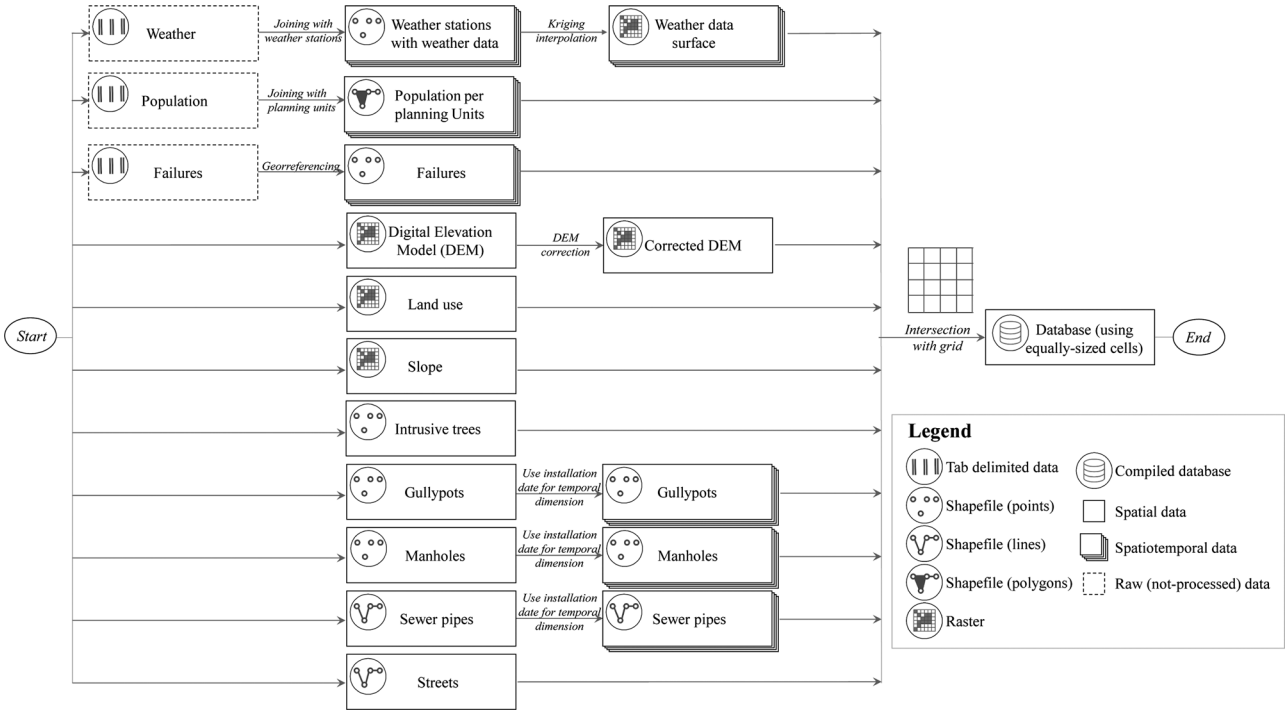


Fig 10. Spatiotemporal database construction.

continuous surfaces over the space applying kriging interpolation for each month (refer to Section 3.1.2).

Some variables were kept as spatial variables as they do not vary over time, namely, slope and elevation. Besides, although land use, streets, and intrusive trees may change over time, there was not enough information available to consider their temporal variations, and thus, they were kept as spatial variables. Fig. 10 illustrates how the information was handled to build the final dataset, specifying the operations to transform raw data to the spatiotemporal dimension.

Following this procedure and the missing data management described in Subsection 3.1.1, 222 explanatory variables were considered when combining all the datasets. Table B1 in the Appendix presents a description of such variables and Fig. A1 shows the correlation matrix for them. The number of observations in the analyzed datasets ranges from 10,000 to 100,000.

4.2.3. Predictive Modeling

The target variable (i.e., response variable **y**) for this study represents the number of failures in each cell every month, which is a spatiotemporal variable (explained in Section 4.2.2). To predict such a response variable, we propose to develop the two-stage prediction model described in Section 3.2.1. As already explained before, both stages are supported by our proposed prediction workflows. In the first stage, **y** is transformed to have only two responses: no risk **N** and risk **R**, where risk is associated with one or more failures. In the second stage, **y** is also coded in two categories: low-risk **L** and high-risk **H**. In doing so, we use the observations classified as **R** in the first stage, and then classify those observations with the number of failures above the third quartile of the failure distribution curve as high risk (**H**), while all the other observations (with the number of failures below the third quartile) as low risk (**L**). True positives (correctly identified data points under the failure risk class **R**) that are obtained from the prediction results of stage 1 act as the input data for the prediction model (**L**/**H**) in stage 2 that is leveraged to predict the failure risk levels.

The workflows in both the stages are defined as observed in Fig. 1. As shown in the figure, the first step in the workflow is scaling the input data. In this step, only two possibilities were available, scale or not scale. In the second step, the workflow involved the utilization of sampling algorithms for balancing the class distribution on training data. We considered

six different alternatives—random oversampling and undersampling (Yap et al., 2014), SMOTE (Chawla et al., 2002), SMOTE with ENN (Batista et al., 2004), near-miss (Mani & Zhang, 2003), and not balancing the data. For more information regarding these methods, the interested reader is referred to Lemaître, Nogueira, and Aridas (2017). In the thirds step, dimensionality reduction was performed. Similar to the first step (scaling data), we considered two possibilities in this step: reduce or not reduce the dimensionality of the data. Reducing the dimension of the input feature space is achieved using PCA. This method uses an orthogonal transformation to construct a low-dimensional representation of the data that describes as much of the variance in the data as possible (Vasan & Surendiran, 2016).

The final step in the workflow consists of training and testing the different parametric and nonparametric (i.e., machine learning) models using the final dataset. The models used include LR, DTs, RFs, and XGBoost (XGB). LR is a classification algorithm that assigns observations to a discrete set of classes. Unlike linear regression, where predictions are continuous values, LR transforms its output using the sigmoid function to return a probability value, which can then be mapped to two or more discrete classes (Dreiseitl & Ohno-Machado, 2002). In DT, the training instances are classified into a tree structure using decision rules that are heuristically derived during the learning phase (Lavanya & Rani, 2012). The tree consists of decision nodes and leaf nodes where each node represents a test over an attribute of the input data and each leaf node has an associated class that is the outcome of the decision for a particular case. RF is the evolution of DT; it is an ensemble of DTs where the training set for each tree is selected using bootstrap sampling from the original sample set. The number of features that are considered for splitting at each tree node is a random subset of the original set of features. The final estimate of RF is the classification result that receives the maximum number of votes across all trees (Breiman, 2001; Mukherjee & Nateghi, 2019). XGB is an ensemble tree-based algorithm using gradient boosted DTs (Chen & Guestrin, 2016). In XGB, a large number of weak learners are built and combined sequentially to produce a strong learner. The difference between XGB and other gradient boosting algorithms like gradient boosting machines or gradient boosting trees (Agarwal et al., 2020) is due to its regularization formalization to control overfitting. Indeed, the name XGB refers to the goal of taking the computational power to its limits; using OpenMP API (Chapman, Jost, & Van Der Pas, 2008) for parallel processing, XGB is able to utilize the multiple cores that are available on a single machine's CPU for parallel computation (Chen et al., 2018).

Based on these steps, we generated a total of 96 different workflows (in order of workflow steps: $2 \cdot 6 \cdot 2 \cdot 4$). In addition to these 96 workflows, we considered the default implementation of four extra models, namely, balanced bagging classifier, balanced RF classifier, easy ensemble classifier, and RUSBoost classifier (Lemaître et al., 2017; Pedregosa et al., 2011). Since these ensemble classifiers use the sampling algorithms for internally balancing the data, it reduces the number of steps in the workflow and parameters that need to be provided by the user. Therefore, for these classifiers, only the options for scaling and dimensionality reduction were provided. Therefore, a total of 16 additional workflows were used ($2 \cdot 2 \cdot 4$). In this manner, we evaluated the performance of $|\mathcal{P}| = 112$ workflows in each stage in each scenario.

To select the best workflow, we used the index based on norm 2 defined in Section 3.2.2. For this study, we considered two measures to be included in the predefined index (i.e., $|\mathcal{M}| = 2$). First, the macroaverage of the F1-score in the test dataset, and second, the absolute value of the difference of the macroaverage of the F1-score [5] between the test and training datasets. The F1-score is a well-established classification performance measure that conveys a balance between precision (P) and recall (R) (Zhang, Wang, & Zhao, 2015). It is known to be more informative and useful than classification accuracy in case of problems with a class imbalance. When only one class is considered, the standard F1-score is defined as the harmonic mean of P and R,

$$F1 = \frac{2PR}{P+R}, \tag{2}$$

where

$$P = \frac{TP_i}{TP_i + FP_i}, \tag{3}$$

$$R = \frac{TP_i}{TP_i + FN_i}, \tag{4}$$

TP_i is the number of test instances correctly assigned to the class i (that is, the number of true positives),

[5] A macroaveraged F1-score is achieved simply by averaging the scores over the classes.

FP_i is the number of test instances the system predicts mistakenly to be a member of the class i (that is, the number of false positives), and FN_i is the number of test instances that belong to the class i in the real data but not in the system output (that is, the number of false negatives). As multilevel classification can be decomposed into distinct binary classification problems, the F1-score can be calculated separately for each class. Macroaverage F1-score treats all classes equally regardless of the number of records within a class, which helps to select the algorithm that performs the best across all of the different labels.

