

UC Berkeley

UC Berkeley Previously Published Works

Title

Identifying schools at high-risk for elevated lead in drinking water using only publicly available data

Permalink

<https://escholarship.org/uc/item/8tq9z3rg>

Authors

Lobo, GP
Laraway, J
Gadgil, AJ

Publication Date

2022

DOI

10.1016/j.scitotenv.2021.150046

Peer reviewed



Identifying schools at high-risk for elevated lead in drinking water using only publicly available data

G.P. Lobo^a, J. Laraway^b, A.J. Gadgil^{a,*}

^a Department of Civil and Environmental Engineering, University of California, Berkeley 94720, United States

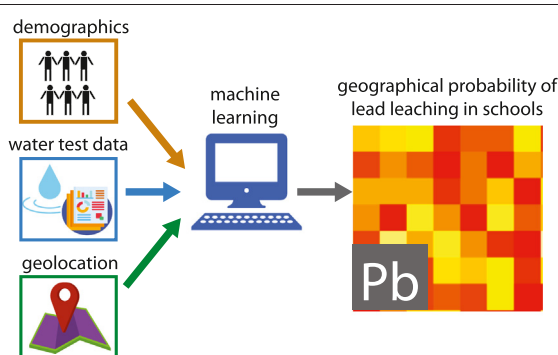
^b Department of Environmental Science, Policy and Management, University of California, Berkeley 94720, United States



HIGHLIGHTS

- A highly accurate machine learning model was developed to predict lead levels in schools.
- The model was implemented using only publicly available data for over 8000 schools.
- The model was used to predict lead exposure from water in schools in California and Massachusetts.
- We estimate that over 16,000 5-year-old children may be exposed to high lead levels in CA and MA.
- The model could help identify hotspots at a state level where lead testing should be a priority.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 23 June 2021

Received in revised form 26 August 2021

Accepted 26 August 2021

Available online xxxx

Editor: Jay Gan

Keywords:

Lead in school drinking water

Lead leaching

Machine learning

Environmental justice

Public data mining

ABSTRACT

Estimating the risk of lead contamination of schools' drinking water at the State level is a complex, important, and unexplored challenge. Variable water quality among water systems and changes in water chemistry during distribution affect lead dissolution rates from pipes and fittings. In addition, the locations of lead-bearing plumbing materials are uncertain. We tested the capability of six machine learning models to predict the likelihood of lead contamination of drinking water at the schools' taps using only publicly available datasets. The predictive features used in the models correspond to those with a proven correlation to the dominant, but commonly unavailable, factors that govern lead leaching: the presence of lead-bearing plumbing materials and water quality conducive to lead corrosion. By combining water chemistry data from public reports, socioeconomic information from the US census, and spatial features using Geographic Information Systems, we trained and tested models to estimate the likelihood of lead contaminated tap water in over 8,000 schools across California and Massachusetts. Our best-performing model was a Random Forest, with a 10-fold cross validation score of 0.88 for Massachusetts and 0.78 for California using the average Area Under the Receiver Operating Characteristic Curve (ROC AUC) metric. The model was then used to assign a lead leaching risk category to half of the schools across California (the other half was used for training). There was good agreement between the modeled risk categories and the actual lead leaching outcomes for every school; however, the model overestimated the lead leaching risk in up to 17% of the schools. This model is the first of its kind to offer a tool to predict the risk of lead leaching in schools at the State level. Further use of this model can help deploy limited resources more effectively to prevent childhood lead exposure from school drinking water.

© 2021 Published by Elsevier B.V.

* Corresponding author.

E-mail address: ajgadgil@berkeley.edu (A.J. Gadgil).

1. Introduction

Childhood lead poisoning is a multifactorial problem that affects as many as 500,000 US children younger than 6 years of age (Hauptman et al., 2017). Drinking water is one of several sources of lead exposure (Triantafyllidou and Edwards, 2012): lead leaching from lead-based water infrastructure affects over 5000 public drinking water utilities across the US, potentially putting over 18 million people at risk (Olson and Fedinick, 2016). While there is good understanding of how different water quality conditions affect lead solubility (Schock, 1990), water quality conditions at the tap are not typically reported. Absence of this knowledge makes it difficult to predict lead levels at the tap and to estimate the number of children exposed to high lead levels in their drinking water. Thus, lead monitoring programs in schools and cities, such as the 3 Ts (Training, Testing and Taking action) and citizen science-based strategies (Redmon et al., 2020) are the only tools currently available to identify the schools with elevated lead tap water levels (Dignam et al., 2019). Widely implemented lead-monitoring programs have been the cornerstone of remediation programs implemented in cities like Flint, MI and Newark, NJ, where thousands of people have been exposed to lead levels well above the 10 ppb guideline value established by the World Health Organization (WHO, 2017).

After the dramatic media exposure received by the 2016 lead water crisis of Flint, MI, several states, including California, implemented mandatory lead testing in all K-12 schools (Kunapuli et al., 2018). Other states, including Massachusetts, implemented voluntary lead testing programs to help schools test for lead and copper in their drinking water (Burlingame et al., 2018). The lead results for some states are publicly available and constitute, to the best of our knowledge, one of the largest publicly available drinking water lead databases spanning entire states. Lead leaching levels are also reported by water utilities in yearly Consumer Confidence Reports (CCRs), as required by the Lead and Copper Rule (LCR) (Ramaley, 1993). However, each utility is required to report only a single number, the 90th percentile lead level, which does not provide information on the actual distribution of lead concentrations, nor on the spatial distribution of lead leaching, unlike the school lead data.

The stark differences between the school lead data and the 90th percentile lead levels reported by the water utilities becomes apparent when they are compared. Even though hundreds of schools in California and Massachusetts reported lead levels above 15 ppb in 2018, none of the utilities in California and only 3 in Massachusetts reported 90th percentile levels above 15 ppb (see Fig. 1). This difference shows that

compliance with the LCR does not necessarily correlate to childhood lead exposure from school drinking water. However, we note that this difference may be a result of differences in sampling protocols between homes and schools (Triantafyllidou et al., 2021) (see Section 2.1 for school drinking water sampling protocols in California and Massachusetts).

In contrast, exhaustive large-scale lead testing programs are expensive and hard to implement (Katner et al., 2016), thus, they have been mostly implemented in cities where lead water crises have already occurred and received wide adverse publicity.

Machine learning is a promising technique that may help identify the locally differentiated risk of lead leaching into the drinking water in schools and elsewhere without the need of exhaustive sampling of entire cities. This technique has been used to identify the presence of lead service lines (Abernethy et al., 2018) and the risk of lead leaching in Flint, MI (Chojnacki et al., 2017). The predictive features used as input data in these studies include demographic and socioeconomic factors, including poverty rates, race, property values, and dates in which properties were built, among others (Switzer and Teodoro, 2017). However, to our knowledge, all the published literature regarding these machine-learning based approaches is limited to only those cities where massive lead testing has been already conducted following a major water crisis. Moreover, extrapolating a model trained on data from one city might not provide accurate results elsewhere because the water chemistry and socioeconomic characteristics might be different across different cities. Another group of researchers developed a model to predict the risk of lead leaching in private drinking water systems in Virginia by using household, geographical and chemical characteristics (Fasaei et al., 2021). However, its applicability is limited to private systems and does not apply to the over 148,000 public drinking water systems in the US that provide water to 90% of Americans.

We hypothesize that a machine learning model meant to predict the risk of lead contamination with the potential of being extrapolated to multiple locations must consider at least three kinds of data (in data science literature these data are called “features” and we will use that term from here onward). These features are: (1) water quality (i.e., water chemistry) information, which includes source and treated water quality, as well as water temperature, (2) demographics, and (3) spatial data. Water quality largely controls lead solubility (Noel et al., 2014; Schock et al., 2014; Masters et al., 2016). Some components of demographic information, such as poverty rates and race, are good predictors of the presence or absence of lead-bearing plumbing materials (Sampson and Winter, 2016). These materials include lead pipes, fixtures, solder and leaded brass plumbing, most of which are found in

