Contents lists available at ScienceDirect

International Journal of Hygiene and Environmental Health

journal homepage: www.elsevier.com/locate/ijheh

Predicting childhood lead exposure at an aggregated level using machine learning

G.P. Lobo [*], B. Kalyan, A.J. Gadgil

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

ARTICLE INFO

Keywords:
Lead poisoning
Machine learning
Lead
Environmental exposure
Aggregated data

ABSTRACT

Childhood lead exposure affects over 500,000 children under 6 years old in the US; however, only 14 states recommend regular universal blood screening. Several studies have reported on the use of predictive models to estimate lead exposure of individual children, albeit with limited success: lead exposure can vary greatly among individuals, individual data is not easily accessible, and models trained in one location do not always perform well in another. We report on a novel approach that uses machine learning to accurately predict elevated Blood Lead Levels (BLLs) in large groups of children, using aggregated data. To that end, we used publicly available zip code and city/town BLL data from the states of New York (n = 1642, excluding New York City) and Massachusetts (n = 352), respectively. Five machine learning models were used to predict childhood lead exposure by using socioeconomic, housing, and water quality predictive features. The best-performing model was a Random Forest, with a 10-fold cross validation ROC AUC score of 0.91 and 0.85 for the Massachusetts and New York datasets, respectively. The model was then tested with New York City data and the results compared to measured BLLs at a borough level. The model yielded predictions in excellent agreement with measured data: at a city level it predicted elevated BLL rates of 1.72% for the children in New York City, which is close to the measured value of 1.73%. Predictive models, such as the one presented here, have the potential to help identify geographical hotspots with significantly large occurrence of elevated lead blood levels in children so that limited resources may be deployed to those who are most at risk.

1. Introduction

Childhood lead exposure is a problem that affects over 500,000 children under 6 years of age in the US (Hauptman et al., 2017). There is no safe level of lead in the bloodstream (Vorvolakos et al., 2016) and even lead levels of 1 μg dL^{-1} have been linked to permanent and irreversible cognitive damage in children under age 6 (Lanphear et al., 2000; Schwartz, 1994). Elevated Blood Lead Levels (BLLs) in US children often result from exposure to lead in paint, soil, dust, and water (Gould, 2009; Mielke and Reagan, 1998; Roy and Edwards, 2019). However, children with elevated BLLs are not evenly distributed in society: those living below the poverty line are four times more likely to have elevated BLLs than their richer counterparts (Vivier et al., 2011). Non-Hispanic Black children are particularly at risk (Whitehead and Buchanan, 2019).

The lifetime social cost of childhood lead exposure (lead blood levels over 1 μg dL^{-1}) is estimated to be $50,000 USD per child (Muennig, 2009). This cost includes all medical costs, loss in IQ, special education, and increased crime rates, among other consequences of low-level lead exposure. However, this estimate does not include the cost of other adverse effects, including immune, cardiovascular, renal, and developmental effects (U.S. Department of Health and Human Services, 2012). Thus, childhood lead exposure in the US is a $25 billion problem (cumulative cost) that not only permanently hinders the livelihood of thousands of children, but is also a preventable, yet persistent, matter of social and environmental justice (Ettinger et al., 2019).

Identifying children at risk is challenging because data on housing with lead-based paint and plumbing components, or with lead-contaminated soils and dust are scarce (Cattle et al., 2002; Mielke, 1999; Triantafyllidou and Edwards, 2012). Instead, socioeconomic features are often used as predictors of elevated BLLs as they account for the unfortunate fact that poor minorities are more likely to live in older housing with multiple lead sources (lead paint and plumbing were commonly used in housing prior to 1978) (Marshall et al., 2020).

These socioeconomic features have been integrated into statistical models meant to predict the risk of lead exposure of individuals and

[*] Corresponding author. 410 O'Brien Hall, Berkeley, CA, 94720, United States.
E-mail address: gplobo@berkeley.edu (G.P. Lobo).

https://doi.org/10.1016/j.ijheh.2021.113862
Received 12 July 2021; Received in revised form 8 October 2021; Accepted 13 October 2021
Available online 18 October 2021

communities; however, their success has been limited. For instance, Taylor et al. (2013) found that, while socioeconomic factors are correlated to elevated BLLs in pregnant women, using a logistic regression to predict their individual exposure risk did not provide accurate results ($R^2 = 0.1$). Bierkens et al. (2011) concluded that even if environmental lead concentrations in air, soil, and water are known, linear regressions are not suited to estimate the average risk of lead exposure of select EU countries. This is due to the multifactorial, non-linear, and region-specific nature of the problem, compounded by the lack of available data (Lanphear et al., 1998).

Machine learning is particularly well-suited for complex nonlinear problems in which traditional statistical methods fail. Machine learning has been used to predict lead concentrations in air (Sethi and Mittal, 2019), water (Chojnacki et al., 2017), and soil (Zhang et al., 2020), as well as the likelihood of housing hazards, including lead paint (Ye et al., 2019). However, this approach has seldom been tested to predict the risk of childhood lead exposure. To the best of our knowledge, only two machine learning models have been reported in literature for this purpose. Potash et al. (2015) developed a gradient boosting model using 2.5 million BLL tests from Chicago, IL, and household characteristics, including year of construction, physical condition, and number of housing units, among others. Socio-demographic characteristics were also included. Their best performing model resulted in a precision of 0.39 and a recall of 0.42 for individual children, (see section 2.4 for technical definitions of the terms precision and recall). Another study by Potash et al. (2020) used a random forest model to predict childhood elevated BLLs. Their best-performing model had a ROC AUC of 0.69 (see section 2.4 for a technical definition of ROC AUC). The performance of these two pioneering machine learning models for childhood BLL predictions is better than that of traditional statistical methods; however, it is still below of that of other models used in other public health applications (Dos Santos et al., 2019). This is likely explained by the high resolution of the predictions, in which the risk of lead exposure of each individual child is predicted. Childhood lead exposure usually involves many environmental factors specific to each child that are difficult to measure (Lanphear et al., 2002), thus, predicting individual lead blood levels is challenging. Moreover, access to individual-based datasets is often limited, particularly for healthcare, where data are protected by patients' privacy laws (Wojtusiak and Baranova, 2011). This makes it hard not only to validate existing models, but to extend their use to locations outside of individual cities where these models are usually trained and tested.

In the context of designing strategies to prevent childhood lead exposure, such as large-scale lead blood testing or source removal programs, using individual-based models with modest predictive power to allocate resources might not provide optimal results. These programs often involve conducting lead blood tests and source removal at a neighborhood scale, and not on an individual basis (Billings and Schnepel, 2017; Magavern, 2018; Zahran et al., 2020). Thus, a model meant to allocate resources for lead blood testing and removal programs should be able to accurately identify geographical areas at risk rather than much less accurately identify individuals at risk (of course, accurately identifying all individuals with elevated BLLs would be even more preferable; however, no model is currently capable of this).