The desired value for our first measure (macroaveraged F1-score of the test dataset) is $\hat{w}_1 = 1$, while for the second (absolute difference between macroaverage F1-scores of test and training datasets) is $\hat{w}_2 = 0$. For each workflow $p \in \mathcal{P}$, the index is calculated as $\|w\|_2^p = \sqrt{(w_{1p} - \hat{w}_1)^2 + (w_{2p} - \hat{w}_2)^2} = \sqrt{(w_{1p} - 1)^2 + (w_{2p} - 0)^2}$. When a tie was obtained through this exercise, we defined the best workflow in terms of complexity—the less the complexity, the better the workflow, per Occam's razor rule. Such complexity refers to the combination of the specific algorithms used in each step of the workflow. For example, a workflow not using a balancing method is simpler than one using a balancing algorithm. Once the best workflow was selected, we performed stratified 10-fold cross-validation to control the bias-variance trade-off. We also used this approach as a part of the hyperparameter tuning process by implementing a random search algorithm. When the workflow for the first stage was optimized, the results from this stage were used as inputs for the workflow in the second stage, where the same process was carried out again (i.e., 112 workflows were trained for each scenario). Thus, we selected the pair (first and second stage) that together produced the best result in predicting the risk of failures.

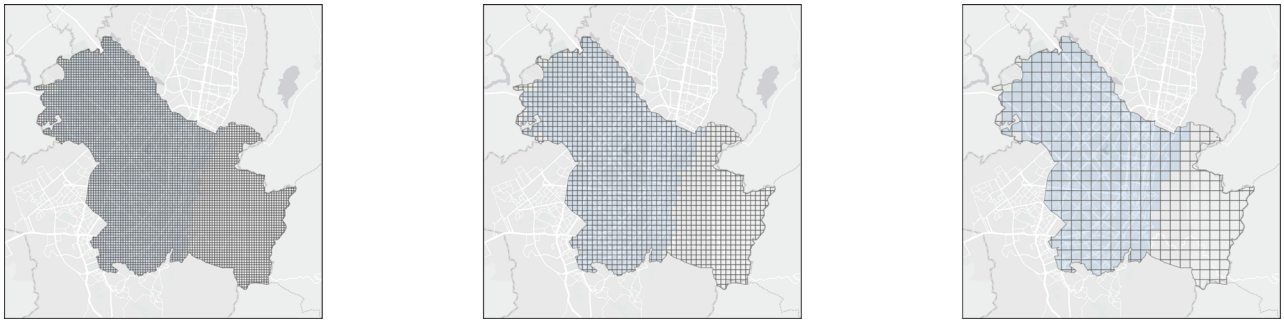
4.2.4. *Managerial Questions and Sensitivity Analyses*

As explained in Section 4.2.2, in the study conducted by Rodríguez et al. (2012), the authors used a set of equally-sized cells of a dimension of 170×170 m^2 each to perform the analysis of the data. Nonetheless, an interesting question arises from the managerial perspective: "What is the adequate size for the cells to make a robust prediction?" The selection of a single cell size renders it to be difficult to evaluate whether this selection influences the quality of the results. Soriano-Pulido et al. (2019) followed a similar procedure to that described by Rodríguez et al. (2012) and found that the spatiotemporal relation of the failures should be studied with cell sizes of at least 283 m of length. Based on this fact, the authors first studied their data using cells of 400×400 m^2 and when the number of failures in the cells was not high enough, they changed the cell size to 800×800 m^2. In our study, we analyzed three possible sizes for the cells, 200×200 m^2, 400×400 m^2, and 800×800 m^2. Fig. 11 shows the granularity of these grids of equally-sized cells over the study area, including 5,818, 1,521, and 418 cells for 200×200 m^2, 400×400 m^2, and 800×800 m^2, respectively. Similarly, Fig. 12 shows the density of cells per number of registered failures through the period of our study. Figs. A2, A3, and A4 present animations of the failures' behavior through the spatiotemporal dimension. To identify the optimal cell size that would serve our purpose, we used two criteria. First, the selected size must help balance the data. Second, it must avoid the extreme sensitivity of our model toward the marginal changes in risk threshold as defined by the managers.

Another important question from the perspective of the utility managers that needs to be addressed is related to the independence of the zones in managing their own resources in a decentralized fashion. As we analyzed the behavior of failures in the two different zones, we also studied whether considering the zones independently or as a whole generated differences in the management of resources and scheduling of preventive maintenance operations. To analyze this, we applied our methodology independently to three datasets, namely, zone 2, zone 3, and zones 2 and 3 combined. Thus, it is important to note that our sensitivity analysis considered separately a total of nine scenarios (i.e., three datasets—grid cell sizes— and three zone types).

5. RESULTS

In Section 5.1, we present the results of the model from the *first stage* (i.e., model predicting the risk of a sewer system failure), and then in Section 5.2, we describe the results of the model from the *second stage* (i.e., model predicting if the severity of the risk corresponds to high or low risk). More specifically, Section 5.1 presents the results for: (1) 112 workflows constructed for each of the nine scenarios mentioned in Section 4.2.4; (2) hyperparameter tuning for the nine best workflows, which serves as a tool for the utility managers to define the



(a) 200x200m^2 (b) 400x400m^2 (c) 800x800m^2

Fig 11. Different grid sizes used for analysis.

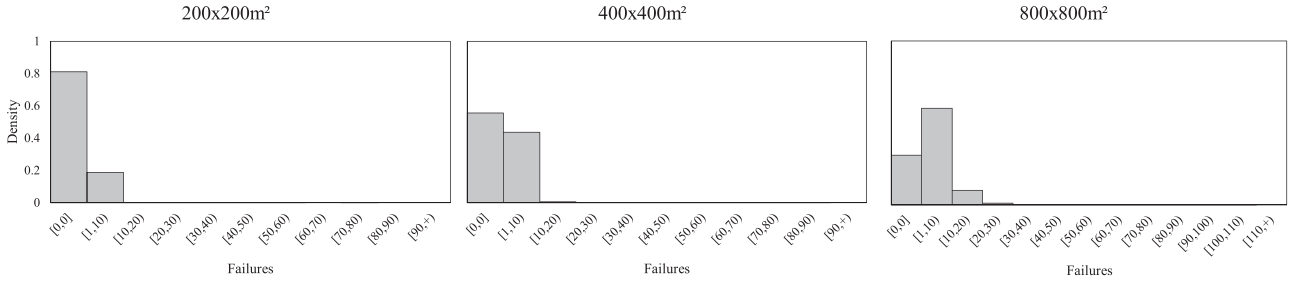


Fig 12. Density curves in terms of the percentage of cells with a number of failures within specific ranges of failures. The figure shows the density plot for the three different sizes for the cells.

discrimination threshold for **N** and **R** (i.e., probability of failure) and also provides insights to select the adequate size of the cells for robust decision making; and (3) variable importance for the *first-stage* model. In Section 5.2, we present the results for the *second-stage* model considering the specific selection regarding the size of the cells made in the first stage. We also provide results for the relevant hyperparameter tuning and feature importance for predicting low **L** and high **H** risk.

5.1. Stage 1

As defined in Section 4.2.3, we used two main performance metrics to construct our performance index $\|w\|_2^p$, which was used to select the best model. The selection was made based on the minimum value of such an index. Fig. 13 presents the results not only for the index but also for the F1-score and accuracy for each of the nine scenarios mentioned in Section 4.2.4. The workflows were ordered alphabetically by—name, scaling, balancing, dimensionality reduction, and machine learning method.

In Fig. 13, we observed that as the cells' size increases, the value of our index decreases, providing the best performance score for the 800×800 m^2 cells. In addition, it was observed that the other two metrics (F1-score and accuracy) increase as the size of the cell increases. Although the increase in cell size demonstrated better performance scores, there is an important trade-off that must be carefully analyzed. As observed from Fig. 12, with the increase in cell size, the number of cells with the risk of failure (one or more failures) increases, leading to a reduction in the number of nonfailure observations from the dataset.

In the plots depicted in Fig. 13, it is possible to observe high values of accuracy for the 800×800 m^2, reaching up to 15% increased accuracy compared to the grid size of 400×400 m^2, while such an increase is about 2% for 400×400 m^2, compared to that of the 200×200 m^2. Note that unlike grid sizes 200×200 m^2 and 800×800 m^2, 400×400 m^2 provided almost similar values for both accuracy and F1-score. In the case of 200×200 m^2, there was a difference of almost 10% on average between these two measures, while in the case of 800x800 m^2, the difference between the scores was almost 3%. The reason behind these observations is particularly related to the imbalanced failure observations in the case of 200×200