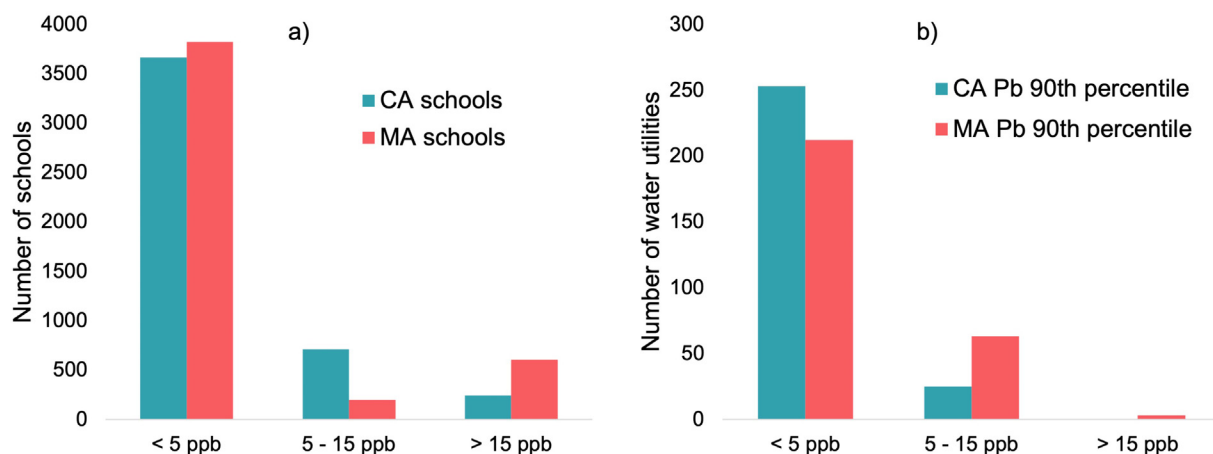


Fig. 1. (a) Number of schools, and (b) number of water utilities, that reported lead levels (at the tap) below 5 ppb, between 5 and 15 ppb and above 15 ppb (the action level of the MCL) for California and Massachusetts in 2018. No water utilities in California and only three in Massachusetts reported 90th percentile lead levels above 15 ppb. In contrast, over 240 and 600 schools exceeded the action level of 15 ppb in California and Massachusetts, respectively. Relying only on the 90th percentile numbers might give a false sense of security for lead in tap water at schools, as demonstrated in California data.

pre-1986 buildings (Sampson and Winter, 2016). Finally lead leaching rates (and therefore lead concentrations in water supply) can also depend on local water quality conditions that develop in the water distribution system in a spatially correlated manner (Aisopou et al., 2012). Moreover, nearby schools may belong to the same administration and be subject to similar policies and contracts for purchasing water fountains or may have been built on similar dates.

To our knowledge, a model based on these key features (that arise from insights in environmental chemistry, history of urban demographics, and history of water infrastructure) has not been reported in the literature. Thus, predicting lead leaching at scales larger than specific, well-sampled cities has remained an unsolved problem. Furthermore, we are not aware of any predictive model of lead leaching in schools in the published literature. Lead leaching in schools is typically associated to brass or bronze plumbing and the use of lead solder (Triantafyllidou and Edwards, 2012). This differs sharply to household lead leaching, which is typically associated to the presence of lead service lines (Cornwell et al., 2016).

We report on the use of machine learning to predict the risk of lead contamination of tap water in schools. Our aim was to implement a robust model that can predict where lead leaching is likely to take place, regardless of the source (lead from pipes, fixtures, or solder etc.), or the form of lead contamination (dissolved or particulate).

Our approach is statistical. The model is not meant to provide a mechanistic explanation of how lead enters the schools' water supply. It only aims to help identify schools with a high risk of lead leaching by relying only on publicly available data.

2. Materials and methods

2.1. Study sites

California and Massachusetts were chosen as study sites because they have two of the largest publicly available and easily accessible datasets of lead levels in schools (Agency, 2021; Executive Office of Energy and Environmental Affairs, 2021). Most states do not have these kinds of data aggregated, organized, and made publicly available. Many others that do, either have incomplete data or present them in ways that are hard to obtain and process (e.g., only paper copies located at various local offices of different agencies). The spatial distributions of all the lead water data used in this study are shown in Fig. 2a and b.

The California dataset contains up to five samples per school, all of which were obtained from regularly used water delivery points, including drinking fountains, cafeteria and food preparation areas, and reusable water bottle filling stations (California Water Boards, 2017). All samples were "first draw" samples: 1 L of water was obtained after a 6 h or longer stagnation period. Only the data corresponding to the year 2018 for 6954 schools were analyzed in this study. We note that private and charter schools are not required to test for lead in California, thus, the dataset contains mostly data from public schools.

The Massachusetts dataset contains a variable number of samples per school (the MassDEP Drinking Water Program recommends sampling all fixtures at least once (MassDEP, 2016)). Two sampling methods were reported: (1) "first draw" samples, which were obtained from drinking water fixtures after an 8 h or longer stagnation time using 250 mL bottles (this method is different to the "first draw" samples described for California), and (2) "flush samples", which were collected after 30 s of flushing using 250 mL bottles. Only the "first draw" samples were analyzed to decrease any bias that may result from including both sampling protocols. Only the data corresponding to the year 2016 for 1151 schools were analyzed.

2.2. Features used

Two factors predominantly control lead leaching into the drinking water: (1) the presence or absence of lead-bearing plumbing materials, and (2) water quality that promotes the formation of soluble or insoluble lead corrosion products. However, there are no studies that we are aware of that report either of those two factors for schools in California or Massachusetts. Thus, in this study we used 94 different, publicly available, features that, to some degree, are proxies for the two factors outlined above. These features fall in one of the three below categories:

2.2.1. Demographic data

Low-income minority groups have been reported to be at a disproportionately larger risk of lead exposure from old and poorly maintained water infrastructure in Flint, MI and the state of New Jersey (Hanna-Attisha et al., 2016; Gleason et al., 2019). Thus, we included measures of race and poverty into the machine learning model to account for the fact that race and poverty may be good proxies for the absence or presence of lead-bearing plumbing materials.

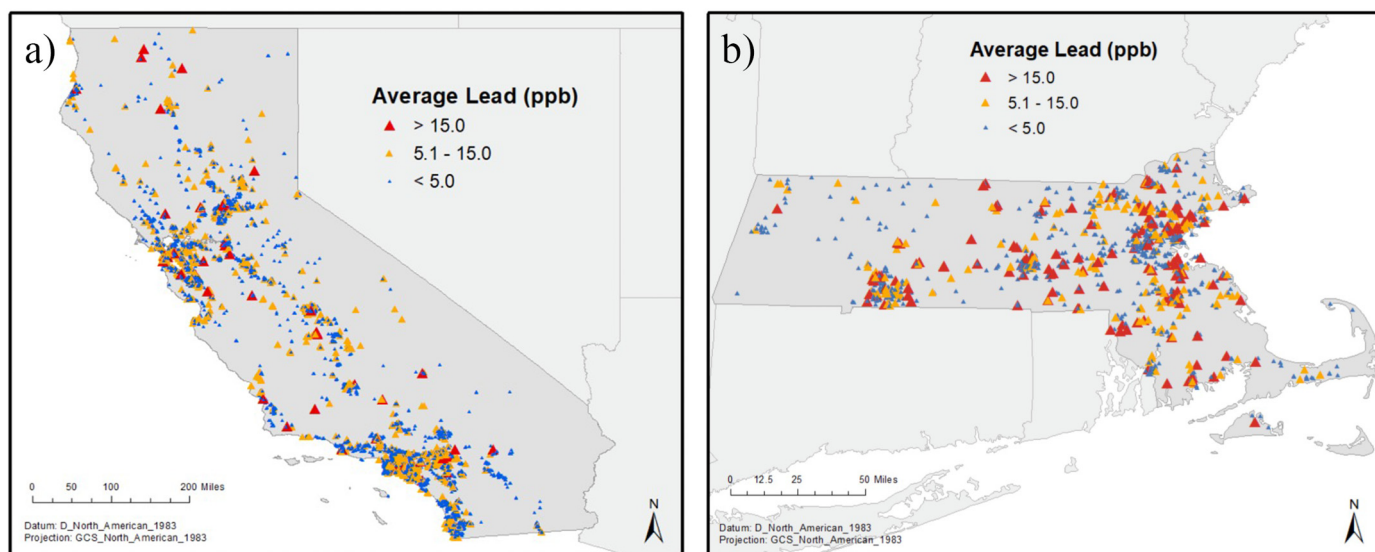


Fig. 2. Maps of schools in (a) California and (b) Massachusetts where lead levels were measured in 2018 and 2016, respectively. These databases were used in this study to develop a predictive machine learning model.

The US Census Bureau's data portal was used to obtain the socioeconomic information for the students attending each school by assigning data from individual census tracts to schools drawing students from those census tracts. All social data are estimates regarding social conditions within California and Massachusetts census tracts and are provided by the American Community Survey for the years 2016 and 2018 (United States Census Bureau, 2021). The selected socioeconomic features correspond to race, poverty levels and unemployment. We note that previous studies have linked demographic data to the risk of lead exposure at the community level, and not to the risk of exposure within schools (Chojnacki et al., 2017; Fasaee et al., 2021). Thus, in this study we assumed that the correlations found between race, poverty and lead exposure hold true for schools. This is a reasonable assumption given that schools with high populations of students from low-income minority families are more likely to have old infrastructure, including lead-bearing plumbing (Jackson and Johnson, 2021).