We hypothesize that using spatially aggregated data (e.g., zip code or city) could significantly improve the performance of existing models meant to predict elevated BLL in children, while still being valuable for designing strategies to prevent childhood lead exposure at a geographically large scale. This is because aggregated data often convey population trends that smoothen out the variability among individuals living under similar conditions, decreasing the noise of individual-based datasets (Rushton, 2003). Moreover, aggregated datasets are often public and readily available (Wojtusiak and Baranova, 2011), increasing data access, transparency, and restrictions on publishing results. Of course, the aggregation level should be relevant for the problem at hand: using a model that predicts city-wide risk of childhood lead exposure might not be as useful as a model that predicts zip codes at risk when designing a program meant to identify neighborhoods where children have elevated BLLs.

To date, there are no published studies that have attempted to process relevant features of aggregated BLL data to predict geographical areas at risk of childhood lead exposure. Thus, to our knowledge, prior literature is unclear whether BLL data aggregation increases the accuracy of childhood BLL predictive models at a spatially aggregated scale.

We report here on the use of machine learning to predict the risk of elevated BLLs in children at a spatially aggregated level. Using zip code and community-level socioeconomic and environmental data, we predicted the risk of childhood lead exposure for the states of New York and Massachusetts. This statistical model is *not* meant to provide a mechanistic understanding of how children are exposed to lead, but to help identify areas where they might be exposed to lead so that limited resources may be allocated more effectively.

2. Materials and methods

2.1. Study sites

New York and Massachusetts were chosen as study sites because they are two of the few states with publicly available BLL surveys in the US. The percentage of children under 6 years of age with elevated BLLs were obtained for 1642 zip codes in New York from the New York State Department of Health for the year 2015 (New York State Department of Health, 2015). These data were missing for all 178 zip codes in New York City because this city belongs to a separate health jurisdiction, NYC Health. In this study we used the Centers for Disease Control and Prevention (CDC) reference value of 5 μg dL^{-1} to determine whether children had elevated BLLs or not. We chose this value because it is the current CDC guideline and because the datasets provided by both states report only the number of children with BLL between 0 and 5, 5–10 and over 10 μg dL^{-1}. This reference value of 5 μg dL^{-1} was adopted in 2012 by the CDC based on the 97.5th percentile of the National Health and Nutrition Examination Survey (NHANES) blood lead distribution in children ages 1–5 years (CDC, 2021). Although more recent lead blood data suggest that this 97.5th percentile is closer to 3.5 μg dL^{-1} (Tsoi et al., 2016), the CDC has not yet updated their guideline value of 5 μg dL^{-1}. In the case of Massachusetts, the percentage of children with elevated BLLs were obtained for all 352 town/cities (also referred to as communities) in the state from the Massachusetts Department of Public Health, also for 2015 (Massachusetts Department of Public Health, 2021a).

We note that not all children within a community or zip code were tested for lead. In this study we assumed that the tested children were representative of all at-risk children within their respective zip codes or cities. BLL testing for children under 3 is mandatory in Massachusetts, after which BLLs are monitored only for children living in at-risk areas (Massachusetts Department of Public Health, 2021). In contrast, New York children are only tested if they are considered to be at risk (New York State Department of Health, 2021). Thus, the datasets used likely overrepresent children with elevated BLLs, which is desirable given the large social cost of failing to identify children at risk of having elevated BLLs.

While other states do report BLLs in children, many of them do so at a county level. Given that the robustness of machine learning models is often dictated by the amount of data used during training, working at a county level will likely not provide enough datapoints per state to develop a robust model. Furthermore, the low resolution of county-level BLLs cannot facilitate narrowly targeted intervention and mitigation strategies.

2.2. Data acquisition

Socioeconomic data for New York and Massachusetts were obtained from the 2015 American Community Survey (US Census Bureau, 2016). These data include, for each census tract, average income, percentage below the poverty line, percentage of property ownership, race, and ethnicity, among others. These socioeconomic features have been linked to elevated BLLs in numerous studies (Schultz et al., 2017; Trimble, 2016).

The 2015 Housing Price Indices (HPI) for each zip code were also obtained from the Federal Housing Finance Agency. HPIs represent the average price changes in repeat sales or refinancings on the same properties. We included HPI in the model because this index reflects, among many other things, housing construction year, which is an important feature given that old housings are more likely to have lead paint and plumbing (Whitehead and Buchanan, 2019). Housing construction year data were publicly available only for Massachusetts from the Massachusetts Department of Public Health, but not for New York.

Lead levels in drinking water were also obtained for each school in New York and Massachusetts from the New York State Department of Health and the Massachusetts Department of Public Health, respectively. Lead in drinking water in schools is an important source of lead exposure in children (Doré et al., 2018); however, it does not provide information on exposure in children under 4 or 5 years of age (children often start going to school at age 5). To the best of our knowledge, no household lead levels in drinking water at a state level are publicly available.

2.3. Data processing

A schematic overview of the steps taken to process the data and implement the machine learning model, described herein, is shown in Fig. 2. The data described in section 2.3 were combined with the BLL data using Geographic Information Systems (GIS). In the case of the New York dataset, the socioeconomic data were converted from a census tract scale to a zip code level by averaging the tract data within each zip code. The lead water levels in schools were first converted to average values per school district and then each zip code was assigned to a school district based on distance. This was done because not all zip codes have schools. Thus, the New York dataset consisted of 1643 datapoints, where each point corresponded to a zip code and all the associated data (percentage of elevated BLLs and socioeconomic, housing and lead water data, among others) with that geolocation. A total of 46 features were associated with each datapoint.

In the case of the Massachusetts dataset, the socioeconomic data were aggregated from a census tract level to a town/city level by taking the weighted average by population density. The HPI index was aggregated from a zip code level to a town/city level also using a weighted average. On the other hand, lead in drinking water levels in schools were averaged for all schools within a town/city. Thus, the Massachusetts dataset consisted of 352 points, all of which corresponded to towns/cities and their corresponding socioeconomic, demographic and water quality data. A total of 38 features were associated with each datapoint.

Both datasets were further processed by first one-hot-encoding (turning into binary values) all categorical features and by normalizing every feature so that all values range from 0 to 1. Missing values were filled in by using the mean value of each feature (e.g. missing income values were filled in using the average income of every city or zip code). Less than 5% of the data were missing for both New York and Massachusetts datasets.

The BLL data was also binarized by using a variable threshold T and the 5 μg dL^{-1} reference value established by the CDC. In the case of the Massachusetts dataset, a 1% threshold ($T = 1\%$) was established: if over 1% children within a town/city had elevated BLLs (over 5 μg dL^{-1}), then the value was set to 1; otherwise, it was set to 0. In the case of the New York dataset, a 6% threshold ($T = 6\%$) was established: if over 6% children within a zip code had BLLs over 5 μg dL^{-1}, then the value was set to 1; otherwise, it was set to 0. We binarized both datasets differently because of the different resolutions of the data and because the percentage of children with high BLLs per zip code in New York are much higher than the percentage per city in Massachusetts, as shown in Fig. 1.

