Water Resources Management
https://doi.org/10.1007/s11269-023-03676-w

Cluster Analysis and Predictive Modeling of Urban Water Distribution System Leaks with Socioeconomic and Engineering Factors

Qing Shuang[1] · **Rui Ting Zhao**[1] · **Erik Porse**[2,3]

Received: 26 March 2021 / Accepted: 17 November 2023

Abstract
Water distribution network (WDN) failures can disrupt operations and cause economic damage. Although leakage has been widely discussed, few studies have integrated spatial clusters with engineering, environmental, and socioeconomic factors simultaneously. This study proposes an approach to explore the role of socioeconomic factors in understanding leak risks. Using a unique data set of more than 4,000 reported leak events within the City of Los Angeles (2010–2013), the analysis (1) assesses the effectiveness of including socioeconomic factors with engineering factors in explaining observed leaks, (2) identifies spatial clusters of leaks, and (3) develops a predictive model with machine learning to identify spatial areas with high risks of failure. Results indicate that distinct clusters of leaks are evident, accounting for 20–30% of all leaks in the study area in a given year. Multivariate regression modeling showed that geography, socioeconomic, and engineering factors are statistically significant in predicting leaks. A predictive model with machine learning was developed, identifying key factors. The model had accuracy rates of 93.29% and 92.45% for interpolation and extrapolation prediction scenarios, respectively. The approach demonstrates the potential value of incorporating socioeconomic indicators into the models for WDN rehabilitation. Moreover, the approach demonstrates how municipal leak loss mitigation programs can consider a broad set of predictive factors to optimize investments.

Keywords Water distribution network · Leakage cluster · Machine learning · Urban water management

✉ Qing Shuang
qings@bjtu.edu.cn

Rui Ting Zhao
19120651@bjtu.edu.cn

Erik Porse
erik.porse@owp.csus.edu; eporse@ioes.ucla.edu

[1] School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China

[2] Office of Water Programs, California State University, 6000 J Street, 95819-6025 Sacramento, CA, USA

[3] Institute of the Environment and Sustainability, University of California, Los Angeles, 619 Charles E. Young Dr. East, La Kretz Hall, Suite 300, Los Angeles, CA 90095-1496, USA

Published online: 02 December 2023

1 Introduction

Critical infrastructures, such as water distribution networks (WDNs), are extremely important to urban communities. In many established cities with aging infrastructure, water losses in the underground pipes of WDNs present a management challenge (Robles-Velasco et al. 2023). Approximately 0.24 million water main breaks occur in the U.S. every year because of the aging WDNs, with losses of more than two trillion gallons of treated drinking water (American Society of Civil Engineers 2017). Considerable investments are needed to uncover and rehabilitate WDNs (Abokifa and Sela 2023). Some urban water utilities cope with future water demands to meet population growth and limited new supply sources. Hence, effort to reduce leak loss are an increasingly popular component of supply and demand management options.

Leaks in WDNs affect utilities and residents. Utilities recover the production costs of these losses, including system repairs, through other means. Leaks have operational impacts on WDNs, where they can reduce system pressure and lead to frequent future supply disruptions. Property damages from catastrophic leaks may incur economic losses for liable utilities and affected residents and businesses. In some cases, leaks can even require utilities to issue public health advisories to boil water before consumption.

Developing trustworthy predictive models requires significant effort to identify factors that affect leakage. Over the years, many potential leakage-causing factors have been assessed using indirect (e.g., water audits) and direct (e.g., gas injection, acoustic monitoring, and visual inspections) techniques (Hunaidi 2006; Misiūnas 2008). These factors can be roughly categorized into two groups: engineering and environment (Barton et al. 2019). Engineering factors include system pressure, pipe age, pipe material, soil type, and system configuration. Meanwhile, environmental factors include temperature and precipitation, which also have an impact on WDN leaks.

In cities, human decisions regarding the design, layout, and operational requirements of infrastructure systems are driven by engineering and socioeconomic factors (Swilling 2011). Hence, infrastructure systems are known to be products of technical design requirements and political and socioeconomic influences (Rahimi-Golkhandan et al. 2022). The engineering and environmental factors that influence the occurrence of urban leaks are well understood. However, the potential socioeconomic influences of leak occurrence, such as the density of connections or income trends, have not been explored. WDNs are infrastructure networks developed by organizations to meet societal needs. Hence, we can consider them as products of institutional planning and technical requirements.

This study analyzes leak failure risks in urban WDNs by incorporating engineering and socioeconomic factors to understand the prevalence of observed leaks in a metropolitan water system. The analysis addresses three questions. First, are socioeconomic and geographic characteristics explanatory factors of observed leaks? Second, can spatial clusters of leaks be observed in non-random patterns across a metropolitan WDN? Third, can predictive models of leak risk be developed that include engineering and socioeconomic factors? To address these questions, we employ multiple linear regression (MLR) to assess the validity of incorporating socioeconomic indicators as explanatory factors explaining leak occurrence. Then, we analyze the spatial high-risk clustering in observed leaks. The results of these procedures are used to develop a predictive model with machine learning (ML), which includes engineering and socioeconomic drivers of urban water leak risks. The procedures are implemented for a case study of the City of Los Angeles, using a unique data set of over 4,000 reported water supply pipe leaks for 4 years (2010–2013). The study

concludes with a discussion of how the results can inform a broader perspective on leak loss mitigation programs in cities to improve utility operations and provide design guidance for urban planners.

2 Methods

Three modeling-related procedures—MLR, scan statistic approach, and ML technique—were established to investigate the influence of socioeconomic factors in predicting leak loss events. Figure 1 shows a summary of the modeling approach, data, and workflow.

2.1 Study Area

LA City is home to four million people. The city is situated in a large coastal basin and is surrounded by mountains, covering approximately 4,000 square miles. LA City has a municipally owned utility, the Los Angeles Department of Water and Power (LADWP), which supplies water and electricity services to residents.