2.2.2. Water quality data

The main water quality parameters that control lead solubility in drinking water are outlined below:

- pH: elevated pH levels tend to decrease lead concentrations due to both the relatively low solubility of lead corrosion products and slower dissolution rates at high pH values in drinking water (Kim et al., 2011).
- Disinfection residual: the use of strong oxidants as disinfection residuals, such as free chlorine, promote the formation of Lead(IV), which is insoluble in drinking water (Tam and Elefsiniotis, 2009). In contrast, moderate disinfectants, such as monochloramines, promote the formation of more soluble Lead(II) minerals (Switzer et al., 2006). Lead concentrations in drinking water depend on both the type of lead mineral in contact with the water, as well as the surrounding water chemistry (Masters et al., 2016; Rajasekharan et al., 2007; Wang et al., 2013).
- Dissolved solids: the presence of several ions in drinking water have a large effect on lead solubility. For instance, chloride ions promote lead corrosion, while sulfates limit lead solubility (Edwards and Triantafyllidou, 2007). Other ions, including calcium and carbonates also have an effect on dissolved and particulate lead levels (Tam and Elefsiniotis, 2009; Desantis et al., 2020). Moreover, numerous dissolved ions make up a large proportion of water alkalinity, which is critically important to aqueous solubility.
- Organic matter: elevated organic matter levels have been associated to the reductive dissolution of Lead(IV) present on the inner surfaces of lead pipes (Masters et al., 2016). The presence of natural organic matter has also been associated to lead release from lead-bearing brass plumbing (Korshin et al., 2000).
- Water temperature: high water temperatures enhance dissolution kinetics and thus, may impact the rate at which lead leaches into the drinking water (Trueman et al., 2016). Moreover, seasonal variations in surface water temperature may be associated to fluctuations of dissolved organic matter concentrations, affecting lead leaching rates (Masters et al., 2016).
- Corrosion inhibitors: many water utilities across the US add corrosion inhibitors, such as orthophosphates, to the drinking water to decrease lead solubility (Stone et al., 2010). Orthophosphates react with dissolved Lead(II) ions and form insoluble Lead(II) phosphate minerals, which then precipitate onto the pipe surfaces. However, under certain conditions the use of orthophosphates may promote the release of particulate lead into the drinking water (Zhao et al., 2018).

Even though these water quality parameters largely determine lead solubility, they are usually not measured at the tap in schools. However, every community water utility in the US is required to develop yearly Consumer Confidence Reports (CCRs) with details on average water quality parameters. Thus, we manually extracted relevant water quality

features from the 2016 and 2018 CCRs of each of the over 600 water utilities serving each school shown in Fig. 2a and b. Each school was assigned to a specific water utility by using service area boundary maps. The extracted features include pH, disinfection residual type and amount, dissolved ions, including sulfate, chloride, and sodium, alkalinity, and water hardness. Trihalomethanes and halo-acetic acids were also included because they are disinfection byproducts from reactions between the disinfectant and organic matter, and thus they may be considered proxies for organic matter in utilities that use a free chlorine as a disinfectant (most water utilities do not report organic matter levels). The use and dose of orthophosphates or other corrosion control strategies were not included because, even though they are key for controlling lead solubility, most water utilities, particularly the small ones, do not report whether they use them or not.

We also extracted data from the Safe Drinking Water Information System (SDWIS) database (EPA, 2021), which contains, for each water utility, information on the water source (ground or surface water), the number of people served, the number of service connections, and the number of violations, among others. Moreover, we obtained average climate information for each city in California and Massachusetts from the National Centers for Environmental Information (NOAA) (NOAA, 2021), as temperature and precipitations may impact the water quality (Masters et al., 2016). These data include minimum, maximum, and average temperatures, as well as precipitation.

We note that water quality may vary significantly over the course of a year, making the CCRs poor indicators of water chemistry. Water quality may also change significantly during distribution and hence, chemical properties measured at the treatment facility may be very different to the ones that reach a particular school. Furthermore, the aforementioned water quality parameters do not consider other mechanisms of corrosion, including galvanic coupling between brass or copper pipes with lead-based fixtures, or lead release caused by physical disturbances (Abokifa and Biswas, 2017). However, in the absence of publicly available water quality data measured at each school, the use of chemical data from CCRs is the only feasible approach to include publicly available chemical information into the model.

2.2.3. Spatial data

Spatial features were included in the model to account for the fact that water quality may change as a function of space (a concept termed "water age" by the EPA (2002)). This may happen as a result of the consumption of the disinfection residuals or leaks, among others (Charisiadis et al., 2015). Thus, we used ArcMap 10.7, a Geographical Information System (GIS), to compute the distance of each school to the nearest school where the reported average lead level in tap water (calculated as the average of lead levels measured at all tested fixtures in the school) exceeds 10 ppb. These distances were calculated as straight lines between schools and were not based on pipe networks or restricted to schools within the same water systems. Preliminary data analysis suggested that lead leaching in schools tends to exist in clusters, regardless of water system boundaries. These clusters may exist because of similar demographics in specific areas, because nearby schools may have been constructed in the same year, or because of similarities in local water quality conditions that are conducive to lead corrosion. We note that, if a school exceeds this 10-ppb threshold, its distance to the nearest school where the threshold is exceeded is not 0; it is the distance to the nearest different school where this level is exceeded. Moreover, these distances were calculated with respect to the schools in the training sets during model implementation. Thus, the distances to the nearest school where lead levels exceed 10 ppb were recalculated every time the data were separated into training and testing sets.

2.2.4. Data acquisition challenges and assumptions

Gathering the data mentioned in Sections 2.2.1–2.2.3 was a challenging process. In the case of the chemical data, there were several

inconsistencies in reporting among different water utilities, and significant data were missing. We only included water quality parameters that were present in over 70% of the water utilities of each state. For instance, water pH was reported by 80% of the water utilities in California. In contrast, pH was reported by only 15% of the utilities in Massachusetts. Thus, we included pH in the water quality dataset for California, but not for Massachusetts. Missing data was not a major problem in California because most of the water utilities consistently reported the same parameters throughout the state (about 10% of the chemical data was missing). However, in Massachusetts only sodium, disinfection byproducts, and the 90th percentile lead and copper levels were consistently reported among water utilities. Thus, in Massachusetts only the aforementioned four water quality parameters were used in the model (the demographic and spatial variables used were the same for both states).

Several water utilities also report water quality parameters of different water sources, which may be blended before delivery to the customer or used seasonally (e.g. groundwater in one season and surface waters in another season). In those cases, we computed the value of each parameter as a weighted average by considering the percentage of each water source. This approach has limitations because the average water quality may not be representative of the water used for distribution. When the water source percentage was missing (15% of the CCRs), we assumed that each water supply contributes equally (i.e., the same volume of water) to the water system. This assumption may not be reasonable for some of the water utilities; however, one of the objectives of this study was to determine whether machine learning may be used to determine the likelihood of lead leaching under uncertainty using publicly available data. Most of the uncertainty in this study come from assumptions and inconsistencies among CCRs, which are the only publicly available data sources on drinking water quality.

A total of 94 and 88 features within the categories described in Sections 2.2.1–2.2.3 were obtained for the CA and MA models, respectively (see full datasets in our Github repository (Lobo et al., 2021)). Given that a different number of features were used in both models and that there are differences in lead sampling protocols in CA and MA, the models are not meant to be compared to each other.

2.3. Model implementation

A schematic overview of the steps taken to process the data and implement the machine learning model, described herein, is shown in Fig. 3.

2.3.1. Data preprocessing

The features described in Section 2.2 and the measured lead levels in each school were combined into two datasets, one for California and one for Massachusetts. The datasets were then preprocessed by one-hot encoding all categorical features, such as water utility name and disinfection type (one-hot encoding is a method used to transform categorical features into binary features (Brownlee, 2017)). The missing data, all of which corresponded to water quality parameters, were filled in by using the mean value of each feature (e.g., the missing pH values were filled in using the average pH value of all water utilities).

The lead datasets described in Section 2.2.1 were processed by first calculating the average lead levels per school measured among all fixtures (several schools had more than one lead measurement). We used the average lead levels because we assumed that each student drinks an equal amount from each of the fixtures tested for lead. These values were then transformed into binary variables by establishing a variable threshold, transforming each average lead measurement to either 0 (below the threshold) or 1 (above the threshold). The values of 10 and 5 ppb were chosen as thresholds because 10 ppb is the current guideline value established by the World Health Organization (WHO, 2017), while 5 ppb is currently considered unsafe by Health Canada (Canada, 2019) and is being considered as a new limit also in the European Union (European Commission, 2018). Several other thresholds may be used, including the Environmental Protection Agency standard for school samples of 20 ppb, or the action limit of the LCR of 15 ppb (the latter applies to the 90th percentile lead levels exclusively).