2.4. Model implementation

Five machine learning models were implemented, and their performance compared using the processed Massachusetts and New York datasets: Random Forest, Logistic Regression, k-Nearest Neighbor (kNN), Decision Trees, and Support Vector Machine (SVM). These models were implemented using the Python *sklearn* package and their hyperparameters (model parameters) optimized using 10-fold cross validation (described below). The process of hyperparameter optimization consisted of iterating through multiple combinations of hyperparameters and finding those that provided the highest cross validation scores. A short description of each of the models, and their optimized hyperparameters are provided in the S.I.

We briefly introduce a method and five 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. 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, was briefly introduced in Section 1. This metric 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 TP and FP, 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 zip code or town/city has more than *T*% of children with elevated BLLs, and the measured value for that zip code or town/city is indeed over *T*%. FP are those cases in which the model predicts that a zip code or town/city has more than *T*% of children with elevated BLLs, while the actual (measured) value for that zip code or town/city is below *T* %

(3) The “Precision Recall Area Under the Curve” (PR AUC) metric, which ranges from 0.5 to 1, is typically used to evaluate the ability of the model to distinguish between TP 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 0.5 indicate that the model is no better than random selection. FNs are those cases in which the model predicts that a zip code or town/city has less than *T*% of children with elevated BLLs, while the measured value for that zip code or town/city is under *T*%.

(4) The F1 metric is the harmonic mean of Precision and Recall, which are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Thus, the F1 metric accounts for the ability of the model to distinguish between TP and FN for a specific probability threshold (0.5 in our

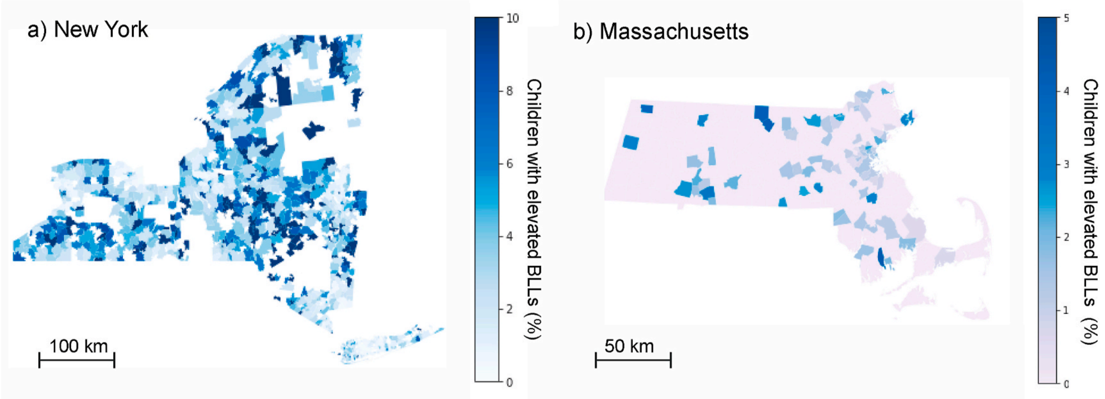
a) New York
100 km
10
8
6
4
2
0
Children with elevated BLLs (%)
b) Massachusetts
50 km
5
4
3
2
1
0
Children with elevated BLLs (%)

Fig. 1. Heat maps of the study sites showing the percentage of children with elevated BLLs (over 5 μg dL^{-1}) within (a) zip codes in New York (n = 1642) and (b) each community (city/town) in Massachusetts (n = 353). There is missing data for some zip codes the in New York dataset; these are shown in white. Note that the color scale in each map is different. New York has a larger percentage of children with elevated BLLs (more than 5 μg dL^{-1}) than Massachusetts. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

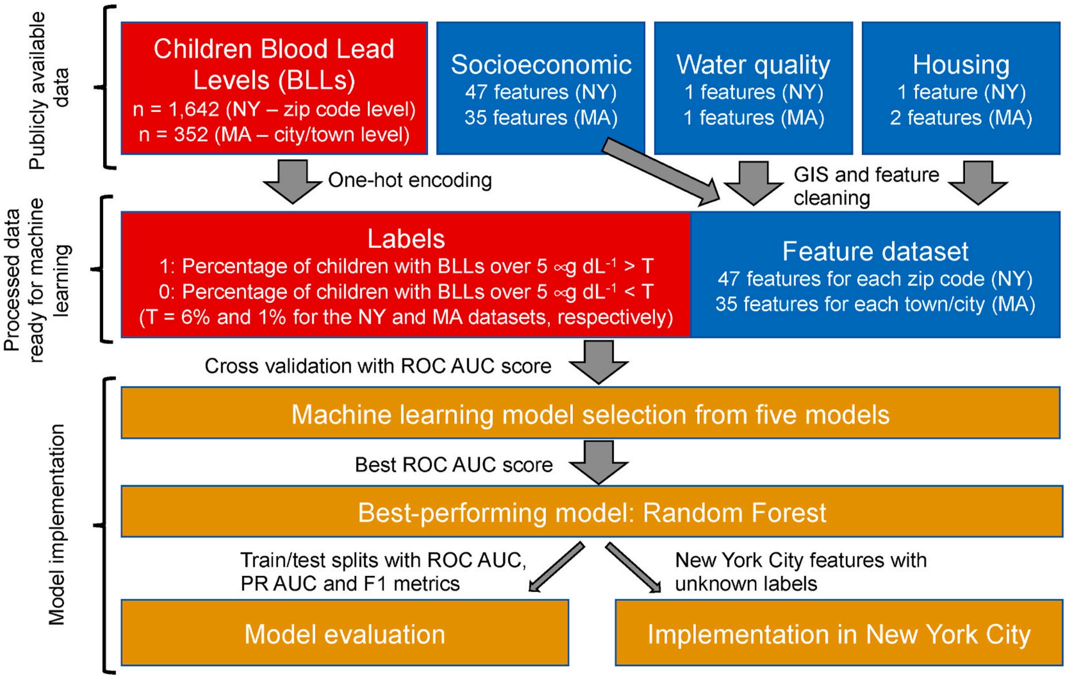
Publicly available data
Children Blood Lead Levels (BLLs)
n = 1,642 (NY – zip code level)
n = 352 (MA – city/town level)
Socioeconomic
47 features (NY)
35 features (MA)
Water quality
1 features (NY)
1 features (MA)
Housing
1 feature (NY)
2 features (MA)
One-hot encoding
GIS and feature cleaning
Processed data ready for machine learning
Labels
1: Percentage of children with BLLs over 5 ∝g dL-1 > T
0: Percentage of children with BLLs over 5 ∝g dL-1 < T
(T = 6% and 1% for the NY and MA datasets, respectively)
Feature dataset
47 features for each zip code (NY)
35 features for each town/city (MA)
Cross validation with ROC AUC score
Model implementation
Machine learning model selection from five models
Best ROC AUC score
Best-performing model: Random Forest
Train/test splits with ROC AUC, PR AUC and F1 metrics
New York City features with unknown labels
Model evaluation
Implementation in New York City