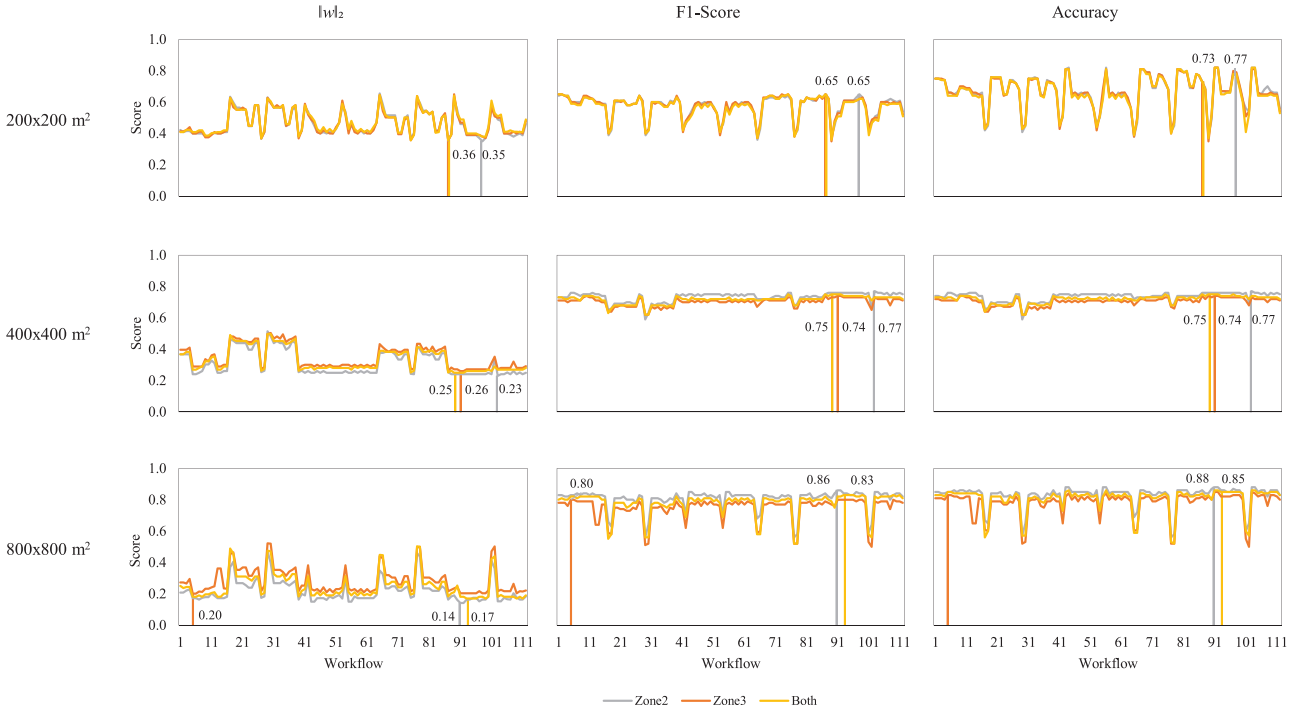


Fig 13. Our performance index, F1-score, and accuracy for the nine scenarios of the case study.

Table 5. Best Workflow for Each Scenario in Stage 1

Grid	Zone	Workflow	Scaling	Balancing method	PCA	ML model
200×200	2	99	Yes	SMOTE	No	XGBoost
	3	88	No	SMOTE with ENN	No	Random Forest
	Both	88	No	SMOTE with ENN	No	Random Forest
400×400	2	104	No	No	Yes	XGBoost
	3	92	No	No	No	XGBoost
	Both	90	No	No	No	XGBoost
800×800	2	92	No	No	No	XGBoost
	3	6	No	No	No	Easy Ensamble
	Both	94	No	Random oversampling	no	XGBoost

m^2 and 800×800 m^2 (see Fig. 12). Higher the extent of imbalance within the datasets, the higher the difference between F1-score and accuracy. These results provide important insight to utility managers on the selection of performance measures. Accuracy can be used when the class-distribution is similar (in case of 400×400 m^2), while F1-score is a better metric when there are imbalanced classes (as observed in 200×200 m^2 and 800×800 m^2). Table 5 presents a summary of the workflows selected and observed in Fig. 13. In this table, it is observed that for grid sizes that have imbalanced failure distributions such as 200×200 m^2 and 800×800 m^2, the best workflow included balancing methods to achieve better performance.

Fig. 14 presents how the F1-score of individual classes and the macroaverage F1-score between classes were affected when the discrimination threshold between failure and nonfailure changed. This figure serves to identify which workflows are more stable between different scenarios (cell size and type of zones). This analysis is conducted in terms of the marginal changes of the F1-score based on the selected discrimination threshold. As it is evident from Fig. 14, regardless of the dataset (zone 2, zone 3, or both), the 400×400 m^2 cell size provided the most stable results for the F1-score. The other two cell sizes presented extremely abrupt changes on the F1-score when changing the threshold, meaning that

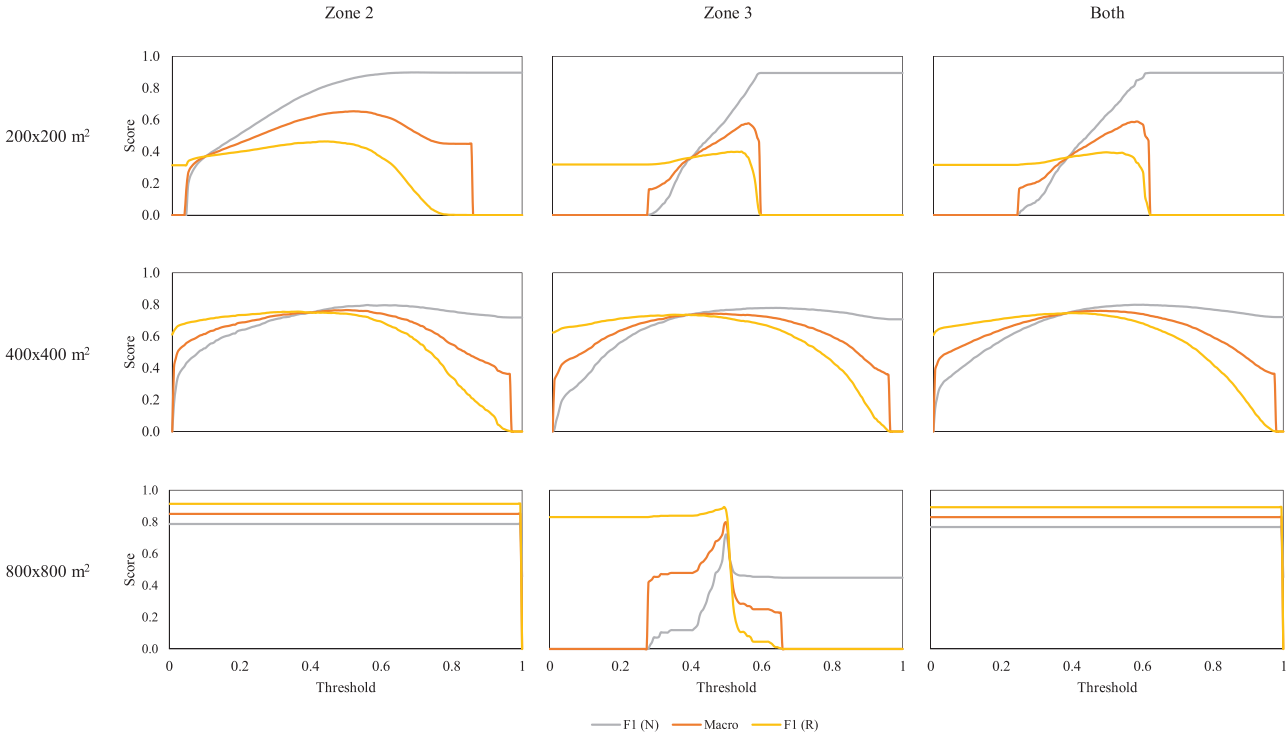


Fig 14. Discrimination threshold and its effect over the F1-score for the nine scenarios of the case study in stage 1.

those models were highly sensitive, and therefore, could not be generalized irrespective of the type of zones being considered for predicting the sewer system failure risk. Thus, we selected the models corresponding to the 400×400 m^2 to continue our analysis.

As shown in previous studies, the receiver operating characteristic (ROC) curve can be used as a diagnostic metric for the prediction models (Li, Wang, Leung, & Jiang, 2010; Terti et al., 2019). Fig. 15 presents the ROC curve to illustrate the diagnostic performance of our selected workflows for 400×400 m^2 as the discrimination threshold was varied. A value of 0.5 for AUC suggests that the diagnostic test has no discriminatory ability (Terti et al., 2019). ROC curves above this diagonal line are considered to have reasonable discriminating ability (Li et al., 2010). In the field of medicine, 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding (Mandrekar, 2010). Using this as a benchmark, we observed that the AUCs for each zone were well above 0.8, thereby demonstrating excellent discriminating ability of the workflow selected for 400×400 m^2 grid.

To identify factors that represent vulnerability for the sewer system infrastructure (Ezell, 2007), Table 6 shows the importance of variables, grouped by different categories, as listed in the first column. An example of a variable from each category is

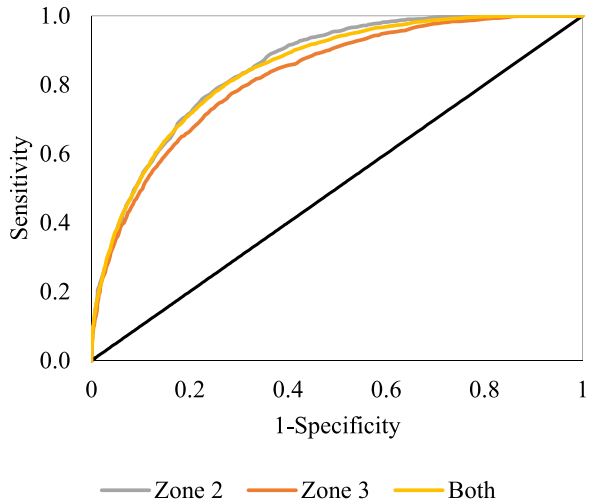


Fig 15. ROC curve for scenarios in stage 1 with cell size of 400×400 m^2. The AUCs are 0.85, 0.82, and 0.85 for Zone 2, Zone 3, and Both, respectively.

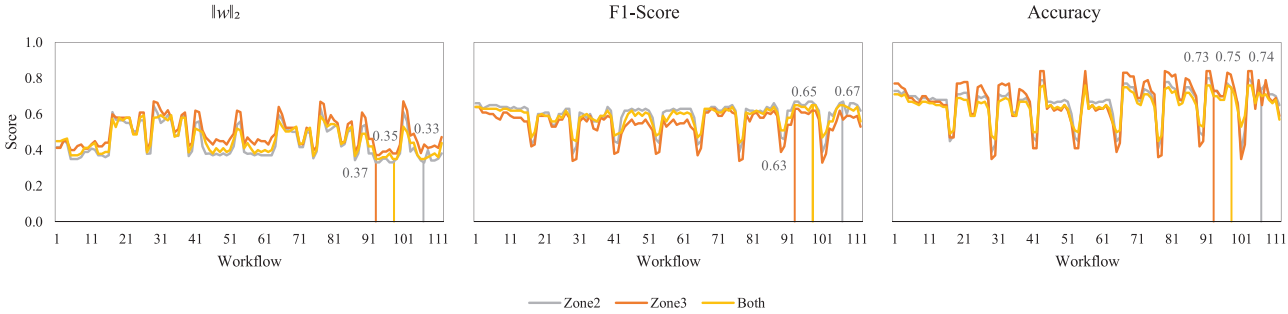


Fig 16. Our performance index, F1-score, and accuracy for the three scenarios of the case study in stage 2.