2.3.2. Evaluating different machine learning models

Six machine learning models (Random Forest, k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Logistic regression, Naïve

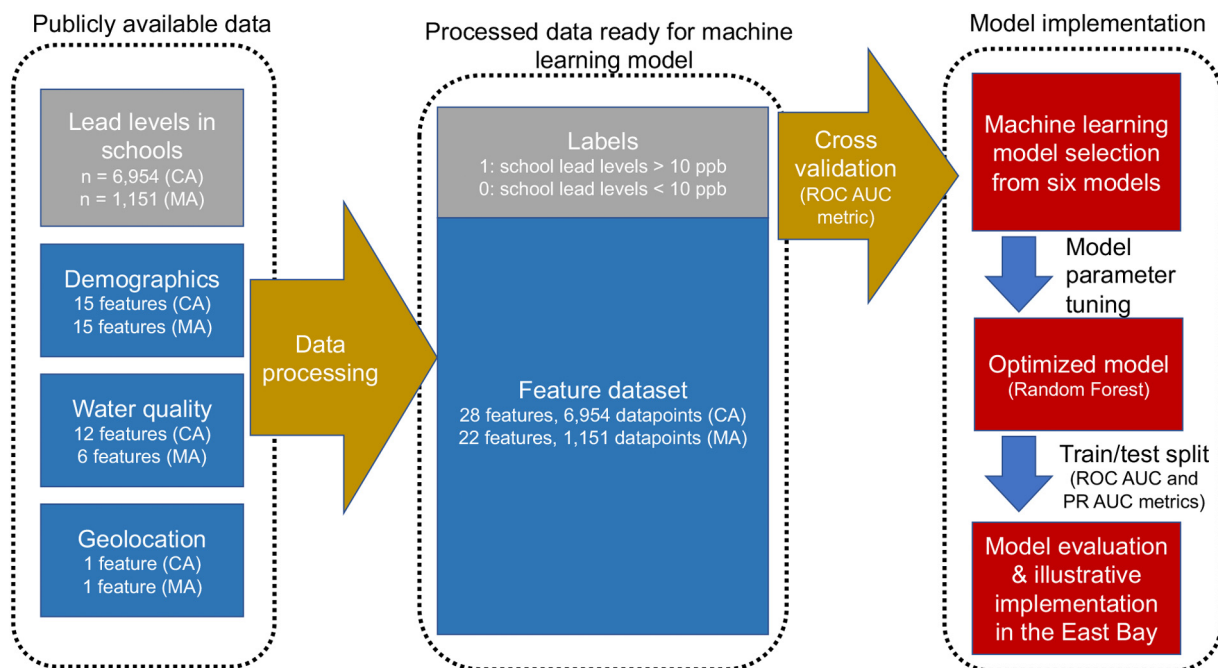


Fig. 3. Schematic overview of the steps taken to process the data and implement the machine learning model. Note that each datapoint in the model corresponds to a school's lead levels (above or below a 10-ppb threshold) and all its associated features (demographic, chemical and geographic) adding up to 28 features for each CA datapoint, and 22 features for each MA datapoint.

Bayes, and Decision Trees) were implemented using the features described in Section 2.2 and the binary labels corresponding to lead levels below and above a threshold of 10 ppb. For more information on how each model works, see the S.I.

We briefly introduce a method and two metrics used to evaluate machine learning models.

- (1) K-fold cross validation is a commonly used method to test how a predictive model will perform in practice with data. Predictive models are typically calibrated with a “training” data set, and their performance must be tested against data from outside this training set. In cross validation, the original data is separated into k independent (i.e., non-overlapping) subsets and then the model is trained with only $k-1$ subsets. The trained model is then tested using the withheld subset. This process is repeated k times by successively withholding a different subset for testing each time, effectively creating k instances of the model that is trained and tested using k different training and testing datasets.
- (2) The “Receiver Operating Characteristic Area Under the Curve” (ROC AUC) metric, which ranges from 0.5 to 1, is commonly used to evaluate the ability of the model to distinguish between True Positives (TP) and False Positives (FP) for different probability thresholds. Thus, ROC AUC values close to 1 indicate that the model can perfectly distinguish between TPs and FPs, while values close to 0.5 indicate that the model is no better than random selection. TPs in our case mean that the model predicts that a school has lead levels over 10 ppb, and the measured value for that school is indeed over 10 ppb. FPs are those cases in which the model predicts that a school has lead levels over 10 ppb however, the actual (measured) lead levels are below 10 ppb.
- (3) The “Precision Recall Area Under the Curve” (PR AUC) metric, which ranges from 0 to 1, is typically used to evaluate the ability of the model to distinguish between TPs and False Negatives (FN) for different probability thresholds. PR AUC values close to 1 indicate that the model can perfectly distinguish between TPs and FNs, while values close to 0 indicate that the model is no better than random selection. FNs are those cases in which the model predicts that a school has lead levels below 10 ppb however, the actual (measured) lead levels are over 10 ppb.

We assessed the performance of the six aforementioned machine learning models by optimizing their respective hyperparameters (the models’ internal parameters) using a 10-fold cross validation. This process consisted of iterating through multiple combinations of hyperparameters and finding those that provided the highest cross validation scores. The ROC AUC metric was used to select the best-performing model (the one with the highest average ROC AUC score for all 10 folds). The methodology used to optimize the hyperparameters, as well as their optimized values are shown in the S.I.

2.3.3. Model selection and feature importance analysis

The best performing model, based on the cross-validation analysis, was used to assess the relevance of all input features in the model. Moreover, we analyzed the importance of each of the three feature types (chemical, demographic, and spatial) by removing all features from one feature type at a time, and then assessing the model performance with only two of the three types of features (e.g., we removed all the chemical features and tested the performance of the model using only the demographic and spatial features). The performance was assessed by splitting the data randomly into a 70% training and a 30% testing dataset by using the *train_test_split* function in the Python *scikit-learn* package. The data was split randomly 1000 times, and the ROC AUC and PR AUC was then calculated for each split. Given the importance of accurately identifying locations with a high risk of lead contamination and the unbalanced nature of our dataset (only about 20% of schools have lead levels over 10 ppb), the PR AUC metric provides

insights into the ability of the model to accurately predict the minority class labels and to avoid predicting FNs.

The compiled datasets used in this study, as well as the code used to implement the machine learning models may be found in our Github repository (Lobo et al., 2021).

2.4. Practical examples of model implementation

2.4.1. Implementation in the East Bay

To test the usefulness of the model in cities with where lead testing is incomplete, the California dataset was separated into two subsets: (1) half of the schools within the East Bay (a part of the San Francisco Bay Area) and (2) the rest of the schools in the state. The best-performing model was trained with set (2) and then tested using set (1). The schools in set (1) were chosen randomly. This was done to simulate a scenario in which not all the schools in a single location have been tested for lead, which is the practical scenario in which we envision this model being used to direct precious testing resources. However, instead of using a binary output to identify individual schools at risk, as explained in Section 2.3, we computed the probability of lead levels exceeding 5 ppb for every school in the East Bay. This method allowed us to categorize each of the schools in the test set as having a simulated low, medium, and high probability of lead concentrations in drinking water exceeding 5 ppb. We defined the modeled probability of lead concentration being above 5 ppb as “low” if below 0.3, as “medium” if between 0.3 and 0.7, and as “high” if over 0.7 (e.g., if a school has a simulated probability of lead concentrations exceeding 5 ppb of 0.1, then it is categorized as “low” risk because its modeled probability is below 0.3). Thus, the model assigned each school a defined modeled probability, and therefore, an assigned risk category (low, medium, and high). Then we counted the actual number of schools observed to have average lead water levels over 5 ppb, as well as the total number of schools within each category. This allowed us to compute, for each risk category, the actual “fraction of schools with lead levels exceeding 5 ppb”. As the wording implies, for each category, this fraction was simply the actual number of schools with lead water levels over 5 ppb divided by the total number of schools in that risk category. We compared the actual fractions, and predicted average probabilities, to assess the predictive power of the model in a single area rather than the entire state. We note that we chose the East Bay to illustrate the model implementation in a particular area. We could have chosen any other city or town in California or Massachusetts.

2.4.2. Implementation in California

The procedure described in Section 2.4.1 (train the model using data from the state excluding half the schools of a single city or town, and then predict the probability of school lead leaching for the excluded schools) was deployed to cover all cities and towns in CA (towns with only one school were ignored). The modeled probability that a school has average lead water levels over 10 ppb was computed for every school excluded from the training sets. Using these probabilities, we categorized each of the predicted school probabilities into low, medium, and high risk of lead leaching following the same procedure described in Section 2.4.1. These probabilities were then compared to the actual fraction of schools with lead levels exceeding 10 ppb within each risk category.

3. Results and discussion

3.1. Model selection and parameter significance

Random Forest (RF) outperformed all other models tested in this study when applied to both the California and Massachusetts datasets (see Table 1). RF has been used to predict lead levels in drinking water in Flint, MI (Chojnacki et al., 2017), arsenic levels in groundwaters in Bangladesh (Tan et al., 2020) and nitrate levels in groundwater

Table 1

Mean and standard deviations of the 10-fold cross validation using the Area Under the Receiver Operating Characteristic curve (ROC AUC) of different machine learning models for the California and Massachusetts lead databases. The optimized hyperparameters were used in all cases (see S.I). Random forest is the best performing model when implemented with data from both states.