Fig. 2. Schematic overview of the steps taken to process the data and implement the machine learning models. Note that each datapoint corresponds to the percentage of children within each zip code or town/city with Blood Lead Levels (BLLs) over 5 μg dL^{-1} in New York and Massachusetts, respectively, and all its associated features (socioeconomic, chemical, and housing). BLLs are encoded as binary variables, where 1 and 0 represent that a zip code or town/city has over or under T% of children with elevated BLLs, respectively, where T% is 6% for NY and 1% for MA datasets.

case).

The five tested models were compared in terms of their mean ROC AUC score obtained with a 10-fold cross-validation. The best-performing model was tested further 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 1000 times, and the ROC AUC, PR AUC and F1 score were then calculated for the test data in each split. Given the importance of accurately identifying locations where children might have elevated BLLs and the unbalanced nature of our dataset (only about 20% of the datapoints have BLLs above the chosen threshold for each State), the PR AUC and F1 metrics provide insights into the ability of the model to accurately predict the minority class labels and to avoid predicting FNs.

2.5. Model implementation in New York city

The best-performing model was implemented with data from New York City to predict, for each zip code, the likelihood that over 6% of the children have elevated BLLs. These results were compared to measured childhood BLLs in New York City, which were obtained at a borough level for the year 2015 (New York City Department of Health and Mental Hygiene, 2020). To compare these data to our modeled results, we first calculated the modeled expected number of children with elevated BLLs for each zip code using the following equation:

$$E(Ch) = T * P(T) * Pop \tag{3}$$

Where *E(Ch)* is the expected number of children with BLLs exceeding 5 μg dL^{-1} within each zip code, *P(T)* is the modeled probability that over *T*% of the children have elevated BLLs (T = 6% for the New York model) and *Pop* is the population of children under 6 years old residing in each zip code. The values of *E(Ch)* were then aggregated for all zip codes within each borough in New York City and the results compared to the measured data. We note that Eq. (3) would underestimate the number of children with elevated BLLs in zip codes where more than 6% of the children have elevated BLLs. However, this threshold is far from the 1.73% average of children with elevated BLLs in New York City in 2015 (New York City Department of Health and Mental Hygiene, 2020).

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).

3. Results and discussion

3.1. Model selection

The best-performing model for both the New York and Massachusetts datasets was a Random Forest (RF), as shown in Fig. 3 (these results were obtained after model hyperparameter optimization). The hyperparameters used in the model for each dataset are shown in Table 1, while the hyperparameters of the other models are shown in the S.I. The RF model outperformed all other models in terms of the average ROC AUC score, demonstrating that it provides more TP and less FP than the other models. Moreover, this model resulted in less variability among the folds, as shown by the lower standard deviations among the folds during cross validation. Thus, the RF model was selected for further testing with data from the study sites and New York City.

We note that all the tested models provided good results during cross validation. Even the logistic regression model, which has been used unsuccessfully in previous studies for similar purposes (Taylor et al., 2013) resulted in high ROC AUC values (0.84 and 0.87 for the New York and Massachusetts datasets, respectively). This supports our hypothesis that predicting elevated BLLs at a geospatially aggregated level (zip codes or towns/cities) is more feasible than predicting BLLs at an individual person-scale, as other studies have attempted.

3.2. Model performance

The optimized RF model was further tested by randomly splitting the datasets into train and test sets 1000 times and by calculating the ROC AUC, PR AUC and F1 scores. These results are shown in Fig. 4. As a reference, the average F1 score reported in the study by Potash et al. (2015) was 0.40 (we calculated this value using the harmonic mean of their average precision and recall of 0.39 and 0.42, respectively). In contrast, our geospatially aggregated model provided an average F1 score of 0.78 and 0.80 for the New York and Massachusetts datasets, respectively. Another model developed by Potash et al. (2020) resulted in an average ROC AUC score of 0.69, while our model's average ROC AUC was 0.87 and 0.91 for the New York and Massachusetts datasets, respectively. It is likely that the difference in performance between our model and those previously reported in literature results from the lower resolution of our datasets. Most of the models reported in literature use socioeconomic and environmental features to predict elevated BLLs of individuals in specific locations. In contrast, our model predicts elevated BLLs in relatively large populations (on the scale of zip codes and cities/towns) at a state level using aggregated data. The aggregated data helps smoothen the variability among individual children, which allows identifying spatial trends.

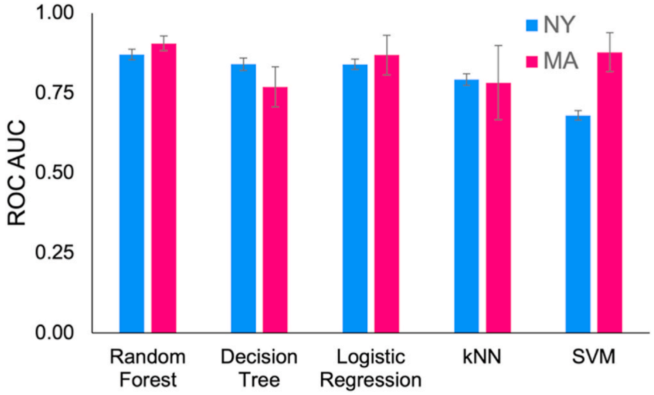

Fig. 3. 10-fold cross validation score of 5 optimized machine learning models using the ROC AUC score for the New York and Massachusetts datasets. The bars represent the mean ROC AUC scores, and the error bars the standard deviation. Random forest outperformed all other models in terms of attaining the highest average ROC AUC score, and demonstrating the smallest value for the standard deviation among the folds.

Table 1
Best combination of hyperparameters for the random forest model implemented using the New York and Massachusetts datasets.

	NY	MA
Number of trees	1000	500
Maximum features	4	6
Maximum depth	12	9

Despite the good average performance of the RF model, significant variations in performance were observed among the 1000 splits for both datasets, as shown in Fig. 2. The variability for all three metrics, F1 score, ROC AUC and PR AUC, was higher when using the Massachusetts dataset. The worst-performing model instance in the Massachusetts dataset had an F1 score of 0.55, a ROC AUC of 0.73 and a PR AUC of 0.62. It is likely that this performance was a product of overfitting, as models trained using small datasets are more likely to overfit the data (Ying, 2019). However, the worst-performing model still has a higher ROC AUC than other models found in literature.