Table 6. Feature Importance Stage 1

Category	Zone 2	Zone 3	Both	Variable example
Age	0.709[01]	0.104[06]	0.065 [10]	Average installation date sanitary gullypots
Diameter	0.000[11]	0.119[05]	0.122 [07]	Number of main stormwater pipes of diameter 1
Elevation	0.000[12]	0.007[13]	0.004 [14]	Terrain elevation
Gullypots	0.188[07]	0.091[08]	0.072 [08]	Sanitary gullypots material 2
Intrusive trees	0.001[10]	0.087[09]	0.069 [09]	Average height intrusive tree
Landuse	0.201[05]	0.038[10]	0.032 [11]	Residential area
Local Pipelines	0.198[06]	0.461[02]	0.312 [02]	Total local sanitary pipes length
Main Pipelines	0.172[08]	0.135[04]	0.154 [05]	Total main sanitary pipes length
Manholes	0.222[04]	0.099[07]	0.133 [06]	Total number of stormwater manholes
Sanitary components	0.325[03]	0.509[01]	0.420 [01]	Number of local sanitary pipes
Slope	0.000[13]	0.003[14]	0.009 [13]	Slope
Stormwater components	0.361[02]	0.205[03]	0.163 [04]	Number of local stormwater pipes
Streets	0.013[09]	0.036[11]	0.176 [03]	Sum of length of streets hierarchy 1
Weather	0.000[14]	0.024[12]	0.014 [12]	Relative humidity
Zone	0.000[15]	0.000[15]	0.003 [15]	Water utility operational zone

Variables are grouped by different characteristics. Interval between 0 - 1. Group of variables are not mutually exclusive. A rank is provided in brackets

provided in the last column of the table (a complete list of the variables and their categories is provided in Table B1). It can also be observed from Table 6 that across the different zones, variables related to characteristics of sewer infrastructure such as pipelines, manholes, sanitary, and stormwater components have a significant effect on the prediction performance of the failure risk models, representing the risk factors that potentially increase the vulnerability of the sewer system infrastructure (Ezell, 2007). The least influential variables are the ones that are grouped under elevation, slope, streets, intrusive trees, and weather categories. Another important observation is that the importance of the variable categories related to age, streets, and intrusive trees is sensitive to the zone type (Zone 1, Zone 2, or Both). This provides important insight to utility managers in identifying the influential risk factors that can help in informed decision-making related to maintenance activities and resources allocations in the different zones.

5.2. Stage 2

In stage 1, it was observed that the values of performance metrics (accuracy, F1-score, and our proposed metric) were approximately similar for 400×400 m^2 grid for the different zones due to a relatively balanced failure class distribution. Additionally, this grid size presented the most stable results when analyzing F1-scores (individual and macroaverage) with respect to the discrimination threshold. As a result, this grid size was chosen for subsequent experimentation in stage 2 prediction. In Fig. 16, it is observed that there is almost a 10% decline in the F1-score as compared to the results presented in Fig. 13, and the values of F1-score and accuracy differ from each other by approximately 10%. This change in

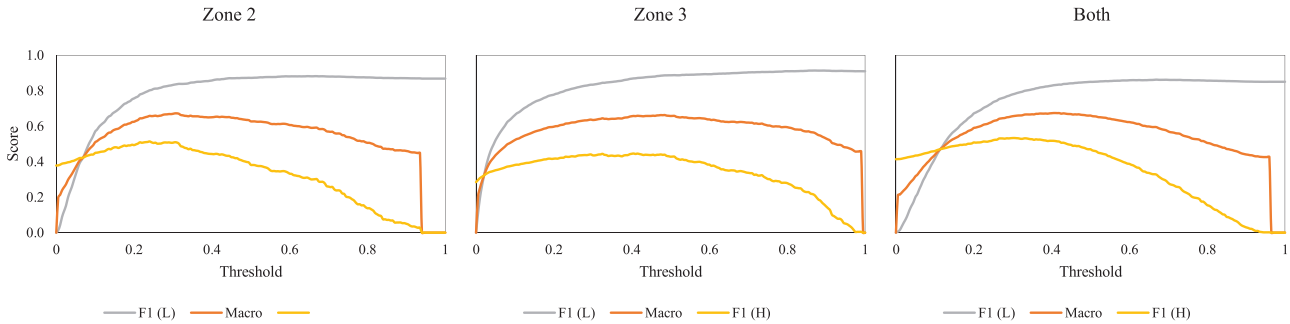


Fig 17. Discrimination threshold and its effect over the F1-score for the three scenarios of the case study in stage 2.

Table 7. Best Workflow for Each Scenario in Stage 2

Zone	Workflow	Scaling	Balancing method	PCA	ML model
2	108	No	Random Under Sampling	Yes	XGBoost
3	94	No	Rand Over Sampling	No	XGBoost
Both	99	Yes	SMOTE	No	XGBoost

the predictive performance while transitioning from stage 1 to stage 2 is due to the imbalanced distribution of low-risk and high-risk failures (see definition of **L** and **H** in Section 4.2.3). In Table 7, it is observed that the imbalanced distribution also causes the workflow to automatically incorporate balancing algorithms to enhance the predictive performance. Furthermore, XGBoost outperforms all the other prediction models for all the scenarios.

Fig. 17 illustrates the sensitivity of F1-score pertinent to the individual classes and the macroaveraged F1-score among classes with respect to the discrimination threshold between low-risk and high-risk failures. Similar to Fig. 14, it is observed that, regardless of the dataset (zone 2, zone 3, or both), the 400×400 m^2 grid provided similar trends for the F1-score. However, in these plots, when transitioning from one zone to the other, the location of the intersection between the three curves changed; this is significantly different compared to that observed in Fig. 14.

Fig. 18 presents the ROC curve to illustrate the diagnostic performance of our selected workflow for 400×400 m^2 as the discrimination threshold was varied. Following the benchmark used in stage 1 prediction, we observed that the AUCs for each zone were well above 0.7, thereby demonstrating the acceptable discriminating ability of the workflow selected for 400×400 m^2 grid in stage 2.

Table 8 shows the importance of variables from

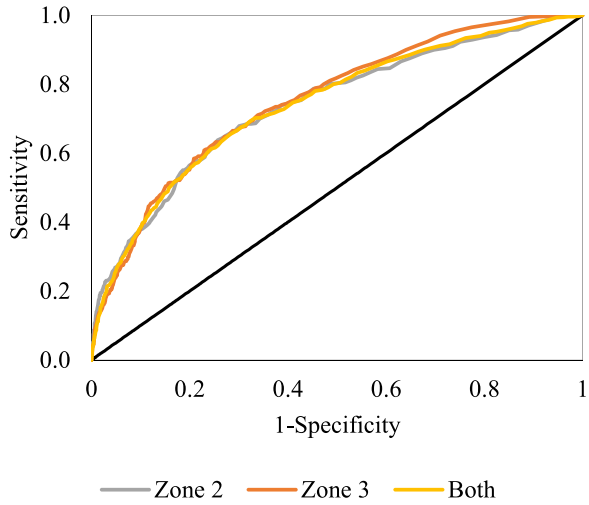


Fig 18. ROC curve for scenarios in stage 2 with cell size of 400×400 m^2. The AUCs are 0.74, 0.75, and 0.74 for Zone 2, Zone 3, and both, respectively.

the stage 2 prediction model grouped by different categories, as listed in the first column. Similar to observations in Table 6, across different zones, variables related to characteristics of sewer infrastructure such as pipelines, manholes, sanitary, and stormwater components have a significant effect on the models' prediction performance. The least influential variables are the ones that are grouped under elevation, slope, streets, intrusive trees, and weather categories. The importance of different variable categories such as age, streets, intrusive trees, and weather are sensitive to zone scenarios (Zone 1, Zone 2, or both). Although insignificant in the case of individual zones, weather characteristics appear to have a comparatively higher significance while taking both the zones into account.

Table 8. Feature Importance Stage 2

Category	Zone 2	Zone 3	Both	Variables Example
Age	0.752[01]	0.125[06]	0.085[08]	Average installation date sanitary gullypots
Diameter	0.000[12]	0.056[09]	0.072[10]	Number of main stormwater pipes of diameter 1
Elevation	0.000[13]	0.006[15]	0.011[14]	Terrain elevation
Gullypots	0.198[06]	0.119[07]	0.169[05]	Sanitary gullypots material 2
Intrusive trees	0.002[11]	0.080[08]	0.134[06]	Average height intrusive tree
Landuse	0.127[08]	0.044[10]	0.045[12]	Residential area
Local pipelines	0.225[05]	0.278[02]	0.209[03]	Total local sanitary pipes length
Main pipelines	0.151[07]	0.194[04]	0.174[04]	Total main sanitary pipes length
Manholes	0.276[04]	0.177[05]	0.123[07]	Total number of stormwater manholes
Sanitary components	0.402[02]	0.430[01]	0.373[01]	Number of local sanitary pipes
Slope	0.000[14]	0.008[14]	0.014[13]	Slope
Stormwater components	0.372[03]	0.261[03]	0.223[02]	Number of local stormwater pipes
Streets	0.015[10]	0.038[11]	0.051[11]	Sum of length of streets hierarchy 1
Weather	0.000[15]	0.037[12]	0.073[09]	Average rainfall
Zone	0.035[09]	0.009[13]	0.003[15]	Water utility operational zone

Variables Are Grouped by Different Characteristics. Interval Between 0 and 1. Group of Variables Are Not Mutually Exclusive

6. DISCUSSION AND MANAGERIAL INSIGHTS

The results described above provide insights about the system's failure risks and can be used to inform decision-makers when to make other decisions related to system operations. For example, pertinent to our case study—*predicting sewer system failure risks in the city of Bogotá*, we learned that the optimal grid size to produce robust prediction models corresponds to 400×400 m^2. The size of a residential neighborhood in our study area ranges from 0.1 to 0.3 km^2 approximately; thus, the 0.16 km^2 of the selected cell size actually is placed within the neighborhood-size interval. This selected size of the cell results convenient for maintenance operations, which can be planned and scheduled based on the geographic delineation of the residential neighborhoods.