Water infrastructure failures have become an issue of public concern for LADWP partly because of multiple prominent water main breaks that resulted in local surface flooding. Within the service territory of LADWP, more than 30% of water supply mains are over 80 years old. Moreover, approximately 6% of the 6,800-mile water supply distribution pipes are classified as a high priority for replacement (LADWP 2023).

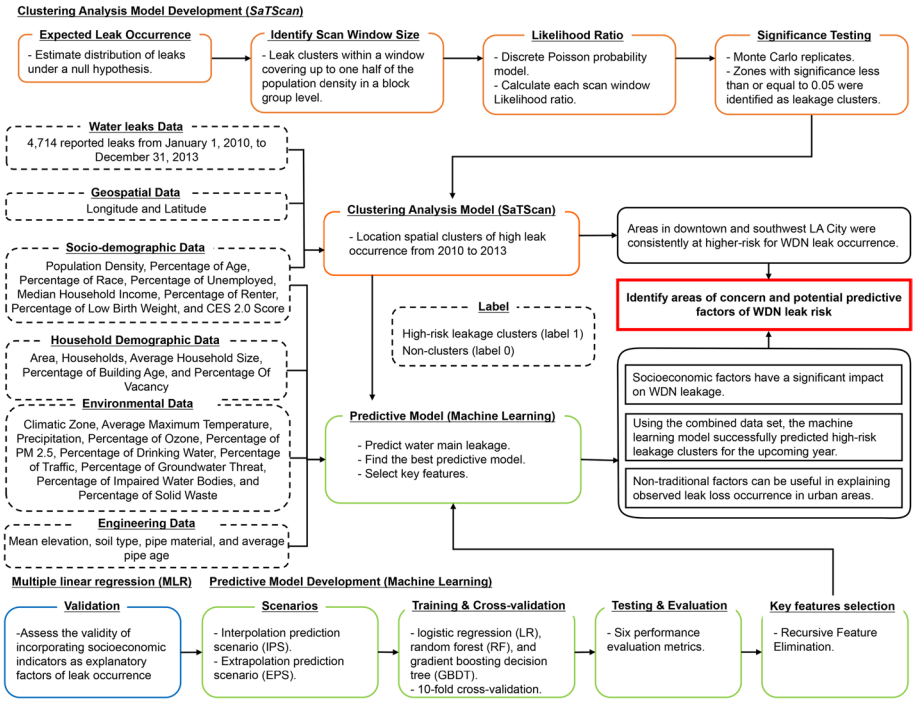

Fig. 1 Analysis procedure

The degraded water pipes in LA City have caused serious failures and localized surface flooding. The problem of reported leak events is of high importance owing to the public scrutiny caused by the leaks and the city's noted water management challenges for supplying water to 4 million residents in a seasonally dry climate. Water managers estimated the annual combined water losses from leakage, firefighting, evaporation, and evapotranspiration to be 30 billion liters, equivalent to the annual supply for up to 50,000 households in the region (Poston and Stevens 2015). In addition, leaks have caused property damage (Reyes-Velarde 2018). Hence, LA City had to spend one billion USD to repair damage following flooding or ruptures (Morrison 2021).

2.2 Data Sources

Publicly available data for historic leak occurrences in LA City were extracted from a published online mapping application from the *Los Angeles Times* (Poston and Stevens 2015). From available data, a total of 4,174 occurrences of leaks in LA City between 2010 and 2013 were obtained. Each record, mapped as a point, contained the address, date of the reported leak (day/month/year), pipe age, leak type, and pipe material (Fig. 2). Leak records were imported into ArcGIS and mapped with LA City boundaries.

Previous studies on pipe failures have shown that environmental and engineering factors can influence WDN leaks (Barton et al. 2019). However, WDNs and other infrastructures are examples of socio-technical systems. Studies on hazards and disasters have indicated that social vulnerability and community resilience should be considered when evaluating the impact of infrastructure failures (Fan et al. 2022; Rahimi-Golkhandan et al. 2022; Da Silveira and Mata-Lima 2021; Hani et al. 2023). Therefore, data from multiple sources, including leak occurrences, socioeconomic characteristics, land use, engineering parameters, and climate, were integrated as explanatory factors used in the statistical and predictive procedures in this study. Potential explanatory factors and their corresponding data sources were described as follows:

1. *Geographic factors*: longitude and latitude. Data were acquired from LA County's geospatial database (LA County 2023).
2. *Socioeconomic factors* constituted by (1) *household demographic variables*: area, households, average household size, percentage of building age by constructed year, and percentage of vacancy. The building age was divided into three groups: constructed before 1959, during 1960–1999, and after 2000; and (2) *sociodemographic variables*: population density, percentages of residents by age groups, percentage of residents by race and ethnicity, unemployment rate, median household income, percentage of renters, percentage of low birth weight, and *CalEnviroScreen (CES)* 2.0 score. The percentage of age was divided into three groups: age under 18, 18–64, and over 65. Race and ethnicity were divided into three groups: white, black and African American, and Hispanic or Latino.

   Socioeconomic data were collected from the American Community Survey (U.S. Census Bureau 2023) and the California Office of Environmental Health Hazard Assessment (OEHHA 2022). The *CES* 2.0 score is a summary index of social and environmental vulnerability for all census tracts in California.
3. *Environmental factors*: climatic zone, average maximum temperature, precipitation, and percentage for relative prevalence of ozone, Particulate Matter 2.5 (PM2.5), drinking