As also shown in Fig. 4, the New York RF model tended to have larger PR AUC than ROC AUC values. This means that this model overestimates the positive label (when over 6% of children within zip codes have elevated BLLs), sacrificing accuracy for precision. This is desirable from a public health perspective, as the cost of predicting FPs are lower than those of predicting FNs (it is more desirable to falsely predict that a zip code has a high risk of childhood lead exposure than to falsely predict that it does not have a high risk of lead exposure). However, the opposite is true for the Massachusetts model: ROC AUC values were greater that PR AUC values. This suggests that the model tends to underestimate the positive label, indicating that it will predict more FNs than FPs. This problem may be addressed by decreasing the probability threshold used to decide the model's outcomes so that more positive predictions (those where the model predicts that over *T*% of the children have elevated BLLs) are made.

3.3. Feature importance analysis using random forest

The features in the New York and Massachusetts RF models were ranked by their Gini score, which represents the loss in entropy (statistical dispersion) resulting from adding each feature to the model (see Fig. 5). We note here that the way RF models work, only the magnitude of reduction in entropy (impact on prediction) is measurable, but not the direction in which the feature impacts the prediction. In both RF models, poverty and race-related features are most important when predicting childhood lead exposure, which aligns with previous studies (Hauptman et al., 2017). In the case of the Massachusetts model, the most important feature corresponds to the percentage of children under 6 living in each zip code or town/city. We have two hypotheses to explain this: 1) It may be the case that not enough BLL tests were performed or that testing was not done randomly, making the results in each zip code or city unrepresentative of the population. This could explain why the model's outcome changes as the number of children tested increases (areas with

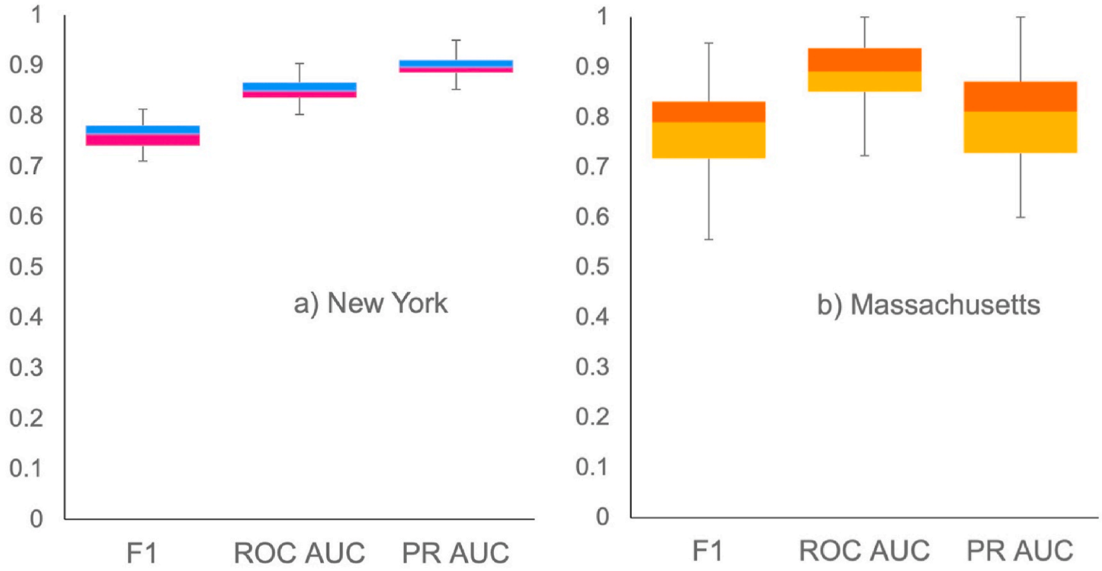

Fig. 4. Box plot of the F1 score and the Area Under the Receiver Operating Characteristic (ROC AUC) and Precision-Recall (PR AUC) curves for 1000 instances of the model using the (a) New York and (b) Massachusetts datasets. The performance of the model is excellent for both the New York and Massachusetts datasets; however, more variability is observed when using the Massachusetts dataset, for reasons discussed in the text.

more children have greater number of tests). 2) Low-income minorities tend to have higher fertility rates (Baughman and Dickert-Conlin, 2009) and this population disproportionally suffers from elevated BLLs (Hauptman et al., 2017). Thus, it may be the case that a town/city with more children is more likely to have more cases of elevated childhood BLLs because more children might indicate less wealth and the presence of minority communities.

Other important features in both random forest models include housing metrics. These results were expected and have been proven to correlate to childhood elevated BLLs in multiple studies. In the case of the Massachusetts RF model, the percentage of pre-1978 housing units is an important feature, which was also expected because old housing is more likely to have lead plumbing and lead paint, two major sources of lead exposure. In the case of the New York model, we did not have access to housing construction year; however, we used Housing Price Index (HPI) as a proxy. As shown in Fig. 5, this index is an important source of information gain in the New York model, indicating that it might be a useful variable for predicting elevated childhood BLLs. Finally, lead levels in schools' drinking water were not an important feature according to the Gini index. This might be because drinking water is not usually the main source of lead exposure in children (Dignam et al., 2019) and because school water quality does not directly impact children under 4 or 5 years of age. These children usually consume water in their homes; however, there are no available public records of lead-levels in household water supplies. It is likely that the housing data used in the model (housing construction date and HPI) indirectly provides information of lead exposure from drinking water at a household level. This is because old buildings are more likely to have lead pipes (Abernethy et al., 2018), which is the main source of lead in drinking water.

3.4. Feature importance analysis using a logistic regression

Even though the Gini index provides information about the importance of each feature, it does not inform on how each of them affect the outcome of the model in terms of magnitude and sign. Given that the logistic regression model provided good results in both datasets during cross validation (see Fig. 3), we ranked the features based on the magnitude of the regression coefficients to gain insights into how each feature affects the outcome of the model but presented our results to also differentiate them by the direction (color-coded for positive or negative) of the influence on the final resulting prediction. These results are shown in Fig. 6.

As seen in the figure, and just like in the random forest model, housing information (HPI and pre-1798 housing) is a very important feature. As expected from prior more narrow studies, HPI is negatively related to childhood elevated BLLs, while pre-1798 housing is positively related to it. Prior literature strongly suggests that older houses, which usually have a lower HPI index, may be sources of lead exposure. In both models, race-related features also behave as expected from prior narrower literature: the percentage of White and Black populations are negatively and positively related to childhood lead exposure, respectively. Furthermore, the prior known fact emerges from both models that low-income communities are disproportionately exposed to toxic levels of environmental lead. Prior researchers have documented in the literature that low-income minority groups tend to have higher rates of childhood lead exposure (Sampson and Winter 2016). Finally, both models agree that lead in school drinking water is less important than other features for children of age 4–5 years (in the New York model this feature is even negatively related to lead exposure). However, the importance of lead in drinking water as a source of childhood lead exposure cannot be disregarded based on these results as we do not have direct information on household water quality.