Model	Mean ROC-AUC score		Standard deviation	
	CA data	MA data	CA data	MA data
Random Forest	0.78	0.88	0.03	0.02
Decision Tree	0.76	0.76	0.09	0.04
k-Nearest Neighbor	0.75	0.72	0.04	0.06
Logistic regression	0.70	0.68	0.05	0.05
SVM	0.70	0.70	0.04	0.06
Naïve Bayes	0.68	0.60	0.07	0.09

(Wheeler et al., 2015), among others. Moreover, its ease of use and comparable results to other machine learning models, such as neural networks, makes it better suited for use by non-experts (Sameen et al., 2019). This study further confirms that RF may be used to predict complex and multifactorial water quality parameters from relevant features, likely because of its ability to generate non-linear decision boundaries. In fact, the best-performing models, RF, decision trees and kNN, are all non-linear classifiers. In contrast, the worst-performing models, SVM, Naïve Bayes and logistic regression, are all linear classifiers.

We emphasize that our approach is not meant to provide a causal relationship between the different input features and the risk of lead leaching in schools. The selection of the input features comes from prior expertise and insights in relevant domains (water chemistry, environmental engineering knowledge related to lead in urban drinking water, urban socio-economic history, and geospatial differentiation) by those developing the model. Thus, even though the model uses socioeconomic information, this does not mean (and nor do we support a misinterpretation) that socioeconomics plays a direct and mechanistic (i.e., causal) role in the lead leaching process.

3.2. Model performance

The performance of the optimized RF model is good in both Massachusetts (average ROC AUC = 0.88) and California (average ROC AUC = 0.77), as shown in Fig. 4. These results were obtained after setting the number of trees to 1000 and 500, and the maximum tree depth to 12 and 9 in the CA and MA models, respectively (the rest of the RF hyperparameters were set to the *sklearn* package default values). As a

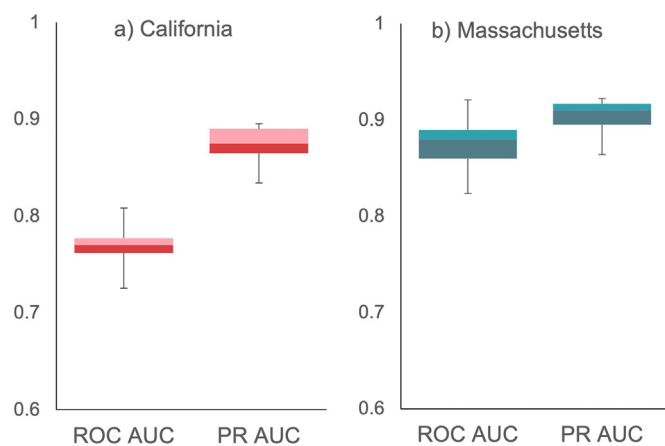


Fig. 4. Box plot of the Area Under the Receiver Operating Characteristic (ROC AUC) and Precision-Recall (PR AUC) curves for 1000 instances of the model in (a) California and (b) Massachusetts. The performance of the model is very good in California, and excellent in Massachusetts. As applied to both the states, the RF models predict more false negatives than false positives, as shown by the large PR AUC values.

reference, to the best of our knowledge, the only other machine learning model that has been used to predict lead levels has been applied to a much more dense, detailed, and richer dataset from a single city (Flint, MI), and nevertheless, it led to an average ROC AUC of 0.72 (Chojnacki et al., 2017) (using a 15 ppb threshold). Thus, our model is more capable over a much larger area, that of the full state, with larger diversity in geography, socio-economic status, and water chemistry, while providing more accurate results. Notably, the RF model in both California and Massachusetts provides precise results. This is shown by the high PR AUC values (average PR AUC is 0.88 for California and 0.91 for Massachusetts). High PR AUC values indicate that the model can accurately distinguish TP from FP. The discrepancy between ROC and PR AUC values is caused by the tendency of the model to predict more FPs than FNs. This is a desirable outcome when predicting lead leaching, given the high social cost of predicting FNs. FNs would be a harmful outcome from a public health perspective, since such schools could be assigned on a lower priority for further monitoring and testing, in a resource-constrained situation.

A major concern when implementing machine learning models is to identify whether the amount of data (i.e., the number of datapoints) used is enough to provide representative results. In the case of ensemble models utilizing bootstrap aggregation, such as RF, the Out-of-Bag error (OOB error) is a metric that is commonly used to evaluate the model stability as a function of datapoints used. The OOB error represents the mean error obtained from evaluating the model's performance on a subset of data that was excluded from the bootstrap sample used to train the model, akin to a cross-validation discussed earlier. Our model's OOB error becomes stable when the number of datapoints (schools) is larger than 400. This was true for both California and Massachusetts models (See S.I). The numbers of datapoints in both State datasets are substantially larger than 400 (over 6000 in California and over 1000 in Massachusetts), therefore it is likely that the error will remain stable as more points are added.

3.3. Feature importance analysis

The distance to the nearest school with average lead levels of over 10 ppb, is the most important feature in the model for both California and Massachusetts, as shown in Fig. 5. This was expected because this distance may incorporate at least four different kinds of relevant information: (1) Water quality conditions existing at a local scale are likely to be similar. Lead leaching is usually associated to corrosive water quality conditions (Schock, 1990), thus, the distance to the nearest school with high lead levels may act as a proxy of the corrosivity of the water at the school being evaluated. The water quality at a local scale may be significantly different from the conditions reported at the water treatment facility, because water quality can and often does change throughout the distribution system as a result of leaks, consumption of the disinfection residual, and consumption of orthophosphates, among others (Sert and Altan-Sakarya, 2017). (2) Similar construction dates may signal similar materials in water transport to the school. Nearby schools are more likely to have been constructed in similar years, thus, if a school was constructed previous to 1986 (before lead-bearing plumbing materials were banned), it is likely that a nearby school was also constructed previous to that year. (3) Nearby schools may belong to the same administration and be subject to similar policies and contracts for purchasing water fountains. And lastly (4) nearby schools may have similar socioeconomic conditions. Given that old schools located in low-income neighborhoods tend to have a higher risk of lead leaching (Lambrinidou et al., 2010), it is likely that many schools within the same low-income communities will have higher risk of lead leaching.

The socioeconomic features are the second most important variables in the model. This is consistent with prior published research showing that poverty rates and race are correlated to the presence or absence of lead-bearing plumbing materials (Sampson and Winter, 2016;

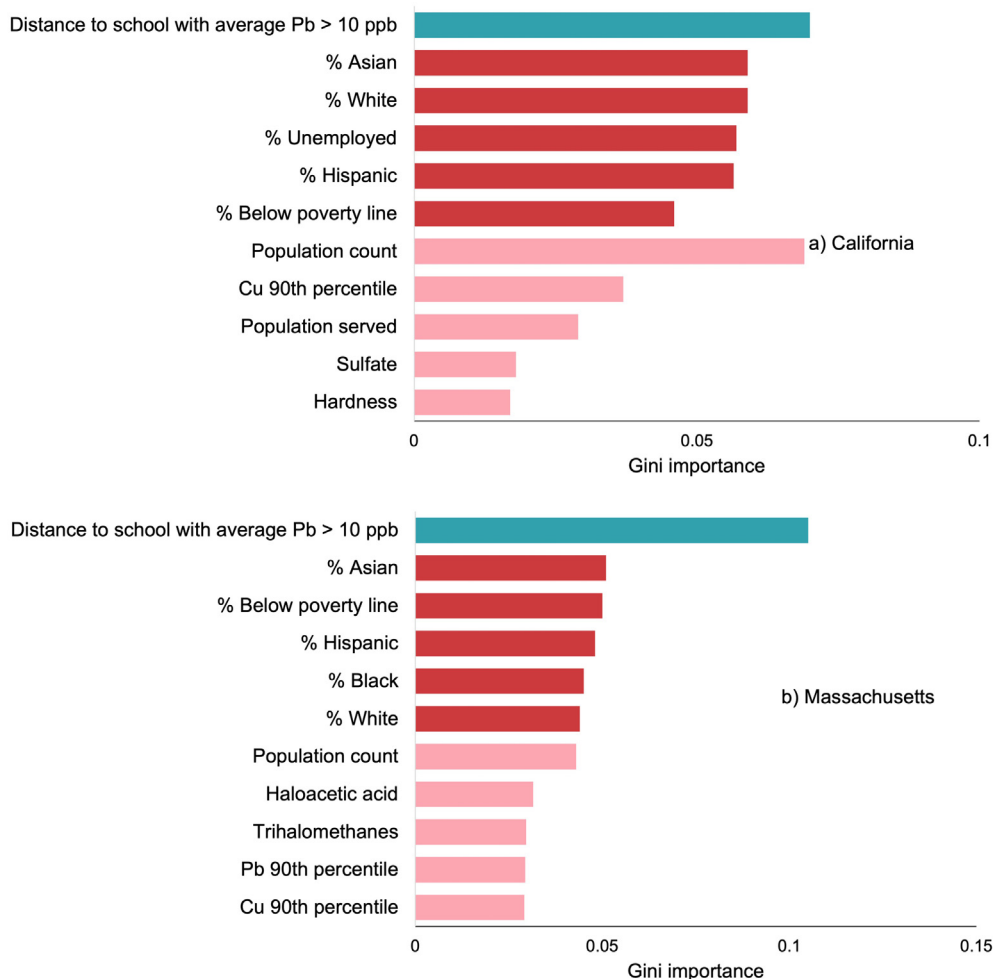


Fig. 5. Feature importance analysis using the Gini importance for the random forest model implemented in (a) California and (b) Massachusetts. The Gini importance represents the loss in entropy (statistical dispersion) resulting from adding each feature to the model. Out of the 94 features used in the California model and the 88 used in the Massachusetts model, only the five most important social and chemical features (red and pink, respectively), and the only spatial feature (green) are shown. The spatial feature contributes the most to the model in both states. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Gleason et al., 2019). Like the distance to the nearest school with average lead levels over 10 ppb, the socioeconomic variables provide an estimate of how lead-bearing plumbing materials are distributed throughout geographical space. The presence of lead-bearing plumbing materials is a necessary condition for lead leaching to take place. Thus, it is expected that any feature that correlates to the presence of such materials will be a good contributing feature for predicting lead contaminated water in schools.