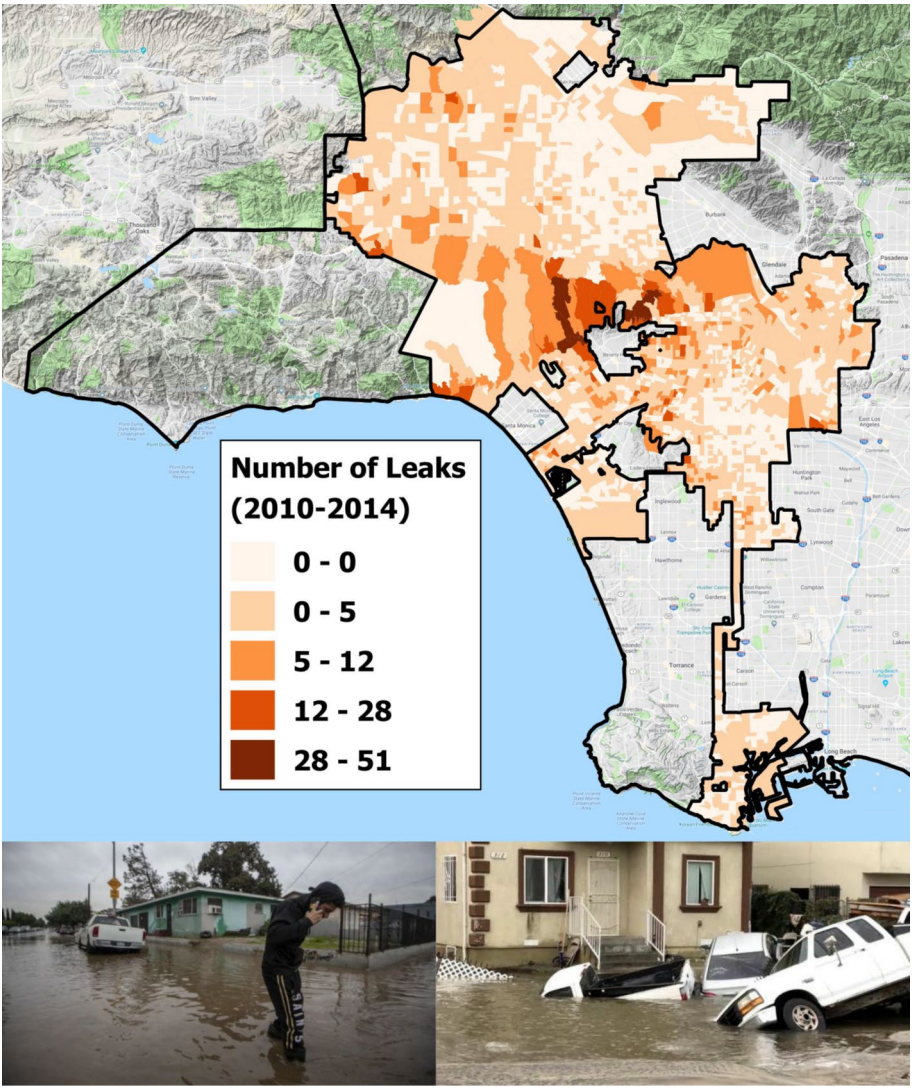

Fig. 2 Reported leak events in the City of Los Angeles. Shaded regions indicate municipal territories. A portion of the larger Los Angeles County boundary is also shown on the left of the map. (Data sources: Krishnakumar and Poston (2016); Poston and Stevens (2015); Picture source: Reyes-velarde (2018); Map created by authors)

water quality, traffic, threats to groundwater as a drinking water source, impaired water bodies, and proximity and exposure of solid waste.

Environmental data were obtained from two sources: local weather monitoring stations and climate zone designations from the California Energy Commission (CEC 2021) and local climate data that are contained within UCLA's *Energy Atlas* (Pincetl and LA Energy Atlas Development Team 2023). Temperature and precipitation data were derived from 17 weather stations near LA City.

4. *Engineering factors*: mean elevation, soil type, pipe material, and average pipe age. Data on mean elevation and soil type were also downloaded from the geospatial database of LA County (LA County 2023). The pipe material and average pipe age were calculated based on the leakage data downloaded from the *Los Angeles Times*.

2.3 Evaluating the Validity of Using Socioeconomic Factors for Leak Occurrence

We developed a model with MLR to evaluate the relative influence of potential explanatory factors. The objective is to determine whether some set of appropriate socioeconomic and urban planning characteristics influence WDN leak occurrence and should be included alongside engineering factors in predictive modeling. First, the number of observed leaks and the failure rate (leaks per 1000 connections) in a census tract were calculated. The failure rate calculation standardized the outcomes for population differences across census tracts. Outliers were removed over the 99th percentile (approximately 68 observed leaks per 1000 connections). Second, the number of failures and the failure rate were plotted against the composite indicator of socioeconomic vulnerability, the *CES* 2.0 score, to assess the overall trends. Third, potential predictive factors of leak events were estimated for census tracts to analyze trends in detail with predictive factors. The engineering operational parameters that could influence the rate of observed leaks included mean elevation and pipe age in a census tract. Other potentially important factors, such as system pressure or pipe diameter, were not available for large enough portions of pipes associated with leaks. The underlying socioeconomic and demographic parameters that make up the dataset were also extracted to be included in the database. Finally, a statistical model was developed to relate how the number of observed leaks (the number of failures and the failure rate) in a census tract varies based on explanatory variables of operational, geographic, and socioeconomic factors at the census tract level.

For the outcome variable of observed leaks per 1000 connections in a census tract, a linear model was constructed in *R*. The model included explanatory variables of poverty rate (those living below two times the federal poverty level), the percentage of low-income households, a ranking of drinking water quality based on 10 indicator contaminants contained within *CES*, total land area, median household income, mean elevation, and average pipe age within a census tract. Multicollinearity tests using a variable inflation factor (VIF) were performed to evaluate multicollinearity between the explanatory variables.

2.4 Identifying Clusters of Leaks from Existing data

We performed a cluster analysis to determine whether the observed leaks were randomly distributed in space and whether the occurrence of leaks exhibited non-random clustering, which may indicate influential factors to correlate with WDN or metropolitan characteristics. Spatial clusters are defined as areas with significantly higher events than the average or higher risks of consequences than the expected value (Roushangar et al. 2022).