3.5. Illustrative example of model implementation in New York city

The zip code-level RF model, previously trained with data from the entire State of New York, excluding New York City, was used to predict the probability that over 6% of the children residing in each zip code in New York City have elevated BLLs (BLLs >5 μg dL^{-1}). The resulting probabilities are shown in Fig. 7. As seen in the figure, our model predicted that most zip codes in New York City are not at risk of having over 6% of children with elevated BLLs. This is consistent with the average 1.73% of children 2015 with elevated BLLs in New York City (New York City Department of Health and Mental Hygiene, 2020). However, the map shown in Fig. 7 has two high risk areas (shown in orange) which are located in the Brooklyn and Queens boroughs. These two boroughs account for most of the cases of elevated BLLs in New York City children, as measured by NYC Environment & Health (New York City Department of Health and Mental Hygiene, 2020). Furthermore, the neighborhoods in

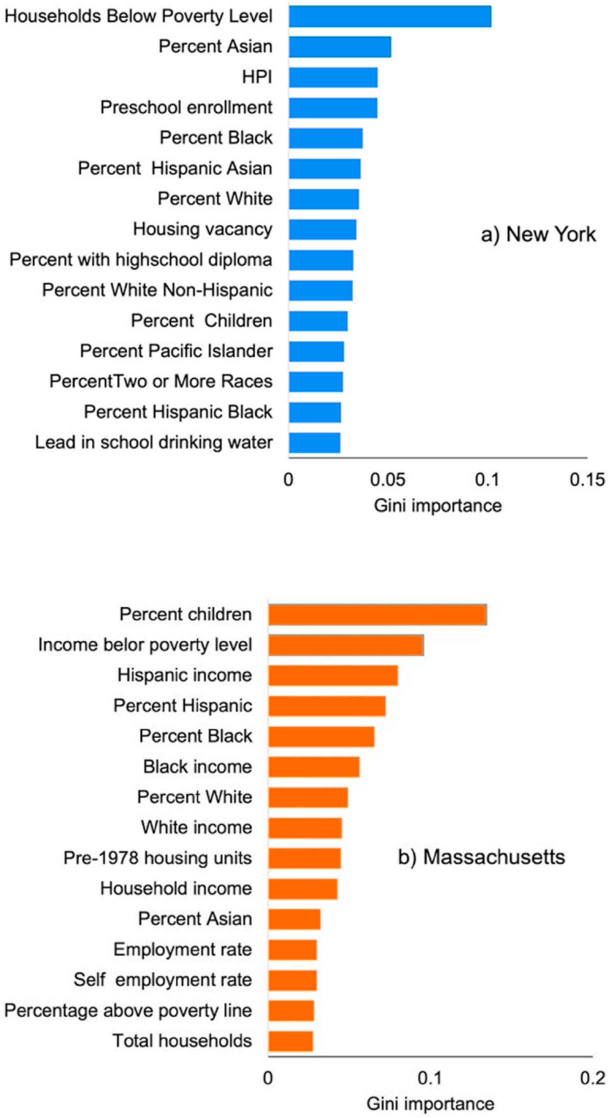

Fig. 5. Top 15 most important features ranked according to the Gini Index when using the optimized random forest model with the a) New York and b) Massachusetts datasets. Race-related features are among the most important predictors of childhood lead exposure.

Brooklyn and Queens with the highest rate of children with elevated BLLs in 2015 were Clinton Hill and Jamaica, respectively (New York City Department of Health and Mental Hygiene, 2020), both of which are contained within the orange areas shown in Fig. 7.

Using Eq. (3) for every zip code, we estimated the expected childhood elevated BLL exposure rates, expressed as cases per 1000 children, for every borough in New York City. These modeled results, as well as the measured childhood elevated BLL exposure rates are shown in Fig. 8. As seen in the figure, the modeled results are within the standard error of every borough. The mean values between measured and modeled data are also similar, except for the Staten Island borough. However, the reported childhood elevated BLL exposure rates in this borough are likely unrepresentative of the population given the small sample size, which explains the large standard error. In fact, the data from two of the sampled neighborhoods in Staten Island have the following warning: “Estimate is based on small numbers so should be interpreted with caution”.

It is worth noting that our model provided accurate BLL estimates for New York City even though this city has taken aggressive measures to combat childhood lead exposure (recall that the model was trained with state data, excluding New York City). In 2015, 2.8% of the children residing in the State of New York, excluding New York City, had BLLs over 5 μg dL^{-1}, one of the largest exposure rates in the US (Centers for Disease Control and Prevention, 2019). In contrast, on the same year, only 1.73% of the children residing in New York City had BLLs over 5 μg dL^{-1} (New York Health, 2021). This is likely a result of wealth distribution: New York City is the wealthiest city in the State of New York (United States Census Bureau, 2011); thus, it has more resources to invest in the removal of potential lead sources. This effect was captured by our model, as wealth is one of the most important features for the model implemented with data from New York State (see Fig. 3).

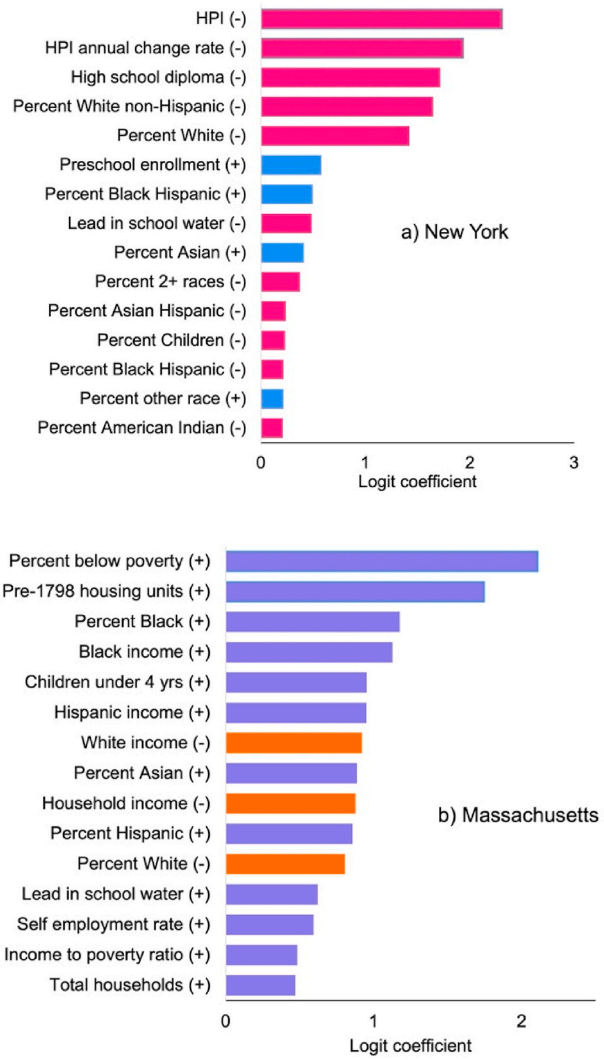

Fig. 6. Top 15 most important features ranked according to the absolute value of the logistic regression coefficients when using the a) New York and b) Massachusetts datasets. The symbols placed in parentheses next to the feature names represent the positive (+) or negative (−) influence of the regression coefficients. As in the random forest model, income and racial features have the largest weights; however, water quality and housing construction date are also important. All features were normalized before the regression so that their magnitudes can be compared.