Second, identifying the factors that drive the risk of failure in the sewer system (i.e., the risk factors) provides important managerial insights. In particular, variables associated with the infrastructure age were found to be an important predictor of the failure risk. Additionally, it was also observed that the risk of failure is affected more by the factors related to the sanitary components rather than those related to the stormwater components. In general, the weather variables apparently have little influence on the risk of failure (Stage 1). However, the weather variables were found to be more important in predicting the severity of the risk of failure (high/low risk), in case the sewer system suffers a failure (Stage 2) (i.e., it acts as an aggravating factor, see Table 8). In this case, we also observed (from Stage 2) that the weather-related variables play a significant role in predicting the high/low failure risk when the study area spreads over a wider geographical region (i.e., considering both zones at the same time). The rationale is, as the geographical area expands, the gradient/variance of the weather-related variables also increases over that region. Thus, our results support the resemblance between the distribution of failures and precipitation, as described in section "Environmental factors."

6.1. Validation of Results and Illustrating Implications of the Results Using a Decision Support Tool

To address additional managerial questions and validate the results of our proposed two-stage risk prediction model, we now consider a real-world scenario in which a manager would leverage the risk prediction model to make risk-informed operational decisions. To do so, we applied our model to a new dataset, which composed of six additional months of data—the last three months in 2004 and the first three months in 2011, therefore allowing us to have data outside the time period of our analysis (i.e., a dataset that was not used for the training and testing of models). Besides validation, results from this example were used to develop a decision support system for risk-informed managerial decision-making.

To explain the applicability and usefulness of our two-stage model in risk-informed managerial decision making, the decision support system is designed in such a way that can incorporate the manager's

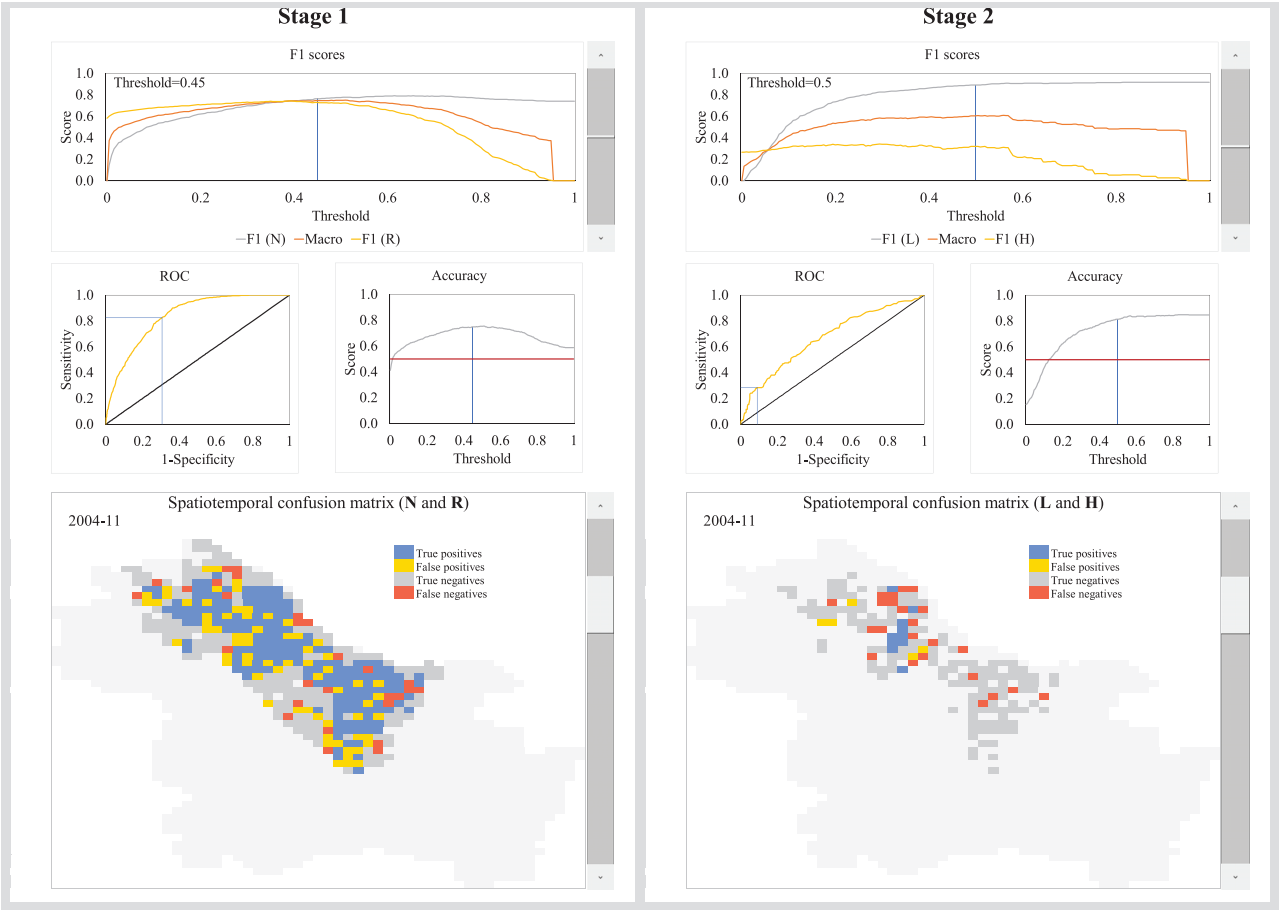


Fig 19. Decision support tool: dashboard with performance metrics as a function of the decision threshold, and a spatiotemporal confusion matrix showing the assertiveness of our model for Zone 2. See also Figs. A5 and A6.

preferences/knowledge/experience in terms of selection of the decision threshold for failure risk (**R**, **N**) as well as that for predicting the severity (low/high) of failure risk (**L**, **H**). The decision support system allows decision-makers to modify the discrimination threshold while showing a dashboard with the models' (Stages 1 and 2) performance metrics and permitting the user to select a particular month to understand the risk of failure and its severity based on a spatiotemporal confusion matrix. Using this system, managers can analyze the risk of failure and the severity of failure corresponding to different decision threshold values for different months as needed. Fig. 19 shows a screenshot of such a tool applied for Zone 2 sewer risk failure prediction.

Note that the developed decision support system helps decision-makers select an optimal decision threshold for predicting both failure risks (Stage 1) and risk severity (Stage 2) and indicates whether the maintenance planning should be conducted in a centralized or decentralized manner. For our case study, the results indicate that the prediction is considerably better when the zones are analyzed independently rather than when they are analyzed together (see Figs. 19, A5, and A6).

6.2. Optimization Exercise: An Illustrative Example Showing Implications of the Results

To show the applicability of our results in a real-world scenario and their importance in risk-informed decision-making for preventive maintenance planning, we advance a step forward and illustrate how the results/outputs from the prediction modeling can be used as the inputs of a prescriptive model.

In this example, the planning and scheduling of the maintenance operations in the sewer system involve two decisions: first, determining the distri-

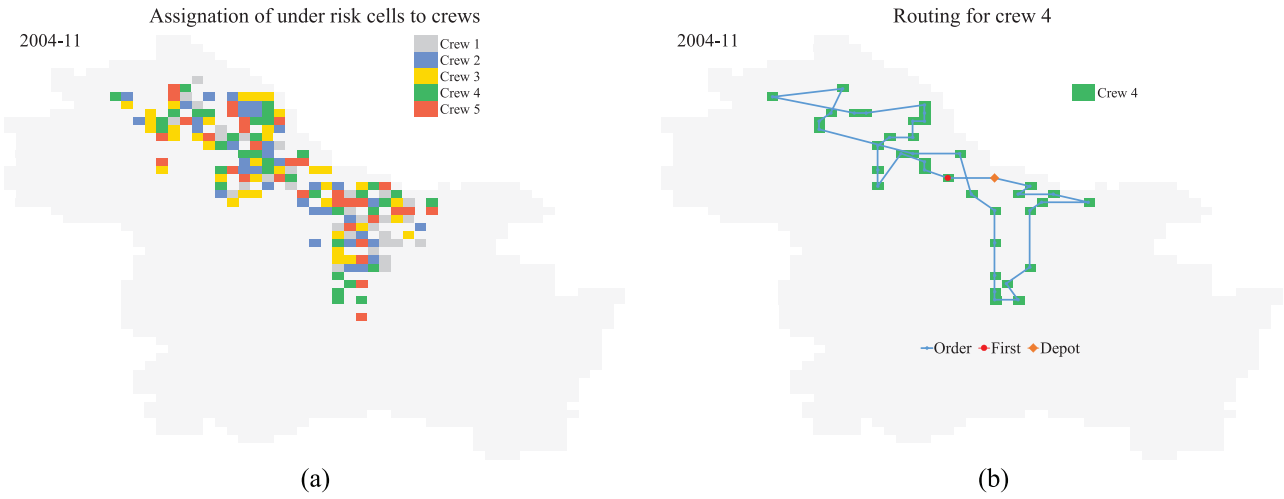


Fig 20. (a) Distribution of under-risk cells to crews. The distribution is made to balance the responsibility of each crew. (b) Service order for crew 4.

bution of tasks of the available crews (technicians and equipment) (i.e., which cells will be served for which crew); and second, identifying the route for each crew to conduct such maintenance operations. From the perspective of a manager, both decisions are to be made over a planning horizon aiming to: (1) balance the responsibility of each crew with respect to their pairs and (2) to minimize the traveled distance while guaranteeing that the cells with higher risk receive priority.