Scan statistic is a widely used method for detecting spatial clusters. This method has been used in many disciplines, including disease, criminology, and disaster risk evaluation (Wadhwa and Thakur 2022; Mondal et al. 2022; Stimers et al. 2022). Robertson and Nelson (2010) compared the popular spatial cluster detection software programs and demonstrated that the scan statistic method had a more robust performance in automatically detecting clusters, among others.

The scan statistic software (*SaTScan*) calculated statistical significance based on data from each zone. A null hypothesis that assumed the risk of the population within a given area being affected by potential leaks is randomly distributed was adopted. A more detailed description is shown in *Supplemental Data*.

1. *Calculate the expected leak occurrence*: The distribution of leaks under the null hypothesis was estimated. The expected number of leaks was the fraction of leaks that fall within the scan window, assuming a uniform distribution of leaks. The population density ratio in the scan window was relative to the entire LA City as a proxy for associated infrastructure and, assuming that leaks were randomly distributed, the expected number of leaks.
2. *Identify the scan window size*: The spatial scan window in *SaTScan* was a circle, varying from 0 to a maximum size. *SaTScan* identified clustering in events by identifying the regions of the dataset with multiple events within a scan window across windows of many sizes. The scan window size in the spatial analysis was set as 50% (Kulldorff 2022). The centroids of the block groups, which were the middle point of the polygon, served as the center point of the scan window.
3. *Estimate the likelihood ratio*: For a given scan window size, a test statistic was used to compare the number of observed and expected leaks inside and outside the window. The discrete Poisson probability model was chosen (Kulldorff 2022).

   The likelihood ratio, which was based on the probability model, was calculated for each scan window and reflected the possibility that the window contained a cluster. High likelihood ratio values represented windows with a great likelihood of continuing leak clusters. *SaTScan* reported the logarithm of likelihood ratio (*LLR*).
4. *Test for statistical significance: SaTScan* used Monte Carlo analysis to test for the significance of identified clusters. *SaTScan* generated a series of randomly distributed leaks and then calculated the maximum *LLR* for each leak. If the *LLR* of the scan window set was ranked at R, then its significance $p = R / (N + 1)$, where N was the total number of Monte Carlo replications (Kulldorff 2022). Small *p-values* indicate that the probability of the clustering being random was small. In this study, the number of Monte Carlo replicates was set to 9999 to gain a steady result (Bailony et al. 2011). The significance level was $p \leq 0.05$.

2.5 Predictive Modeling to Evaluate Contributing Factors of Leak Risk

The ML technique is increasingly important in the analysis of modeling failure risk in WDNs (Li et al. 2022; Mazumder et al. 2021; Fan et al. 2022). Compared with single ML methods, such as support vector machines and decision trees, ensemble learning combines results from multiple weak learners to improve prediction. In addition, ensemble learning provides more explanation regarding the model structure, strategies, and mechanism explorations (Diamantopoulou 2023; Tripathi et al. 2023).

We addressed classification problems where the inputs were conducted by m-dimensional explanatory factors. The output, that is, the target variables, showed whether a census block group should be classified as a high-risk leak cluster (labeled with 1) or non-cluster (labeled with 0).

The predictive process was divided into four steps: scenario modeling, training and cross-validation (CV), model testing and evaluation, and key explanatory factor extraction.

2.5.1 Scenario Modeling

Two scenarios were conducted in this study: the interpolation prediction scenario (IPS) and the extrapolation prediction scenario (EPS). IPS randomly selected 80% of the samples for training and CV. The remaining 20% of the samples were used for testing. Then, EPS used 2010–2011 explanatory factors and 2011–2012 labels for training and CV and then tested on 2011–2012 explanatory factors and 2012–2013 labels, with a one-year interval.

2.5.2 Training and CV

This procedure identified which ML models were appropriate for leakage prediction based on three models, which were evaluated for effectiveness: logistic regression (LR), random forest (RF), and gradient boosting decision tree (GBDT). The mechanism of each model is briefly described in *Supplemental Data*.

The models' performance was judged by a 10-fold CV. The model to be adopted was determined by CV results because the validation set was an invisible set to the model.

2.5.3 Model Testing and Evaluation

The models were tested with the two prediction scenarios. In addition, six performance metrics, that is, accuracy, precision, recall, *F1* score, the area under the curve, and Brier, were calculated. A detailed description of metrics and corresponding equations are concluded in *Supplemental Data*.

2.5.4 Extraction Key Explanatory Factors

Recursive feature elimination (RFE) was proposed to find the key factors from the established ML prediction model (Guyon et al. 2002). The basic steps are described as follows: train the classifiers, calculate the ranking criterion for all features, and remove the features with the minimum ranking criteria. We used a 10-fold RFE CV to obtain the key feature combinations. *F1* score was chosen as the performance metric.

3 Results

The results are presented below for (1) trends in leak events, (2) regression models to evaluate the validity of using socioeconomic factors to understand leak occurrences, (3) analysis of spatial clustering of reported leak events, and (4) key explanatory factors of leak risks based on predictive modeling.

3.1 Frequency of Leak Occurrence

There were identifiable spatial and temporal trends among the 4,714 leakage incidents that occurred during the study period (Fig. S1 and Tables S1 and S2 in *Supplemental Data*). The number of leaks in a month fluctuated from 52 to 243, with an average of 98 leaks. Leaks occurred more often in December and January compared with other months. These totals were more than twice the values in March, April, May, and June. A small difference existed between summer and autumn occurrences, which are warm

seasons in Southern California's Coastal Mediterranean climate. However, minimal differences existed in the number of reported leaks across each year. The greatest number of leaks occurred in 2011 (1,317), followed by 2010 (1,159), whereas the number of leaks was equal in 2011 and 2013 (1,119).