The chemical features contribute the least towards reducing the entropy of the model, likely because they are the same for every school within the same water system (water utilities issues a single Consumer Confidence Report (CCR) for the entire water system under its purview). However, the “population count” feature, which indicates the number of people served by each water utility, was important in the CA dataset. This feature is likely a good predictor of lead in drinking water in schools because very small water systems (<500 people) are more likely to violate health-related water quality regulations than larger systems (Rubin, 2013). This is particularly true for small water systems in rural California (Balazs and Ray, 2014). Given that the chemical features are constant for each water utility, they do not explain the variability within any given water system, but instead provide a water quality baseline for each city. This allows representing the fact that cities with similar demographics (or with the same distribution of lead-bearing plumbing materials), will have different lead leaching levels if

their water quality is different. However, we also note that the water quality dataset had the largest amount of missing information, which may have affected its relevance in the model. A similar feature importance analysis using the regression coefficients of a logistic model is shown in the S.I.

Retraining the model with any two of the three feature types (i.e., chemical and social, social and spatial, or spatial and chemical) highlights the importance of using all three to optimize performance, as shown in Fig. 6. When using only chemical and socioeconomic data, the mean ROC AUC drops over 25% in both states (mean ROC AUC decreases to 0.75 and 0.65 for Massachusetts and California, respectively), making the model inadequate for use in California. In the case of Massachusetts, the ROC AUC decreases to a level comparable to the model developed by Chojnacki et al. for the city of Flint, MI (Chojnacki et al., 2017). This supports the idea that socioeconomic features by themselves may be good predictors of lead contamination in individual cities served by a single water utility, like Flint, MI. However, extrapolating results to a state level requires the use of chemical information, as the baseline water quality is different for different water systems. This is because cities with highly corrosive tap water can be expected to have higher lead leaching levels in schools than cities where corrosion control strategies are implemented, everything else being held equal.

Using only the social and spatial data, the ROC AUC decreases by 15% in both states (mean ROC AUC is 0.79 and 0.72 for Massachusetts and

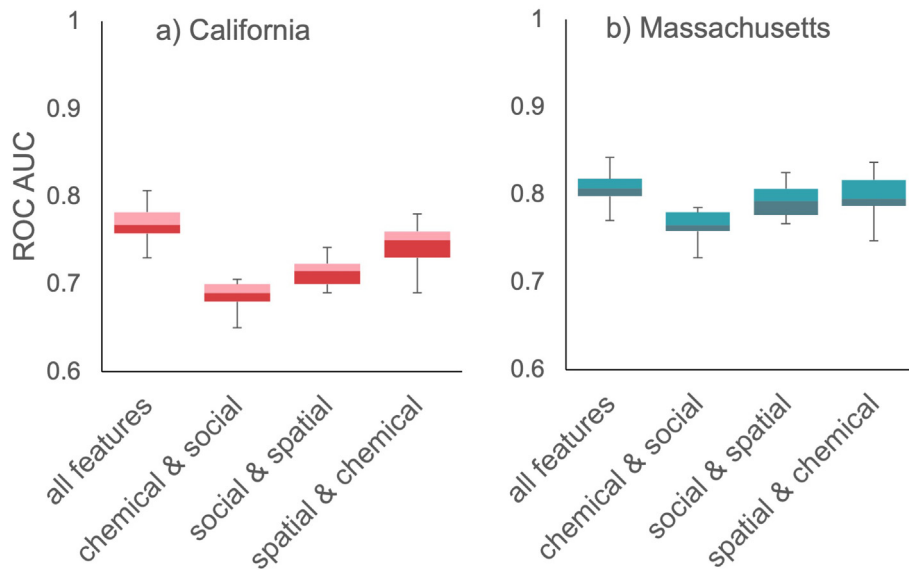


Fig. 6. Box plot of the Area Under the Receiver Operating Characteristic (ROC AUC) curve for 1000 instances of the model in (a) California and (b) Massachusetts when using different feature types (chemical, social and spatial) to train and test the model. The model performance is best when all three kinds of data are used; however, good results are also observed with the use of only spatial and chemical data, for both states.

California, respectively). Even if these results are better than those obtained when using the chemical and socioeconomic data, the performance in California is still borderline adequate. This supports our hypothesis that the chemical data is necessary as a baseline that adjusts how the model responds to spatial and social variables. Both the presence of lead-bearing plumbing materials and drinking water quality conducive to lead corrosion are needed for lead leaching to take place, thus, features that predict both factors are needed to predict the likelihood of lead leaching in schools.

When using only the spatial and chemical data, the mean ROC AUC decreases by less than 5% in both states (mean ROC AUC is 0.84 and 0.74 for Massachusetts and California, respectively); however, the standard deviation increases, as shown in Fig. 6. It is likely that the spatial and the social data provide redundant information in many cases, as the former may provide information on three different underlying causal factors: local water quality conditions, socioeconomic conditions among similar close-by schools, and the presence or absence of lead-bearing plumbing materials (Wescoat et al., 2007). However, the social data decreases the variability of the model and slightly increases the average performance. This supports our hypothesis that all three kinds of variables, social, chemical, and spatial, are relevant to model lead leaching at a local level, as they provide insights into different phenomena. The socioeconomic data provide information on the presence of lead-bearing plumbing materials, while the chemical data provide insights into the corrosivity of the water. Finally, the spatial data may provide both kinds of information.

3.4. Practical examples of model implementation

3.4.1. Implementation in the East Bay

We tested the ability of the model to assess the risk of lead leaching in a scenario where only partial lead data is available for a single city, but other data exist for other cities within the same State (see Section 2.4.1 for more details). As shown in Fig. 7a, the model accurately categorized the risk of above-threshold lead concentrations in over 87% of the schools in the East Bay that were not included in the training set. In the figure there is good agreement between the predicted probabilities and actual measured fractions of schools with lead concentrations above the 5-ppb threshold. Recall that these are defined as the probability that a certain school categorized as low, medium, and high risk

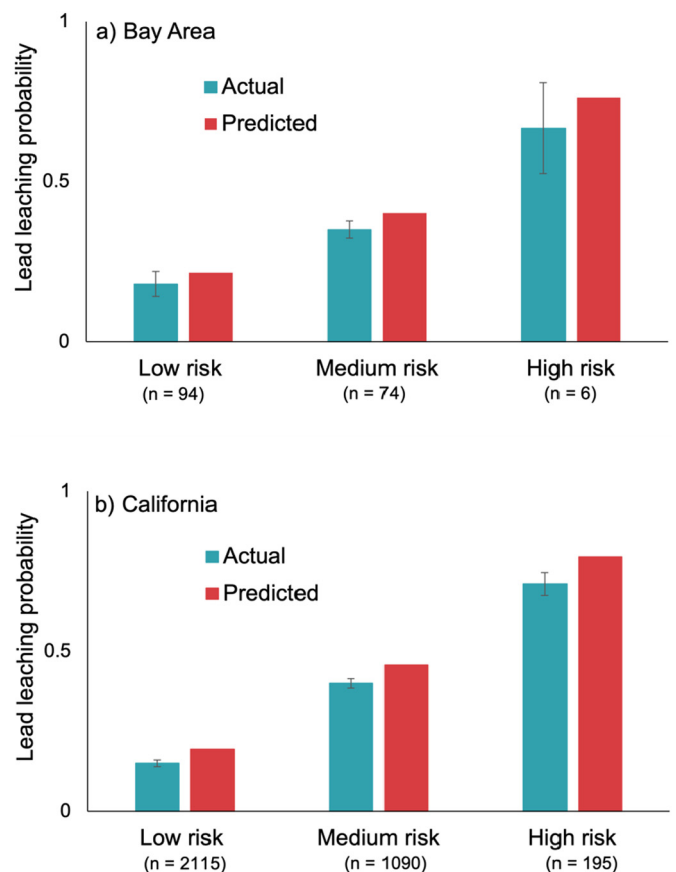


Fig. 7. Actual fraction and predicted probability of (a) lead in drinking water levels exceeding 5 ppb in half of the schools within the East Bay (a part of the San Francisco Bay Area) and (b) lead in drinking water levels exceeding 10 ppb in half of the schools in California, for three modeled risk categories: low, medium and high. Each risk category corresponds to modeled lead leaching probabilities of less than 0.3 (low), between 0.3 and 0.7 (medium), and over 0.7 (high). The model is in good agreement with the observed data for each risk category; however, it tends to slightly overpredict the average likelihood of lead leaching for each category. Error bars were added to the observed (actual) fraction in each category to account for noise caused by finite sample size.

predicted by the model, will have average lead leaching levels over 5 ppb. These good results were obtained even though only 50% of the schools in the East Bay (and the rest of the cities in CA) were used to train the model. We note that the model tends to slightly overpredict the risk of lead leaching in schools, which is a bias towards predicting more FP than FN (see Section 3.2). This, however, is desirable from a public health perspective where the cost of a FN is higher than that of a FP.