This problem is defined over a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \mathcal{N}_c \cup \{0\}$ represents the set of cells and the depot (where the crews are based and dispatched from). The set of cells $\mathcal{N}_c = \{1, 2, \ldots, n\}$ is geographically distributed and possess a risk level associated with the probability p_i of high risk (**H**) for each $i \in \mathcal{N}$. There is a set $\mathcal{R}$ of identical crews available to conduct the maintenance operations that are dispatched from a single depot (indexed by $i = 0$). The set $\mathcal{E} = \{\{i, j\} : i, j \in \mathcal{N}, j > i\}$ contains the edges connecting all cells. The travel distance d_e for each edge $e \in \mathcal{E}$ is deterministic and known. The goal is to: (1) define the distribution of cells per crew that maximizes the minimum of the reliability for which each crew is responsible, and (2) find a set of routes ζ over $\mathcal{G}$ that minimizes the total expected traveled distance while giving priority to those cells with a higher probability of high risk.

In a practical setting, one would expect the following optimization approach to be used dynamically using a rolling horizon. That is, the cell distribution and maintenance routes are generated via the optimization models using as input the risk-level probabilities at a given point in time. Then, after a predefined time window elapses and while possibly some maintenance tasks are still being executed, the optimization models are resolved using as input the updated failure risk probabilities for the new period. This produces a reoptimized set of routes that considers the spatiotemporal characteristics of the predicted data.

Generating this type of dynamic solutions is of high importance because the service time duration of some maintenance operations is long enough so that the completion of the entire set of tasks assigned to each crew may take multiple days, and the failure risk levels of some cells may change while the maintenance tasks of some others are still being completed. This particular consideration is of notable importance for the maintenance scheduling problem. However, given the illustrative purposes of this section, we limit our discussion to a *static model* to depict one of the many uses of the predicted failure risk probabilities to support the strategic management of sewer systems.

6.2.1. *A Model to Assign the Tasks for Each Crew*

Let $1 - p_i$ be the reliability of cell $i \in \mathcal{N}$, β be a parameter indicating the maximum accepted percentage of imbalance in the number of cells assigned to each crew, z be a variable that represents the minimum of the average reliability for which a crew is responsible for, x_{ij} be a variable that takes the value of

1 if cell $i \in \mathcal{N}$ is assigned to crew $j \in \mathcal{R}$, the following model represents the maximization of the minimum average reliability for which each crew is responsible for:

$$\max \quad z \tag{5}$$

$$\text{s.t.} \quad z \leq \frac{\sum_{i \in \mathcal{N}} (1 - p_i) x_{ij}}{\sum_{i \in \mathcal{N}} x_{ij}} \qquad \forall j \in \mathcal{R}, \tag{6}$$

$$\sum_{j \in \mathcal{R}} x_{ij} = 1 \qquad \forall i \in \mathcal{N}, \tag{7}$$

$$\sum_{i \in \mathcal{N}} x_{ij} \leq \frac{|\mathcal{N}|}{|\mathcal{R}|}(1 + \beta) \qquad \forall j \in \mathcal{R}, \tag{8}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i \in \mathcal{N},\ j \in \mathcal{R}. \tag{9}$$

$$z \geq 0 \tag{10}$$

The expression (5) maximizes the minimum average reliability for each crew. The set of constraints (6) traps the minimum of the average reliability among the crews. The set of constraints (7) guarantees that each cell is assigned to one crew. The set of constraints (8) guarantees that the number of cells assigned to each crew will be imbalanced at most for β percent. Constraints (9) and (10) represent the nature of the variables. It is worth noting that the ratios present in constraints 6 can be linearized by traditional methods (Wolsey, 2020).

6.2.2. A Model to Route Crews

With the previous model's output, we will have the distribution of cells that should be served for each crew. With this at hand, another optimization model can be built to define the order in which the cells must be served per the crews. One additional goal from the manager's perspective is the minimization of the traveled distance (as it consumes resources) while prioritizing those cells with a higher risk of sewer system failure. To do so, let us first define an auxiliary parameter $c_e = d_e(1 - p_i)(1 - p_j)$ where i and j represent the two cells connected by edge e. With this at hand, we formulate a traveling salesman problem (TSP). Here, we lend the TSP to define the ordering in the service for the cells. The TSP is a well-studied problem in the operations research literature and several formulations and methods have been developed to solve it. For the purpose of this project, we used the symmetric version of the Dantzig–Fulkerson–Johnson formulation (Applegate, Bixby, Chvatal, & Cook, 2006).

6.2.3. Optimization Results

To build this exercise, some additional information such as the number of available crews to perform maintenance was needed and collected from the previous work by Fontecha et al. (2020). With this information at hand and the results of our prediction modeling in Stage 2 for November 2004 (as shown in Fig. 19), we first assigned the cells that should be served for each crew (see Section 6.2.1), and then we identified the order in which each of the cells needs to be served based on their probability of high risk (see Section 6.2.2).

Fig. 20(a) presents the distribution of cells per crew. Note that the distribution, in this case, was made in such a way that the risk/reliability for each of the responsible crews is balanced (i.e., every crew is responsible for a similar number of cells and for a similar total value of risk). Fig. 20(b) shows the order in which the cells should be served by crew 4, as an example. Note that the distance itself is not the minimum—this is because cells with a higher probability for a high risk are prioritized.

7. CONCLUSIONS

In this article, we proposed a novel two-stage data-driven framework to predict the risk of sewer system failures while considering spatiotemporal data and intrinsic data imperfections (i.e., imbalanced data, missing values, and outliers). We tested the performance of our methodology by predicting the risk of sediment-related failures in two out of the five operational zones of Bogotá (Colombia). The results obtained validate the capabilities of our framework to analyze the failure risk of large-scale infrastructure systems with limited data while providing key managerial insights about the system. Furthermore, we provided a decision support tool that uses a dashboard to show the performance of the predictive models under different discrimination thresholds. Finally, we developed an illustrative optimization example to use the results from the data-driven risk assessment predictive models to plan and schedule the maintenance operations in the study area.

Our framework has proven to be flexible, providing more versatility to utility managers to decide how maintenance should be carried out with respect to grid size, zone scenarios, discrimination threshold, and performance metric selection in the prediction models. Also, the failure risk prediction model proved to be robust, validating its performance on a completely new dataset that was not used for training the algorithms.

The sensitivity analysis developed for the case study showed their usefulness to aid managers in identifying whether a decentralized or centralized management of the failure risk in sewer system infrastructure is more adequate or not. Although weather variables showed to have no influence on the risk of failure, once the system fails, they do contribute to increasing the severity of the risk (i.e., changing the risk level from low to high risk). The exercise of identifying which cell size renders the best predictive accuracy for the models also revealed that the managerial decisions can be taken using residential neighborhoods, whose area is similar to that covered by a cell. In that sense, the selection of the size of the cell can be crucial not only for the stability and robustness of the prediction models but also for the decision-making process in practice and for the planning of the maintenance operations.

This article also exemplifies how our results can assist operational decisions and how the insights derived from our analysis can support strategic decisions. The classification of zones (cells or areas) into risk levels could be used to prioritize the sewer sections that require engineering interventions to increase their reliability. Examples of such interventions are expansions of the sewer system capacity or replacements of piped sections with self-cleansing sewer pipes that guarantee sediment transport (Montes, Berardi, Kapelan, & Saldarriaga, 2020; Montes, Kapelan, & Saldarriaga, 2019). Alternatively, Sustainable Urban Drainage Systems (Ghodsi, Zahmatkesh, Goharian, Kerachian, & Zhu, 2020; Torres, Fontecha, Zhu, Walteros, & Rodríguez, 2020) can be placed to alleviate the sediments load from runoff (Maringanti, Chaubey, & Popp, 2009). From a larger perspective, results presented herein constitute a building block to the planning of urban renewal interventions. The latter, in synchrony with other urban structures analyses (e.g., trees, highways, cycling paths, or sidewalks [Rodriguez-Valencia, Barrero, Ortiz-Ramirez, & Vallejo-Borda, 2020; Vallejo-Borda, Cantillo, & Rodriguez-Valencia, 2020]), represent an opportunity for urban renewal, while decreasing the risk of failure and improving the perception of sewer system quality of service (Vallejo-Borda, Ortiz-Ramirez, Rodriguez-Valencia, Hurtubia, & de D. Ortúzar, 2020).

Finally, our proposed framework can also be applied for data-driven preventive maintenance of other public facilities such as roads, bridges, electricity transmission lines, and water supply lines by providing valuable information about location, time, and type of failures that are likely to happen. For example, in the case of maintenance of roads and highways, the presented framework can be leveraged using volumes of available data that are based on several factors such as climatic variables, structural characteristics, traffic load, age, and drainage condition to provide answers related to different parts of the transportation system such as when to carry out maintenance, what resources are required, and how to carry out maintenance for maximum efficiently at reduced costs.