3.2 MLR Models to Evaluate Sociodemographic Factors

With trends in clusters identified, MLR was used to explore the validity of constructing a predictive model for WDN leak risks that include sociodemographic factors. Plotting the number of failures and the failure rate (per 1,000 connections) against the overall index of socioeconomic vulnerability yielded opposite results. The number of failures and the *CES* score were positively correlated, whereas areas with high *CES* scores (low risk) had high failures. Alternatively, areas with high *CES* scores were inversely correlated with the rate of failures per 1000 connections (Fig. 3). Census tracts in *CES* with low scores (indicating less risk) tended to have a high failure rate. The main reason is that these areas have few WDN connections, which is a function of urban planning and density. The areas are less "dense," not having as many properties per unit area. The negative relationship between failure rate and density indicates that although a dense area with many connections and pipes should have high failures, the rate of increase in failures is less than expected.

The model with MLR ($r^2 = 0.17$) provided further insight into the observed trends in failure rates (Table 1). The typical operational indicators of elevation and pipe age were statistically significant, with elevation being negatively correlated and pipe age being positively correlated with the failure rate. Multiple socioeconomic indicators were also significantly correlated, including poverty rate (positive), indicators of drinking water quality (positive), area (positive), and population (negative). A VIF calculation for the model indicates low to moderate levels of multicollinearity, with values ranging from 1.06 to 2.95. The results provide evidence for the use of socioeconomic characteristics as potential predictive factors of WDN leak occurrences.

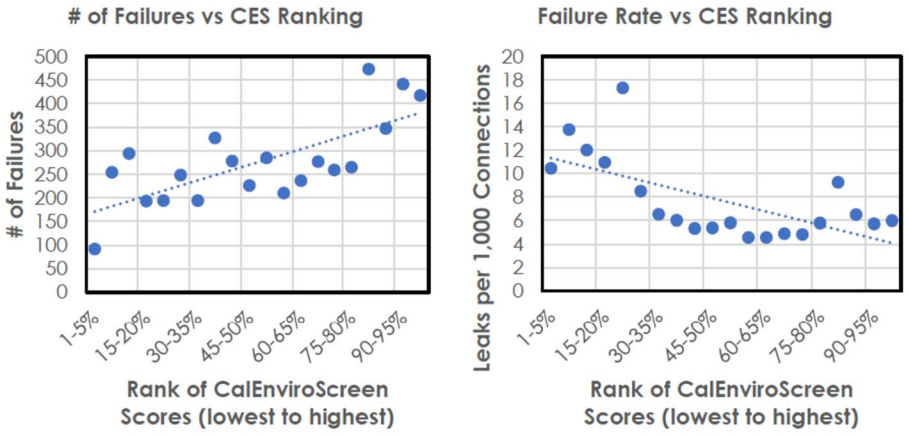

Fig. 3 Comparing the number of observed failures and the failure rate (per 1000 connections) in a census tract to its *CES* ranking. Lower rankings indicate the most vulnerable census tracts

Table 1 Statistical outputs from MLR, for the outcome variable of observed WDN leaks in a census tract per 1,000 connections ($r^2 = 0.17$, *F-statistic = 20.07 on 8 & 786 Degrees of Freedom)*

Explanatory Variable	Coefficient	Standard Error	t-value
Intercept	-1.01e + 00	2.59e + 00	-0.391
Poverty rate**	7.64e-02	2.50e-02	3.054
Low-income household burden	-4.51e-02	5.21e-02	-0.867
Indicators of adequate drinking water quality***	8.03e-02	1.87e-02	4.288
Area**	6.38e-07	2.00e-07	3.192
Population***	-1.64e-03	2.87e-04	-5.732
Median Household Income	2.51e-05	1.33e-05	1.89
Mean Elevation***	-3.45e-03	1.02e-03	-3.377
Average Pipe Age***	1.17e-01	1.72e-02	1.97e-11

** significant at 0.01 level; *** significant at < 0.001 level

3.3 Spatial Clustering

Statistically significant clusters of leaks ($p \leq 0.05$) were found throughout the region (Fig. 4). With census blocks as the smallest unit, the proportion of high-risk leakage blocks in LA City from 2010 to 2013 was 32.64%, 29.07%. 23.67%, and 26.19%, respectively. Downtown and southwest LA City were high-risk areas for WDN leaks.

3.4 Prediction Modeling

CV analysis revealed the algorithms that most accurately predicted leaks in the training set (Fig. S2 in *Supplemental Data*). The difference in CV scores between RF and GBDT was small; hence, both models were selected as the proposed models. In the test set, the RF model outperformed the GBDT model in all performance metrics in IPS and EPS (Table 2). The RF model was selected as the most suitable model and further calculated for feature selection.

Feature selection identified the explanatory factors of EPS leaks because it was greatly practical for predicting future leakage clusters. A total of 10 major factors were extracted: 10% related to geography (latitude and longitude), 30% to sociodemography (median household income, percentage of Hispanic or Latino population, and *CES* 2.0 Score), 40% to the environment (average maximum temperature, precipitation, percentage of ozone, and percentage of PM2.5), and 10% to engineering (mean elevation). The removal of multicollinearity among explanatory variables following RFE resulted in a small boost in the RF model's performance metrics on EPS. Its accuracy was 94.25% after three-quarters of the features were eliminated.

Figure 5 shows a comparative analysis of differences between the predicted results of considering and not considering socioeconomic factors. Evidently, the prediction performance decreased by approximately 10% on IPS and EPS. This result corresponded in part with the statistical analysis using MLR in that the high risk of WDN leaks can be explained by multiple factors that include engineering and socioeconomic indicators. In other words, the results from this study indicate that leak risk may not necessarily result from a single type of influence,