Of course, the map shown in Fig. 7 does not show any new information given that New York City already tests children for BLLs. However, given that our modeled results closely match those reported in lead blood tests, we expect that this model may be trained with data from other states and then applied to cities within those same states where BLL testing data is lacking.

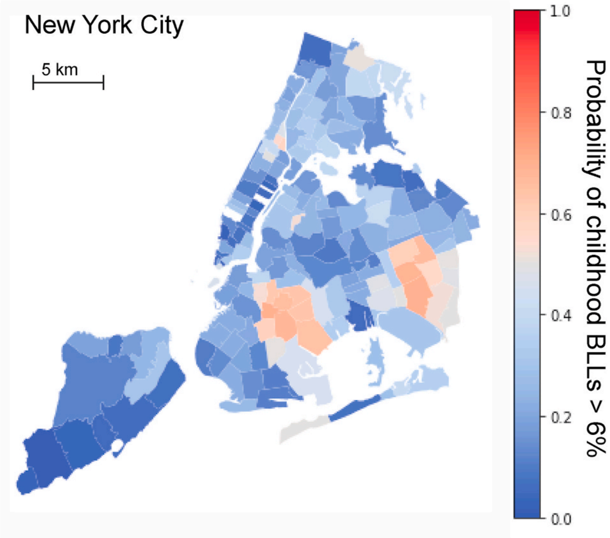

Fig. 7. Predicted probabilities of childhood lead exposure (BLLs over 5 μg dL^{-1}) exceeding 6% in each New York City zip code. The areas with the highest modeled risk are in the Brooklyn and Queens boroughs.

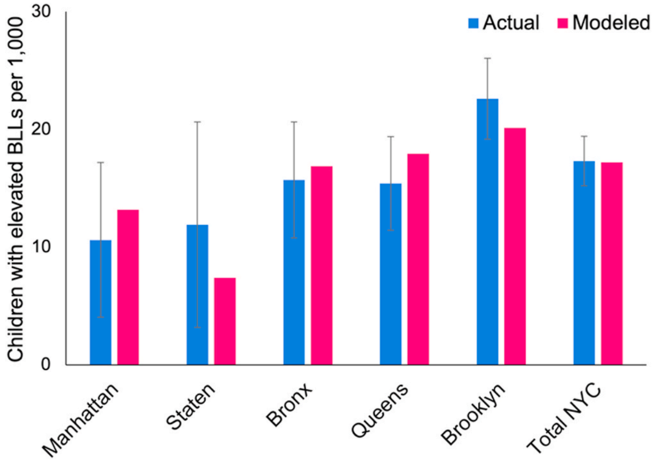

Fig. 8. Modeled and measured childhood elevated BLL exposure rate, expressed as cases per 1000 children, for every borough in New York City in 2015. Standard error bars are shown for the measured data to account for differences in the number of children tested in each borough. The modeled results are within the standard error margins of every borough.

3.6. Model limitations

The RF models presented thus far were implemented by establishing an arbitrary exposure threshold T (1% and 6% for Massachusetts communities and New York zip codes, respectively) and the current CDC reference value of 5 μg dL^{-1}. It is likely that these models may be applied elsewhere and using different values for the threshold T; however, in all cases they must be retrained to account for socioeconomic differences and their effect on lead exposure rates. We hypothesize that other reference values for determining whether children have elevated BLLs or not may also be used; however, this remains untested given that the datasets used only report the number of children with BLLs exceeding 5 μg dL^{-1}. Given that the models rely heavily on socioeconomic features, we do not expect that a model trained with data from one state will be directly applicable to another state or country because different locations have different policies to address childhood lead exposure. However, training the model with partial data, like in the case of the state of New York, may provide useful insights into locations where BLLs have not been measured so that targeted testing and mitigation strategies may be implemented.

Another disadvantage of relying on socioeconomic data is that the model does not directly reflect the mechanisms by which children ingest lead. Ideally, a model to predict childhood lead exposure will rely exclusively on environmental variables, such as lead in air, soil and drinking water. However, those variables are rarely measured, thus, socioeconomic features are needed because, unfortunately, they correlate with lead exposure. We hope that environmental justice efforts will make models like the one presented in this study obsolete (income and race should not be related to BLLs); however, at present they constitute useful tools for predicting, and thus, preventing childhood lead exposure.

While the model accurately predicts the risk of childhood lead exposure for a given area (zip code or community), it does not provide information regarding the variability within each area. This might limit its applicability in large and heterogenous communities where aggregated data fails to accurately describe the population.

Finally, the current version of our model is also limited by the lack of publicly available data. For instance, water utilities in the US test for lead in drinking water in several locations yearly; however, they are only required to report the 90th percentile lead level, per the Lead and Copper Rule. Knowing the locations where high lead water levels were detected would provide invaluable information to models such as the ones presented in this study. Not only would this benefit modeling the impacts of lead in drinking water on childhood lead exposure, but it would also increase the transparency and accountability of water distribution systems. Good quality BLL data is also scarce, as, to the best of our knowledge, only New York and Massachusetts have published data at a high enough resolution to be useful. Other states have published data at a county or state level, which are unsuitable for data science applications or even for people interested in knowing if their children are likely to have elevated BLLs or not. Making high quality BLL data public and easily accessible is not only key for the development of models like the one presented in this study but will also increase transparency and accountability of local authorities.

4. Conclusions

Even though this work focused on developing a machine learning model to predict elevated BLLs at an aggregate level in the states of New York and Massachusetts, we envision future applications in states that do not routinely monitor childhood BLLs. Only 14 states currently recommend universal screening (Michel et al., 2020), thus, this model has the potential to fill in the gaps in the other 36 states that perform partial screening. By testing a fraction of children in these states, this modeling approach may help identify areas at the state level where children are at high risk of having elevated BLLs so that targeted testing and mitigation strategies may be adopted. Moreover, the data collected can be can then be added to the training data, successively improving its accuracy. We believe that modelling approaches using machine learning have the potential to help identify and mitigate childhood lead exposure, a preventable heath crisis that affects the most vulnerable members of our communities.