ACKNOWLEDGMENTS

We would like to thank Bogotá's water utility (EAAB) for providing the sewer information and the consumer complaints database. In particular, we would like to acknowledge the help of Engineer Daniel Rodríguez, from EAAB, in providing guidance and feedback through the development of this project. We thank Bogotá's meteorology institute (IDEAM), urban planning secretariat (SDP), and planting authority (JBB) for granting access to their databases. We thank scikit-learn and imbalanced-learn APIs for providing the access to the necessary machine learning tools used through this study. Finally, we would like to thank Gurobi for providing us with an academic license of their linear optimizer.

AUTHOR CONTRIBUTIONS

- John E. Fontecha: Conceptualization, Methodology, Software, Validation, Formal Analysis, Data curation, Writing—Original Draft, Visualization.
- Puneet Agarwal: Methodology, Software, Formal Analysis, Writing—Original Draft.
- María N. Torres: Conceptualization, Formal Analysis, Data curation, Writing - Original Draft, Visualization.
- Sayanti Mukherjee: Methodology, Writing Review, Writing Editing, Supervision.

- Jose L. Walteros: Formal Analysis, Writing Editing, Supervision, Funding acquisition.
- Juan P. Rodríguez: Conceptualization, Resources, Writing Review.

REFERENCES

Agarwal, P., Tang, J., Narayanan, A. N. L., & Zhuang, J. (2020). Big data and predictive analytics in fire risk using weather data. *Risk Analysis*, *40*(7), 1438–1449.

Alipour, P., Mukherjee, S., & Nateghi, R. (2019). Assessing climate sensitivity of peak electricity load for resilient power systems planning and operation: A study applied to the Texas region. *Energy*, *185*, 1143–1153.

Allison, P. D. (2002). *Missing data*, Vol. Quantitative Applications in the Social Sciences, *136*. Thousand Oaks, CA: Sage Publications.

Allouche, E. N., & Freure, P. (2002). *Management and maintenance practices of storm and sanitary sewers in Canadian municipalities*. ICLR Research, Paper Series, *18*(1–52). London, ON: Dept. of Civil and Environmental Engineering, Univ. of Western Ontario.

Ana, E., Bauwens, W., Pessemier, M., Thoeye, C., Smolders, S., Boonen, I., & De Gueldre, G. (2009). An investigation of the factors influencing sewer structural deterioration. *Urban Water Journal*, *6*(4), 303–312.

Anbari, M. J., Tabesh, M., & Roozbahani, A. (2017). Risk assessment model to prioritize sewer pipes inspection in wastewater collection networks. *Journal of Environmental Management*, *190*, 91–101.

Applegate, D. L., Bixby, R. E., Chvatal, V., & Cook, W. J. (2006). *The traveling salesman problem: A computational study*. Princeton, NJ: Princeton University Press.

Baah, K., Dubey, B., Harvey, R., & McBean, E. (2015). A risk-based approach to sanitary sewer pipe asset management. *Science of the Total Environment*, *505*, 1011–1017.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Caradot, N., Riechel, M., Fesneau, M., Hernandez, N., Torres, A., Sonnenberg, H., Rouault, P. (2018). Practical benchmarking of statistical and machine learning models for predicting the condition of sewer pipes in Berlin, Germany. *Journal of Hydroinformatics*, *20*(5), 1131–1147.

Carvalho, G., Amado, C., Brito, R. S., Coelho, S. T., & Leitão, J. P. (2018). Analysing the importance of variables for sewer failure prediction. *Urban Water Journal*, *15*(4), 338–345.

Chapman, B., Jost, G., & Van Der Pas, R. (2008). *Using OpenMP: Portable shared memory parallel programming*, Vol. *10*. Cambridge: MIT press.

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.

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). ACM.

Chen, Z., Jiang, F., Cheng, Y., Gu, X., Liu, W., & Peng, J. (2018). XGBoost classifier for DDoS attack detection and analysis in SDN-Based cloud. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 251–256). IEEE.

Chughtai, F., & Zayed, T. (2008). Infrastructure condition prediction models for sustainable sewer pipelines. *Journal of Performance of Constructed Facilities*, *22*(5), 333–341.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359.

Duchesne, S., Beardsell, G., Villeneuve, J.-P., Toumbou, B., & Bouchard, K. (2013). A survival analysis model for sewer pipe structural deterioration. *Computer-Aided Civil and Infrastructure Engineering*, *28*(2), 146–160.

Duran, O., Althoefer, K., & Seneviratne, L. D. (2002). State of the art in sensor technologies for sewer inspection. *IEEE Sensors Journal*, *2*(2), 73–81.

Egger, C., Scheidegger, A., Reichert, P., & Maurer, M. (2013). Sewer deterioration modeling with condition data lacking historical records. *Water Research*, *47*(17), 6762–6779.

Elmasry, M., Hawari, A., & Zayed, T. (2017). Defect based deterioration model for sewer pipelines using Bayesian belief networks. *Canadian Journal of Civil Engineering*, *44*(9), 675–690.

Empresa de Acueducto y Alcantarillado de Bogotá. (2015). Delimitación por zonas de servicio.

Ezell, B. C. (2007). Infrastructure vulnerability assessment model (I-VAM). *Risk Analysis*, *27*(3), 571–583.

Faisal, S., & Tutz, G. (2017). *Nearest neighbor imputation for categorical data by weighting of attributes*. arXiv preprint arXiv:1710.01011.

Fontecha, J. E., Akhavan-Tabatabaei, R., Duque, D., Medaglia, A. L., Torres, M. N., & Rodríguez, J. P. (2016). On the preventive management of sediment-related sewer blockages: A combined maintenance and routing optimization approach. *Water Science and Technology*, *74*(2), 302–308.

Fontecha, J. E., Guaje, O. O., Duque, D., Akhavan-Tabatabaei, R., Rodríguez, J. P., & Medaglia, A. L. (2020). Combined maintenance and routing optimization for large-scale sewage cleaning. *Annals of Operations Research*, *286*, 441–474.

Franke, R. (1982). Scattered data interpolation: Tests of some methods. *Mathematics of Computation*, *38*(157), 181–200.

Ghodsi, S. H., Zahmatkesh, Z., Goharian, E., Kerachian, R., & Zhu, Z. (2020). Optimal design of low impact development practices in response to climate change. *Journal of Hydrology*, *580*, 124266.

Hahn, M. A., Palmer, R. N., Merrill, M. S., & Lukas, A. B. (2002). Expert system for prioritizing the inspection of sewers: Knowledge base formulation and evaluation. *Journal of Water Resources Planning and Management*, *128*(2), 121–129.

Harvey, R. R., & McBean, E. A. (2014). Predicting the structural condition of individual sanitary sewer pipes with random forests. *Canadian Journal of Civil Engineering*, *41*(4), 294–303.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer Science & Business Media.

Hernández, N., Caradot, N., Sonnenberg, H., Rouault, P., & Torres, A. (2018). Support tools to predict the critical structural condition of uninspected pipes for case studies of Germany and Colombia. *Water Practice & Technology*, *13*(4), 794–802.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*, Vol. *112*. Boston: Springer.

Jiang, G., Keller, J., Bond, P. L., & Yuan, Z. (2016). Predicting concrete corrosion of sewers using artificial neural network. *Water Research*, *92*, 52–60.

Jin, Y., & Mukherjee, A. (2010). Modeling blockage failures in sewer systems to support maintenance decision making. *Journal of Performance of Constructed Facilities*, *24*(6), 622–633.

Jung, Y. (2018). Multiple predicting k-fold cross-validation for model selection. *Journal of Nonparametric Statistics*, *30*(1), 197–215.

Kabir, G., Balek, N. B. C., & Tesfamariam, S. (2018). Consequence-based framework for buried infrastructure systems: A Bayesian belief network model. *Reliability Engineering & System Safety*, *180*, 290–301.

Karnieli, A. (1990). Application of kriging technique to areal precipitation mapping in Arizona. *GeoJournal*, *22*(4), 391–398.

Kleidorfer, M., Möderl, M., Tscheikner-Gratl, F., Hammerer, M., Kinzel, H., & Rauch, W. (2013). Integrated planning of rehabilitation strategies for sewers. *Water Science and Technology*, *68*(1), 176–183.

Korving, H., & Van Noortwijk, J. (2008). Bayesian updating of a prediction model for sewer degradation. *Urban Water Journal*, *5*(1), 51–57.

Korving, H., Van Noortwijk, J. M., Van Gelder, P. H., & Clemens, F. H. (2009). Risk-based design of sewer system rehabilitation. *Structure and Infrastructure Engineering*, *5*(3), 215–227.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232.

Kuliczkowska, E. (2016). Risk of structural failure in concrete sewers due to internal corrosion. *Engineering Failure Analysis*, *66*, 110–119.

Laakso, T., Kokkonen, T., Mellin, I., & Vahala, R. (2018). Sewer condition prediction and analysis of explanatory factors. *Water*, *10*(9), 1239.

Lameski, P., Zdravevski, E., Mingov, R., & Kulakov, A. (2015). SVM parameter tuning with grid search and its impact on reduction of model over-fitting. In *Proceedings of the Intertnational Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing* (pp. 464–474). Springer.

Lavanya, D., & Rani, U. (2012). Ensemble decision tree classifier for breast cancer data. *International Journal of Information Technology Convergence and Services*, *2*(1), 17–24.

Le Gat, Y. (2008). Modelling the deterioration process of drainage pipelines. *Urban Water Journal*, *5*(2), 97–106.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1–5.

Li, L., Wang, J., Leung, H., & Jiang, C. (2010). Assessment of catastrophic risk using Bayesian network constructed from domain knowledge and spatial data. *Risk Analysis*, *30*(7), 1157–1175.