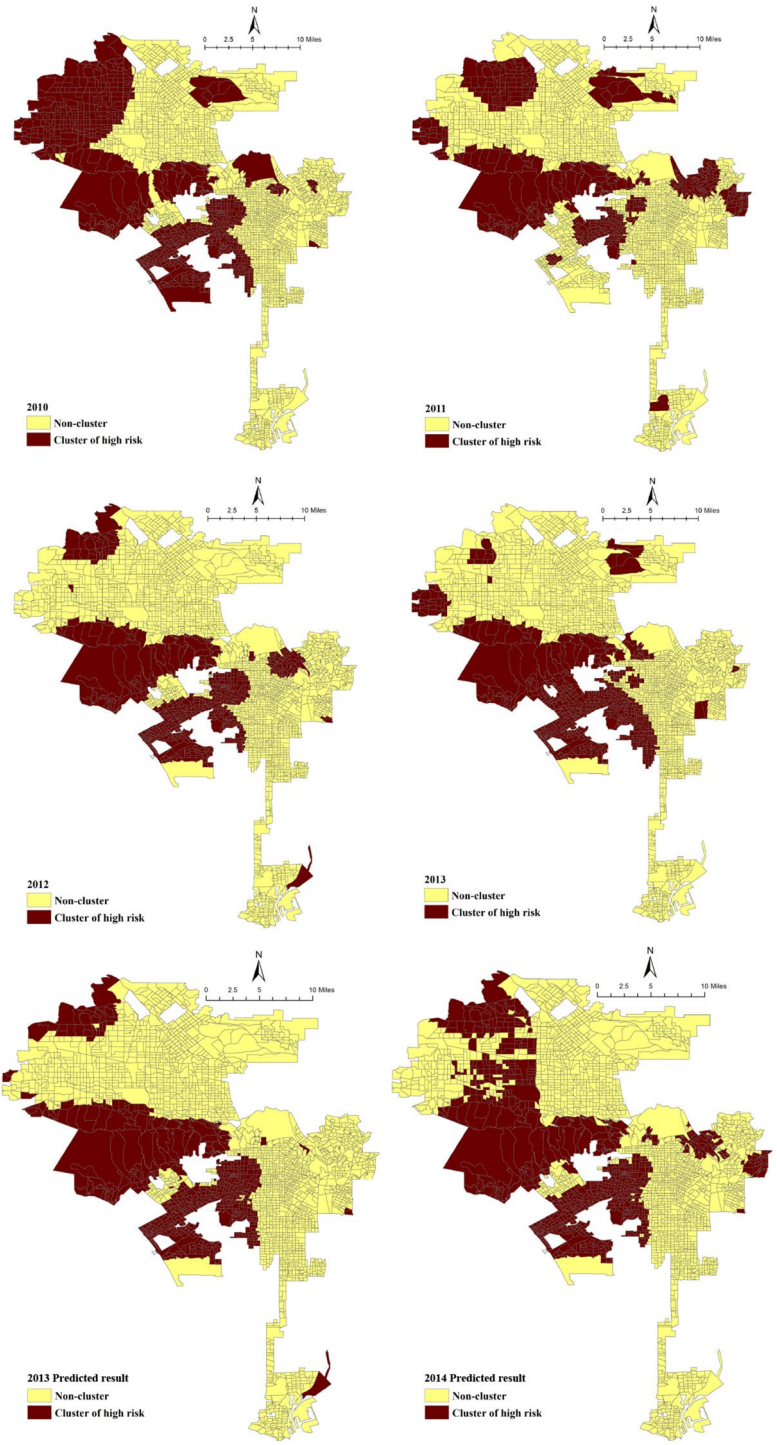

Fig. 4 Spatial location of Clusters of high leakage risk clusters (in brown) and non-clusters (in yellow) identified in LA City

Table 2 Prediction of leak events based on performance measures

IPS	RF	GBDT	**EPS**	RF	GBDT
Accuracy (%)	93.29%	92.29%	Accuracy (%)	92.45%	88.19%
Precision (%)	88.95%	86.70%	Precision (%)	89.96%	82.20%
Recall (%)	86.71%	85.46%	Recall (%)	78.47%	67.15%
F1 Score (%)	93.27%	92.28%	*F1* Score (%)	92.27%	87.76%
AUC	0.9128	0.9020	AUC	0.8778	0.8116
BS	0.0540	0.0562	BS	0.0568	0.0839

such as engineering operations. Rather, leak risk may relate to the many ways that infrastructure and operations are influenced by socioeconomic and climatic influences. Some additional parameters of interest, such as pipe pressure zones or pipe sizes, were not included owing to the lack of data but could be incorporated in future studies.

We further provided the leakage clustering prediction results for 2013 and 2014 and compared the 2013 prediction results with the actual clusters (Fig. 4). Owing to the limited amount of leakage data available, that is, only leakage data from 2010 to 2013 were disclosed, we could only provide short-term predictions with one-year intervals. The validation in 2013 showed that 88.51% of the census block group matched the actual leakage clusters, indicating the practical application of the predictive model. The prediction result in 2014 showed that the leakage risk was more severe than that in 2013, reaching 28.59%. If utilities could further open leakage point data over the years, then this model can be used for long-term leakage clustering prediction by adjusting the interval years.

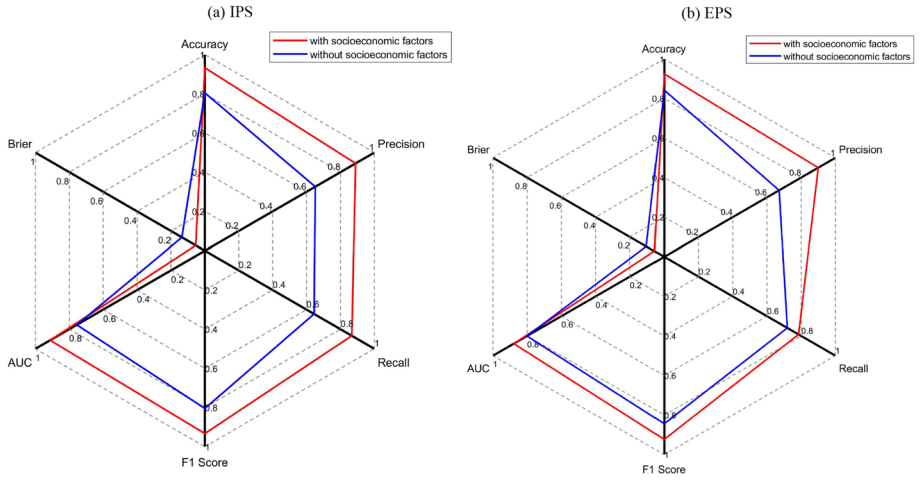

Fig. 5 Comparative analysis of differences between the predicted results of considering and not considering socioeconomic factors

4 Discussion

The results show that prediction modeling can help utilities develop asset management programs that proactively engage in operations and maintenance activities before damage occurs. The analysis illustrated how a collection of socioeconomic and environmental variables could be used to identify geographic clusters of leaks. Working in geographic areas can have tangible benefits, such as cost savings by minimizing mobilization costs and travel time between work sites for field crews.