Using the probabilities shown in Fig. 7a and the total number of schools within each risk category, we calculated that, out of the 176 schools in the East Bay where the model was used to predict the risk of lead leaching, 54 are expected to have leaching levels of over 5 ppb. This is close to the 48 schools in this subset of schools where lead levels exceeded 5 ppb in 2018.

3.4.2. Implementation in California

The same approach used to test the model in the East Bay was used to test its applicability in every city and town in CA, but this time using a 10-ppb threshold. As shown in Fig. 7b, the model accurately categorized the risk of lead levels exceeding 10 ppb of most schools in CA that were not included in the test set (like in the previous case, half of the schools in each city were excluded during training). Using the predicted lead leaching probabilities and the total number of schools within each risk category, we calculated that out of the 3748 schools in CA where the model was used to predict the risk of lead leaching, 551 are expected to have leaching levels of over 10 ppb. This is reasonably close to the 423 schools in this subset of schools where average lead concentrations in water exceeded 10 ppb in 2018.

The results shown in Fig. 7 give an example of how this approach may be used to identify schools where lead testing should be a priority. This approach may be used to identify schools at high risk of lead leaching so that limited resources may be deployed more efficiently. Of course, these results for the East Bay and other towns and cities in CA do not provide any new information regarding which schools should be tested, because (as noted earlier in this paper) most of the schools in these two states have already been tested for lead concentrations in their water. However, we note that charter and private schools in California are not required to test their drinking water for lead. For those schools, this model could be used to predict their risk of lead leaching so that appropriate actions can be taken.

These results provide a testing ground for the proposed methodology, since the model accurately categorized the lead leaching risk of over 85% of the schools in California when relying on incomplete publicly available datasets. To our knowledge, this model is the first of its kind to predict the risk of lead leaching in geographical areas where the lead source is not predominantly from lead service lines (lead leaching in schools is usually caused by the use of leaded brass plumbing (Triantafyllidou and Edwards, 2012; Boyd et al., 2008)). Moreover, our approach could potentially be used to identify schools at risk anywhere in the US where partial school lead water surveys exist (this remains to be tested). Lastly, our results show that the model provides accurate predictions (for risk of lead in school waters) in specific locations, even when trained using state-level data. This, to the best of our knowledge, has not been previously reported in literature. This success is likely a result of the relying on features that capture the two main components of lead leaching: the presence of lead-bearing plumbing materials and drinking water quality conducive to lead corrosion. By (partially) capturing these phenomena, it becomes easier to extrapolate the model to different cities, unlike existing models that mostly use socioeconomic features to predict lead leaching.

3.5. Model challenges and limitations

As stated earlier, the model reported in this work does not intend to provide a mechanistic explanation of the factors that govern lead leaching. The nature of the RF development process, by training on

diverse data, incorporates the complex non-linear correlations of lead leaching with socioeconomic, chemical, and spatial features. We chose these features based on our knowledge and expertise gained from study of various relevant fields relevant to the lead leaching process, and relevant to how different mechanistic variables correlate to publicly available data. Like many prior studies, our model too captures the unfortunate fact that low-income, nonwhite communities are more likely to be exposed to lead in drinking water. This fact is entirely separate from the chemistry of lead dissolution. We emphasize that, if progress is made towards an equitable and fair access to safe drinking water, the accuracy of our model should decrease. We included social variables because they are currently good predictors of lead-bearing infrastructure, thus, the reported accuracy is meant to show a picture of today and not of some immutable laws of nature. In an ideal world, race and income-levels should not be correlated with lead contamination of drinking water, and thus, a model to predict lead leaching should not depend on such social data; however, that is not currently the case.

The accuracy of our model is also likely limited by the quality of the chemical data. We used all the available information provided in the water utility CCRs; however, the amount of relevant information is limited. For instance, most of the utilities in California and Massachusetts do not report whether they use corrosion control (e.g., add orthophosphates to the water supply) or other strategies to prevent lead leaching, therefore, these variables could not be included in the model. Furthermore, the CCRs are often hard to find and vary greatly in terms of quality and content among different utilities. This makes the construction of a water quality database a slow and painstaking process prone to human error. Thus, the performance of the developed model might be hindered both by the lack of relevant water quality data and by human errors when transcribing the CCRs into the database.

4. Conclusions

Although machine learning models meant to predict lead leaching in drinking water have been developed for individual cities, to our knowledge no model has attempted to predict lead leaching in schools at the state level. In this study we provided a methodology to build such a model from publicly available datasets and we tested it using data from California and Massachusetts. We found that using water-chemistry, socioeconomic, and spatial data provided the best results despite the low quality of the publicly available water-chemistry data. This research also highlights the need of a nation-wide database for drinking water quality, akin to the existing water quality database created by the US Geological Survey for water bodies across the US, that is compatible with the data needs of the 21st century.

Our results suggest that applications in states are also possible where lead testing is still incomplete or seriously lacking, since the model predictions are useful to predict high-risk schools even with incomplete state level data. Moreover, the data from the newly tested schools can then be added to the training data for the model, successively improving its accuracy. Our results suggest that machine learning can play a significant role in designing future urban water management strategies to support equitable access to safe drinking water for all, supporting the goals of Environmental Justice related to safe drinking water, and aligned with Sustainable Development Goal 6.

CRedit authorship contribution statement

G.P. Lobo: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **J. Laraway:** Methodology, Software, Investigation, Data curation, Writing – review & editing, Visualization. **A.J. Gadgil:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing 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.

Acknowledgements

The authors are grateful for key funding support that enabled this work. This includes support from the Fulbright Program, the Hellman Foundation, Prof. Gadgil's Rudd Chair funds, and a gift to support this research from the Barbara and Gerson Bakar Foundation. We thank Joseph Mella and David Sanchez, undergraduate students at UC Berkeley, for participation in data mining and analysis. Some of the effort for this work was enhanced by the InFEWS program at UC Berkeley, supported by the National Science Foundation under Grant No. DGE-1633740.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.scitotenv.2021.150046>.