López-Kleine, L., Hernández, N., & Torres, A. (2016). Physical characteristics of pipes as indicators of structural state for decision-making considerations in sewer asset management. *Ingeniería e Investigación*, *36*(3), 15–21.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316.

Mani, I., & Zhang, I. (2003). kNN approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*, Vol. *126*.

Maringanti, C., Chaubey, I., & Popp, J. (2009). Development of a multiobjective optimization tool for the selection and placement of best management practices for nonpoint source pollution control. *Water Resources Research*, *45*(6), 1–15.

Mashford, J., Marlow, D., Tran, D., & May, R. (2011). Prediction of sewer condition grade using support vector machines. *Journal of Computing in Civil Engineering*, *25*(4), 283–290.

McDonald, S., & Zhao, J. (2001). Condition assessment and rehabilitation of large sewers. In *Proceedings of International Conference on Underground Infrastructure Research* (pp. 361–369). Citeseer.

Micevski, T., Kuczera, G., & Coombes, P. (2002). Markov model for storm water pipe deterioration. *Journal of Infrastructure Systems*, *8*(2), 49–56.

Montes, C., Berardi, L., Kapelan, Z., & Saldarriaga, J. (2020). Predicting bedload sediment transport of non-cohesive material in sewer pipes using evolutionary polynomial regression–multi-objective genetic algorithm strategy. *Urban Water Journal*, *17*, 154–162.

Montes, C., Kapelan, Z., & Saldarriaga, J. (2019). Impact of self-cleansing criteria choice on the optimal design of sewer networks in South America. *Water*, *11*(6), 1148.

Montes, C., Vanegas, S., Kapelan, Z., Berardi, L., & Saldarriaga, J. (2020). Non-deposition self-cleansing models for large sewer pipes. *Water Science and Technology*, *81*(3), 606–621.

Mukherjee, S., & Nateghi, R. (2017). Climate sensitivity of end-use electricity consumption in the built environment: An application to the state of Florida, United States. *Energy*, *128*, 688–700.

Mukherjee, S., Vineeth, C. R., & Nateghi, R. (2019). Evaluating regional climate-electricity demand nexus: A composite Bayesian predictive framework. *Applied Energy*, *235*(2019), 1561–1582.

Mukherjee, S., & Nateghi, R. (2019). A data-driven approach to assessing supply inadequacy risks due to climate-induced shifts in electricity demand. *Risk Analysis*, *39*(3), 673–694.

Mukherjee, S., Nateghi, R., & Hastak, M. (2018). A multi-hazard approach to assess severe weather-induced major power outage risks in the US. *Reliability Engineering & System Safety*, *175*, 283–305.

Obringer, R., Mukherjee, S., & Nateghi, R. (2020). Evaluating the climate sensitivity of coupled electricity-natural gas demand using a multivariate framework. *Applied Energy*, *262*(114419), 1–11.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Post, J., Pothof, I., ten Veldhuis, M.-C., Langeveld, J., & Clemens, F. (2016). Statistical analysis of lateral house connection failure mechanisms. *Urban Water Journal*, *13*(1), 69–80.

Rodríguez, J. P., McIntyre, N., Díaz-Granados, M., & Maksimović, Č. (2012). A database and model to support proactive management of sediment-related sewer blockages. *Water Research*, *46*(15), 4571–4586.

Rodriguez-Valencia, A., Barrero, G. A., Ortiz-Ramirez, H. A., & Vallejo-Borda, J. A. (2020). Power of user perception on pedestrian quality of service. *Transportation Research Record*, *2674*, 250–258.

Roehrdanz, P. R., Feraud, M., Lee, D. G., Means, J. C., Snyder, S. A., & Holden, P. A. (2017). Spatial models of sewer pipe leakage predict the occurrence of wastewater indicators in shallow urban groundwater. *Environmental Science & Technology*, *51*(3), 1213–1223.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326.

Salman, B., & Salem, O. (2012a). Modeling failure of wastewater collection lines using various section-level regression models. *Journal of Infrastructure Systems*, *18*(2), 146–154.

Salman, B., & Salem, O. (2012b). Risk assessment of wastewater collection lines using failure models and criticality ratings. *Journal of Pipeline Systems Engineering and Practice*, *3*(3), 68–76.

Santos, P., Amado, C., Coelho, S. T., & Leitão, J. P. (2017). Stochastic data mining tools for pipe blockage failure prediction. *Urban Water Journal*, *14*(4), 343–353.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference* (pp. 517–524).

Shortridge, J., & Camp, J. S. (2019). Addressing climate change as an emerging risk to infrastructure systems. *Risk Analysis*, *39*(5), 959–967.

Sirkiä, J., Laakso, T., Ahopelto, S., Ylijoki, O., Porras, J., & Vahala, R. (2017). Data utilization at finnish water and wastewater utilities: Current practices vs. state of the art. *Utilities Policy*, *45*, 69–75.

Soley-Bori, M. (2013). *Dealing with missing data: Key assumptions and methods for applied analysis*. Technical Report, Boston University.

Soriano-Pulido, E., Valencia-Arboleda, C., & Rodríguez Sánchez, J. P. (2019). Study of the spatiotemporal correlation between sediment-related blockage events in the sewer system in Bogotá (Colombia). *Water Science and Technology*, *79*(9), 1727–1738.

Sousa, V., Matos, J. P., & Matias, N. (2014). Evaluation of artificial intelligence tool performance and uncertainty for predicting sewer structural condition. *Automation in Construction*, *44*, 84–91.

Terti, G., Ruin, I., Gourley, J. J., Kirstetter, P., Flamig, Z., Blanchet, J., Arthur, A., & Anquetin, S. (2019). Toward probabilistic prediction of flash flood human impacts. *Risk Analysis*, *39*(1), 140–161.

Torres, M. N., Fontecha, J. E., Zhu, Z., Walteros, J. L., & Rodríguez, J. P. (2020). A participatory approach based on stochastic optimization for the spatial allocation of sustainable urban drainage systems for rainwater harvesting. *Environmental Modelling & Software*, *123*, 104532.

Torres, M. N., Rodríguez, J. P., & Leitao, J. P. (2017). Geostatistical analysis to identify characteristics involved in sewer pipes and urban tree interactions. *Urban Forestry & Urban Greening*, *25*, 36–42.

Tran, D. H., Ng, A., & Perera, B. (2007). Neural networks deterioration models for serviceability condition of buried stormwater pipes. *Engineering Applications of Artificial Intelligence*, *20*(8), 1144–1151.

Tscheikner-Gratl, F., Caradot, N., Cherqui, F., Leitão, J. P., Ahmadi, M., Langeveld, J. G., Clemens, F. (2019). Sewer asset management–state of the art and research needs. *Urban Water Journal*, *16*(9), 662–675.

Ugarelli, R., Kristensen, S. M., Røstum, J., Sægrov, S., & Di Federico, V. (2009). Statistical analysis and definition of blockages-prediction formulae for the wastewater network of Oslo by evolutionary computing. *Water Science and Technology*, *59*(8), 1457–1470.

Vallejo-Borda, J. A., Cantillo, V., & Rodriguez-Valencia, A. (2020). A perception-based cognitive map of the pedestrian perceived quality of service on urban sidewalks. *Transportation Research Part F: Traffic Psychology and Behaviour*, *73*, 107–118.

Vallejo-Borda, J. A., Ortiz-Ramirez, H. A., Rodriguez-Valencia, A., Hurtubia, R., & de D.Ortúzar, J. D. D. (2020). Forecasting the quality of service of Bogota's sidewalks from pedestrian perceptions: An ordered probit MIMIC approach. *Transportation Research Record*, *2674*(1), 205–216.

Vasan, K. K., & Surendiran, B. (2016). Dimensionality reduction using principal component analysis for network intrusion detection. *Perspectives in Science*, *8*, 510–512.

Wang, F., Lan, M., & Wu, Y. (2017). ECNU at SemEval-2017 Task 8: Rumour evaluation using effective features and supervised ensemble models. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 491–496).

Wolsey, L. A. (2020). *Integer programming*. New York: John Wiley & Sons.

Yang, X., Xie, X., Liu, D. L., Ji, F., & Wang, L. (2015). Spatial interpolation of daily rainfall data for local climate impact assessment over greater Sydney region. *Advances in Meteorology*, *2015*, 1–12.

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In *Proceedings of the 1st International Conference on Advanced Data and Information Engineering (DaEng-2013)* (pp. 13–22). Springer.

Younis, R., & Knight, M. A. (2010). A probability model for investigating the trend of structural deterioration of wastewater pipelines. *Tunnelling and Underground Space Technology*, *25*(6), 670–680.

Yuan, Y. C. (2010). *Multiple imputation for missing data: Concepts and new development (Version 9.0)*. Technical report, SAS Institute Inc, Rockville, MD.

Zhang, D., Wang, J., & Zhao, X. (2015). Estimating the uncertainty of average F1 scores. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (pp. 317–320).

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. A1. Correlation matrix for the 222 variables. Blue indicates a correlation under 0.7, and red indicates otherwise.

Fig. A2. Distribution of sediment-related failures over the spatiotemporal dimension on a grid of 200×200 m^2. Press play to see the distribution.

Fig. A3. Distribution of sediment-related failures over the spatiotemporal dimension on a grid of 400×400 m^2. Press play to see the distribution.

Fig. A4. Distribution of sediment-related failures over the spatiotemporal dimension on a grid of 800×800 m^2. Press play to see the distribution.

Fig. A5. Decision support tool: dashboard with performance metrics as a function of the decision threshold, and a spatiotemporal confusion matrix showing the assertiveness of our model for Zone 3.

Fig. A6. Decision support tool: dashboard with performance metrics as a function of the decision threshold, and a spatiotemporal confusion matrix showing the assertiveness of our model for both zones together.

Table B1. Name, Description, and General Information of the Variables used in Our Analysis