The results are subject to several limitations. The publicly accessible data were restricted from 2010 to 2013, constraining future leakage predictions. The associated repair time for LADWP was included in the data set. However, entries for many records are missing (23.3%), which inhibits including repair time in the analysis. Additional granular data for socioeconomic and engineering operations could refine results. In addition, this study focuses on two classification problems. If continuous leak observation data can be obtained, then the regression model in ML technology can be performed.

Leaks in urban water systems cost utilities time and money. In California, statewide regulations have focused on improving efficiency and reducing leaks in cities. For instance, the California Senate Bill (SB) 1420 in 2014 required urban water supplies to begin submitting annual water loss audits with regular reporting on system-wide operations. Subsequently, in 2015, SB 555 required urban water supplies to submit validated, detailed audits of losses. As agencies expand efforts to improve water use efficiency beyond indoor and outdoor conservation, advanced methods for leak loss prediction based on openly available data can reduce the cost of implementing leak loss reduction programs.

5 Conclusions

This study analyzed spatially explicit explanatory and predictive factors in observed high-risk leakage clusters in LA City using multiple quantitative and modeling procedures. The primary contributions are as follows:

1. This study greatly broadens and deepens the understanding of how socioeconomic factors affect the spatial leakage clusters of WDN. Following Fan et al. (2022), this study proposes a multi-source dataset and further enriches the socioeconomic factors with significant effects on WDN leakage. Census tracts in the socioeconomic vulnerability index with less risk tended to have a high failure rate.
2. The best-performed RF model successfully predicted high-risk leakage clusters in 2013 and could be employed to predict for the upcoming year. This model was thoroughly validated and compared on two scenarios with six performance metrics. The prediction period can be longer by adjusting EPS interval years if the utilities publish more open leakage data.
3. This study identified significant socioeconomic factors. The outcomes of the comparison analyses also demonstrated that considering socioeconomic factors had a positive impact on prediction accuracy. The contributions of the engineering and environmental factors were in line with the findings of previous studies. This finding implies that non-traditional factors—that is, socioeconomic ones—have a significant impact on WDN leakage, which can be useful in explaining observed leak loss occurrence in urban areas.

Complicated non-linear interactions among multiple components lead to WDN leakage. Future research should consider advanced ML methods and further information collection to increase the accuracy of WDN leakage predictions. As cities target programs to reduce water losses in distribution systems, data-driven studies combining engineering, environmental, and socioeconomic data can help inform utility management. Such studies may also potentially facilitate a move toward decision support for proactive management with sustainability goals.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11269-023-03676-w.

Acknowledgements Many thanks to Eric Fournier, Dan Cheng, Claire Hirashiki, and Stephanie Pincetl at the California Center for Sustainable Communities within UCLA's Institute of the Environment and Sustainability for guidance and technical assistance.

Authors' Contributions Shuang Q. and Porse E. designed the research, collected the data, and programed the code; Zhao R.T. tested the model; Shuang Q. and Zhao R.T. wrote the original draft preparation; Porse E. reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding Support for this research was provided by the Beijing Social Sciences Foundation (grant number 18GLC070), the Fundamental Research Funds for the Central Universities (grant number 2019JBW007), and China Scholarship Council (grant number 201707095080). The Office of Water Programs at California State University, Sacramento also supported the research.Beijing Social Sciences Foundation,18GLC070,Qing Shuang,Fundamental Research Funds for the Central Universities,2019JBW007,Qing Shuang,China Scholarship Council,201707095080,Qing Shuang.

Data Availability The data is available in an online repository: https://github.com/erikporse/artes. Data used in the study was compiled from multiple sources, including: (CEC 2021); (County's Enterprise GIS (eGIS) Steering Committee 2018); (Krishnakumar and Poston 2016); (OEHHA 2022); (Pincetl and LA Energy Atlas Development Team 2023); (Poston and De Groot 2014); (Poston and Stevens 2015).

Code Availability Code generated or used during the study are available in an online repository: https://github.com/erikporse/artes.

Declarations

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent to Publish Not applicable.

Conflicts of Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abokifa AA, Sela L (2023) Integrating spatial clustering with predictive modeling of pipe failures in water distribution systems. URBAN WATER J 20:465–476. https://doi.org/10.1080/1573062X.2023.2180393

American Society of Civil Engineers (2017) Infrastructure report card: drinking Water. American Society of Civil Engineers (ASCE), Reston, VA (Reprinted)

Bailony MR, Hararah MK, Salhab AR, Ghannam I, Abdeen Z, Ghannam J (2011) Cancer registration and healthcare access in West Bank, Palestine: a GIS analysis of childhood cancer, 1998–2007. INT J CANCER 129:1180–1189. https://doi.org/10.1002/ijc.25732

Barton NA, Farewell TS, Hallett SH, Acland TF (2019) Improving pipe failure predictions: factors affecting pipe failure in drinking water networks. WATER RES 164:114926. https://doi.org/10.1016/j.watres.2019.114926

CEC (2021) California Building Climate Zones. https://cecgis-caenergy.opendata.arcgis.com/datasets/CAEnergy::california-building-climate-zones/explore

County's Enterprise GIS (eGIS) Steering Committee (2018) Los Angeles County GIS Data Portal. https://egis3.lacounty.gov/dataportal/. Accessed 12 Jan 2020