References

- Abernethy, J., Chojnacki, A., Farahi, A., Schwartz, E., Webb, J., 2018. Active remediation: the search for lead pipes in Flint, Michigan. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 5–14 <https://doi.org/10.1145/3219819.3219896>.
- Abokifa, A.A., Biswas, P., 2017. Modeling soluble and particulate lead release into drinking water from full and partially replaced lead service lines. *Environ. Sci. Technol.* 51, 3318–3326.
- Agency, C.N.R., 2021. Lead Sampling of Drinking Water in California Schools: Lab Results. <https://data.cnra.ca.gov/dataset/drinking-water-results-of-lead-sampling-of-drinking-water-in-california-schools>.
- Aisopou, A., Stoianov, I., Graham, N.J.D., 2012. In-pipe water quality monitoring in water supply systems under steady and unsteady state flow conditions: a quantitative assessment. *Water Res.* 46, 235–246.
- Balazs, C.L., Ray, I., 2014. The drinking water disparities framework: on the origins and persistence of inequities in exposure. *Am. J. Public Health* 104, 603–611.
- Boyd, G.R., Pierson, G.L., Kirmeyer, G.J., English, R.J., 2008. Lead variability testing in Seattle public schools. *J. Am. Water Work. Assoc.* 100, 53–64.
- Brownlee, J., 2017. Why One-Hot Encode Data in Machine Learning? *Machine Learning Mastery*. <https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>.
- Burlingame, G.A., et al., 2018. Lessons learned from helping schools manage lead in drinking water to protect children's health. *J. Am. Water Works Assoc.* 110, 44–53.
- California Water Boards, 2017. Collecting Drinking Water Samples for Lead Testing At K-12 Schools.
- Canada, H., 2019. Guidelines for Canadian Drinking Water Quality. National Meeting - American Chemical Society, Division of Environmental Chemistry. 24.
- Charisiadis, P., et al., 2015. Spatial and seasonal variability of tap water disinfection by-products within distribution pipe networks. *Sci. Total Environ.* 506–507, 26–35.
- Chojnacki, A., et al., 2017. A data science approach to understanding residential water contamination in flint. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. Part. F1296*, pp. 1407–1416.
- Cornwell, D.A., Brown, R.A., Via, S.H., 2016. National survey of lead service line occurrence. *J. Am. Water Works Assoc.* 108, E182–E191.
- Desantis, M.K., Schock, M.R., Tully, J., Bennett-Stamper, C., 2020. Orthophosphate interactions with destabilized PbO₂ scales. *Environ. Sci. Technol.* 54, 14302–14311.
- Dignam, T., Kaufmann, R.B., LeSturgeon, L., Brown, M.J., 2019. Control of lead sources in the United States, 1970–2017. *J. Public Health Manag. Pract.* 25, S13–S22.
- Edwards, M., Triantafyllidou, S., 2007. Chloride-to-sulfate mass ratio and lead leaching to water. *J. Am. Water Work. Assoc.* 99, 96–109.
- EPA, 2002. Effects of Water Age on Distribution System Water Quality.
- EPA, 2021. SDWIS Federal Reports Advanced Search. <https://ofmpub.epa.gov/apex/sfdw/f?p=108:1:::1>.
- European Commission, 2018. Proposal for a Directive on the Quality of Water Intended for Human Consumption.
- Executive Office of Energy & Environmental Affairs, 2021. Lead and Copper Drinking Water Results in Schools/Childcare. <https://eeonline.eea.state.ma.us/portal#/1/search/leadandcopper>.
- Fasaee, M.A.K., et al., 2021. Developing a framework for classifying water lead levels at private drinking water systems: a Bayesian belief network approach. *Water Res.* 189, 116641.
- Gleason, J.A., Nanavaty, J.V., Fagliano, J.A., 2019. Drinking water lead and socioeconomic factors as predictors of blood lead levels in New Jersey's children between two time periods. *Environ. Res.* 169, 409–416.
- Hanna-Attisha, M., LaChance, J., Sadler, R.C., Schnepf, A.C., 2016. Elevated blood lead levels in children associated with the flint drinking water crisis: a spatial analysis of risk and public health response. *Am. J. Public Health* 106, 283–290.
- Hauptman, M., Bruccoleri, R., Woolf, A., 2017. An update on childhood lead poisoning. *Clin. Pediatr. Emerg. Med.* 18, 181–192.
- Jackson, V., Johnson, N., 2021. America's School Infrastructure Needs a Major Investment of Federal Funds to Advance an Equitable Recovery. *Cent. Budg. Policy Priorities*.
- Katner, A., et al., 2016. Weaknesses in federal drinking water regulations and public health policies that impede lead poisoning prevention and environmental justice. *Environ. Justice* 9, 109–117.
- Kim, E.J., Herrera, J.E., Huggins, D., Braam, J., Koshowski, S., 2011. Effect of pH on the concentrations of lead and trace contaminants in drinking water: a combined batch, pipe loop and sentinel home study. *Water Res.* 45, 2763–2774.
- Korshin, G.V., Ferguson, J.F., Lancaster, A.N., 2000. Influence of natural organic matter on the corrosion of leaded brass in potable water. *Corros. Sci.* 42, 53–66.
- Kunapuli, A., Whiting-Hill, M., Beardsley, E., 2018. Perspectives on state legislation concerning lead testing in school drinking water. 36. U.S. Green Build. Council.
- Lambrinidou, Y., Triantafyllidou, S., Edwards, M., 2010. Failing our children: lead in U.S. school drinking water. *New Solut.* 20, 25–47.
- Lobo, G.P., Laraway, J., Gadgil, A.J., 2021. Lead in schools' drinking water data and code. https://github.com/gadgil-group/school_Pb_water.git.
- MassDEP, 2016. Sampling for Lead and Copper at Schools and Childcare Facilities.
- Masters, S., Welter, G.J., Edwards, M., 2016. Seasonal variations in lead release to potable water. *Environ. Sci. Technol.* 50, 5269–5277.
- NOAA, 2021. Climate at a Glance. National Centers for Environmental Information (NCEI).
- Noel, J.D., Wang, Y., Giammar, D.E., 2014. Effect of water chemistry on the dissolution rate of the lead corrosion product hydrocerussite. *Water Res.* 54, 237–246.
- Olson, E., Fedinick, K.P., 2016. What's in your water? flint and beyond. *Natl. Resour. Def. Council*.
- Rajasekharan, V.V., Clark, B.N., Boonsalee, S., Switzer, J.A., 2007. Electrochemistry of free chlorine and monochloramine and its relevance to the presence of Pb in drinking water. *Environ. Sci. Technol.* 41, 4252–4257.
- Ramaley, B.L., 1993. Monitoring and control experience under the lead and copper rule. *J. Am. Water Work. Assoc.* 85, 64–67.
- Redmon, J.H., Levine, K.E., Aceituno, A.M., Litzenberger, K., Macdonald, J., 2020. Lead in drinking water at North Carolina childcare centers: piloting a citizen science-based testing strategy. *Environ. Res.* 183, 09126.
- Rubin, S.J., 2013. Evaluating violations of drinking water regulations. *Am. Water Work. Assoc.* 105, 51–52.
- Sameen, M.I., Pradhan, B., Lee, S., 2019. Self-learning random forests model for mapping groundwater yield in data-scarce areas. *Nat. Resour. Res.* 28, 757–775.
- Sampson, R.J., Winter, A.S., 2016. The racial of lead poisoning: toxic inequality in Chicago neighborhoods, 1995–2013. *Du Bois Rev.* 13, 261–283.
- Schock, M.R., 1990. Factors governing the leaching rate of lead from plumbing materials. *Environ. Monit. Assess.* 15, 59–82.
- Schock, M.R., Cantor, A.F., Triantafyllidou, S., Desantis, M.K., Scheckel, K.G., 2014. Importance of pipe deposits to lead and copper rule compliance. *J. Am. Water Works Assoc.* 106, 87–88.
- Sert, Ç., Altan-Sakarya, A.B., 2017. Optimal scheduling of booster disinfection in water distribution networks. *Civ. Eng. Environ. Syst.* 34, 278–297.
- Stone, E.D., Duranceau, S.J., Lintereur, P.A., Taylor, J.S., 2010. Effects of orthophosphate corrosion inhibitor on lead in blended water quality environments. *Desalin. Water Treat.* 13, 348–355.
- Switzer, D., Teodoro, M.P., 2017. The color of drinking water: class, race, ethnicity, and safe drinking water act compliance. *J. Am. Water Works Assoc.* 109, 40–45.
- Switzer, J.A., Rajasekharan, V.V., Boonsalee, S., Kulp, E.A., Bohannon, E.W., 2006. Evidence that monochloramine disinfectant could lead to elevated Pb levels in drinking water. *Environ. Sci. Technol.* 40, 3384–3387.
- Tam, Y.S., Elefsiniotis, P., 2009. Corrosion control in water supply systems: effect of pH, alkalinity, and orthophosphate on lead and copper leaching from brass plumbing. *J. J. Environ. Sci. Health A Tox. Hazard. Subst. Environ. Eng.* 44, 1251–1260.
- Tan, Z., Yang, Q., Zheng, Y., 2020. Machine learning models of groundwater arsenic spatial distribution in Bangladesh: influence of holocene sediment depositional history. *Environ. Sci. Technol.* 54, 9454–9463.
- Triantafyllidou, S., Edwards, M., 2012. Lead (Pb) in tap water and in blood: implications for lead exposure in the United States. *Crit. Rev. Environ. Sci. Technol.* 42, 1297–1352.
- Triantafyllidou, S., et al., 2021. Variability and sampling of lead (Pb) in drinking water: assessing potential human exposure depends on the sampling protocol. *Environ. Int.* 146.
- Trueman, B.F., Camara, E., Gagnon, G.A., 2016. Evaluating the effects of full and partial lead service line replacement on lead levels in drinking water. *Environ. Sci. Technol.* 50, 7389–7396.
- United States Census Bureau, 2021. American Community Survey. <https://data.census.gov/cedsci/>.
- Wang, Y., Wu, J., Wang, Z., Terenyi, A., Giammar, D.E., 2013. Kinetics of lead(IV) oxide (PbO₂) reductive dissolution: role of lead(II) adsorption and surface speciation. *J. Colloid Interface Sci.* 389, 236–243.
- Wescoat, J.L., Headington, L., Theobald, R., 2007. Water and poverty in the United States. *Geoforum* 38, 801–814.
- Wheeler, D.C., Nolan, B.T., Flory, A.R., DellaValle, C.T., Ward, M.H., 2015. Modeling groundwater nitrate concentrations in private wells in Iowa. *Sci. Total Environ.* 536, 481–488.
- WHO, 2017. Guidelines for Drinking-water Quality.
- Zhao, J., et al., 2018. Formation and aggregation of lead phosphate particles: implications for lead immobilization in water supply systems. *Environ. Sci. Technol.* <https://doi.org/10.1021/acs.est.8b02788>.