Da Silveira APP, Mata-Lima H (2021) Assessing energy efficiency in water utilities using long-term data analysis. Water Resour Manag 35:2763–2779. https://doi.org/10.1007/s11269-021-02866-8

Diamantopoulou MJ (2023) Machine learning in environmental modeling: A case study with groundtruth data from Seich–Sou suburban forest, Greece. Paper presented at the 12th World Congress on Water Resources and Environment (EWRA 2023), Thessaloniki, Greece

Fan X, Wang X, Zhang X, Yu XB (2022) Machine learning based water pipe failure prediction: the effects of engineering, geology, climate and socio-economic factors. Reliab Eng Syst Safe 219:108185. https://doi.org/10.1016/j.ress.2021.108185

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer classification using support Vector machines. MACH LEARN 46:389–422. https://doi.org/10.1023/A:1012487302797

Hani A, Nechem D, Hani S, Bougherira N, Toumi F, Djabri L, Chaffai H (2023) Multi-criteria analysis and characterization of the integrated water resources management model in the Annaba region. Paper presented at the 12th World Congress on Water Resources and Environment (EWRA 2023), Thessaloniki, Greece

Hunaidi O (2006) (Condition assessment of water pipes) Proceedings of the EPA Workshop on Innovation and Research for Water Infrastructure in the 21st Century, EPA Workshop, Arlington, VA, USA

Krishnakumar P, Poston B (2016) Los Angeles water main leaks since 2010, Los Angeles Times, http://graphics.latimes.com/los-angeles-pipe-leaks/

Kulldorff M (2022) SaTScanTM User Guide. http://www.satscan.org/techdoc.html. Accessed: 14 Nov 2023

LADWP (2023) 2022-23 Water infrastructure plan. Los Angeles. Reprinted

LA County (2023) Los Angeles County GIS Data Portal. https://egis3.lacounty.gov/dataportal/

Li Z, Wang J, Yan H, Li S, Tao T, Xin K (2022) Fast Detection and Localization of Multiple Leaks in Water Distribution Network jointly driven by Simulation and Machine Learning. J Water Res Plan Man 148:5022005. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001574

Mazumder RK, Salman AM, Li Y (2021) Failure risk analysis of pipelines using data-driven machine learning algorithms. Struct Saf 89:102047. https://doi.org/10.1016/j.strusafe.2020.102047

Misiūnas D (2008) Failure monitoring and asset condition asssessment in water supply systems. Vilniaus Gedimino technikos universitetas

Mondal S, Singh D, Kumar R (2022) Crime hotspot detection using statistical and geospatial methods: a case study of Pune City, Maharashtra, India. GeoJournal 87:5287–5303. https://doi.org/10.1007/s10708-022-10573-z

Morrison R (2021) The Aging Pipes Dilemma. https://www.contractormag.com/piping/article/21153216/the-aging-pipes-dilemma. Accessed: 2 Nov 2023

OEHHA (2022) CalEnviroScreen v 2.0. Sacramento, CA: California Office of Environmental Health Hazard Assessment. https://oehha.ca.gov/calenviroscreen. Accessed: 1 Nov 2023

Pincetl S, LA Energy Atlas Development Team (2023) LA Energy Atlas. California Center for Sustainable Communities. UCLA, Los Angeles, CA. (Reprinted)

Poston B, De Groot L (2014) Water pipe leaks in Los Angeles, Los Angeles Times, http://www.latimes.com/visuals/graphics/la-me-g-water-leaks-20141107-htmlstory.html

Poston B, Stevens M (2015) LA's Aging Water Pipes; a $1-Billion Dilemma, Los Angeles Times, Accessed 13 Jan 2020. http://graphics.latimes.com/la-aging-water-infrastructure/

Rahimi-Golkhandan A, Aslani B, Mohebbi S (2022) Predictive resilience of interdependent water and transportation infrastructures: a sociotechnical approach. Socio-Econ Plan Sci 80:101166. https://doi.org/10.1016/j.seps.2021.101166

Reyes-Velarde A (2018) Century-old water main breaks in South Los Angeles, submerging streets and cars and spurring evacuations. https://www.latimes.com/local/lanow/la-me-ln-south-la-water-main-break-20181221-story.html. Accessed 2 Nov 2023

Robertson C, Nelson TA (2010) Review of software for space-time Disease surveillance. Int J health Geogr 9:16. https://doi.org/10.1186/1476-072X-9-16

Robles-Velasco A, Cortés P, Muñuzuri J, De Baets B (2023) Prediction of pipe failures in water supply networks for longer time periods through multi-label classification. Expert Syst Appl 213:119050. https://doi.org/10.1016/j.eswa.2022.119050

Roushangar K, Ghasempour R, Nourani V (2022) Spatiotemporal analysis of droughts over different climate regions using hybrid clustering method. Water Resour Manag 36:473–488. https://doi.org/10.1007/s11269-021-02974-5

Stimers M, Lenagala S, Haddock B, Paul BK, Mohler R (2022) Space-time clustering with the space-time permutation model in SaTScan™ Applied to Building Permit Data following the 2011 Joplin, Missouri Tornado. Int J Disast Risk Sc 13:962–973. https://doi.org/10.1007/s13753-022-00456-9

Swilling M (2011) Reconceptualising urbanism, ecology and networked infrastructures. Soc Dyn 37:78–95. https://doi.org/10.1080/02533952.2011.569997

Tripathi V, Mohanty MP, Singh H (2023) Fidelity of machine learning models in capturing flood inundation through geomorphic descriptors over Ganga sub-basin, India. Paper presented at the 12th World Congress on Water Resources and Environment (EWRA 2023), Thessaloniki, Greece

U.S. Census Bureau (2023) American Community Survey. https://www.census.gov/data.html. Accessed: 1 Nov 2023

Wadhwa A, Thakur MK (2022) Rapid surveillance of COVID-19 by timely detection of geographically robust, alive and emerging hotspots using particle Swarm Optimizer. Appl Geogr 144:102719