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 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.ijheh.2021.113862.

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. 5–14 https://doi.org/10.1145/3219819.3219896.

Baughman, R., Dickert-Conlin, S., 2009. The earned income tax credit and fertility. J. Popul. Econ. 22, 537–563. https://doi.org/10.1007/s00148-007-0177-0.

Bierkens, J., Smolders, R., Van Holderbeke, M., Cornelis, C., 2011. Predicting blood lead levels from current and past environmental data in Europe. Sci. Total Environ. 409, 5101–5110. https://doi.org/10.1016/j.scitotenv.2011.08.034.

Billings, S.B., Schnepel, K.T., 2017. The value of a healthy home: lead paint remediation and housing values. J. Publ. Econ. 153, 69–81. https://doi.org/10.1016/j.jpubeco.2017.07.006.

Cattle, J.A., McBratney, A.B., Minasny, B., 2002. Kriging method evaluation for assessing the spatial distribution of urban soil lead contamination. J. Environ. Qual. 31, 1576–1588. https://doi.org/10.2134/jeq2002.1576.

CDC, 2021. Blood lead reference value. accessed 10.4.21. https://www.cdc.gov/nceh/lead/data/blood-lead-reference-value.htm.

Centers for Disease Control and Prevention, 2019. Blood Lead Levels (μg/dL) Among U.S. Children < 72 Months of Age, by State, Year, and Blood Lead Level (BLL) Group Year State.

Chojnacki, A., Dai, C., Farahi, A., Shi, G., Webb, J., Zhang, D.T., Abernethy, J., Schwartz, E., 2017. A data science approach to understanding residential water contamination in flint. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. Part F1296, 1407–1416. https://doi.org/10.1145/3097983.3098078.

Dignam, T., Kaufmann, R.B., LeStourgeon, L., Brown, M.J., 2019. Control of lead sources in the United States, 1970-2017. J. Publ. Health Manag. Pract. 25 https://doi.org/10.1097/phh.0000000000000889. S13–S22.

Doré, E., Deshommes, E., Andrews, R.C., Nour, S., Prévost, M., 2018. Sampling in schools and large institutional buildings: implications for regulations, exposure and management of lead and copper. Water Res. 140, 110–122. https://doi.org/10.1016/j.watres.2018.04.045.

Dos Santos, B.S., Steiner, M.T.A., Fenerich, A.T., Lima, R.H.P., 2019. Data mining and machine learning techniques applied to public health problems: a bibliometric analysis from 2009 to 2018. Comput. Ind. Eng. 138, 106120. https://doi.org/10.1016/j.cie.2019.106120.

Ettinger, A.S., Ruckart, P.Z., Dignam, T., 2019. Lead poisoning prevention: the unfinished agenda. J. Publ. Health Manag. Pract. 25, S1–S2. https://doi.org/10.1097/PHH.0000000000000902.

Gould, E., 2009. Childhood lead poisoning: conservative estimates of the social and economic benefits of lead hazard control. Environ. Health Perspect. 117, 1162–1167. https://doi.org/10.1289/ehp.0800408.

Hauptman, M., Bruccoleri, R., Woolf, A., 2017. An update on childhood lead poisoning. Clin. Pediatr. Emerg. Med. 18, 181–192. https://doi.org/10.1117/12.2549369. Hyperspectral.

Lanphear, B.P., Burgoon, D.A., Rust, S.W., Eberly, S., Galke, W., 1998. Environmental exposures to lead and urban children's blood lead levels. Environ. Res. 76, 120–130. https://doi.org/10.1006/enrs.1997.3801.

Lanphear, B.P., Dietrich, K., Auinger, P., Cox, C., 2000. Cognitive deficits associated with blood lead concentrations < 10 μg/dL in US children and adolescents. Publ. Health Rep. 115, 521–529. https://doi.org/10.1093/phr/115.6.521.

Lanphear, B.P., Hornung, R., Ho, M., Howard, C.R., Eberle, S., Knauf, K., 2002. Environmental lead exposure during early childhood. J. Pediatr. 140, 40–47. https://doi.org/10.1067/mpd.2002.120513.

Lobo, G.P., Kalyan, B., Gadgil, A.J., 2021. Childhood BLL model data and code. GitHub. https://github.com/gadgil-group/childhood_BLL_model.git.

Magavern, S., 2018. Policies to reduce lead exposure: lessons from buffalo and rochester. Int. J. Environ. Res. Publ. Health 15. https://doi.org/10.3390/ijerph15102197.

Marshall, A.T., Betts, S., Kan, E.C., Mcconnell, R., Lanphear, B.P., Sowell, E.R., 2020. Association of lead-exposure risk and family income with childhood brain outcomes. Nat. Med. 26, 91–97. https://doi.org/10.1038/s41591-019-0713-y.Association.

Massachusetts Department of Public Health, 2021a. PHIT data: childhood lead poisoning. accessed 8.7.21. https://www.mass.gov/guides/phit-data-childhood-lead-poisoning.

Massachusetts Department of Public Health, 2021. Learn about lead screening and reporting requirements. accessed 10.4.21. https://www.mass.gov/service-details/learn-about-lead-screening-and-reporting-requirements.

Michel, J.J., Erinoff, E., Tsou, A.Y., 2020. More Guidelines than states: variations in U.S. lead screening and management guidance and impacts on shareable CDS development. BMC Publ. Health 20, 1–10. https://doi.org/10.1186/s12889-020-8225-8.

Mielke, H.W., 1999. Lead in the inner cities. Am. Sci. 87, 62–73. https://doi.org/10.1511/1999.1.62.

Mielke, H.W., Reagan, P.L., 1998. Soil is an important pathway of human lead exposure. Environ. Health Perspect. 106, 217–229. https://doi.org/10.1289/ehp.98106s1217.

Muennig, P., 2009. The social costs of childhood lead exposure in the post-lead regulation era. Arch. Pediatr. Adolesc. Med. 163, 844–849. https://doi.org/10.1001/archpediatrics.2009.128.

New York City Department of Health and Mental Hygiene, 2020. Environment & health data portal. accessed 8.7.21. http://a816-dohbesp.nyc.gov/IndicatorPublic/Subtopic.aspx?theme_code=2,3&subtopic_id=108.

New York Health, 2021. Childhood Blood Lead Level Surveillance Quarters 1-3 2020.

New York State Department of Health, 2021. Childhood lead poisoning prevention. accessed 10.4.21. https://www.health.ny.gov/environmental/lead/.

New York State Department of Health, 2015. Childhood blood lead testing and elevated incidence by zip code: beginning 2000. accessed 8.7.21. https://health.data.ny.gov/Health/Childhood-Blood-Lead-Testing-and-Elevated-Incidenc/d54z-enu8.

Potash, E., Brew, J., Loewi, A., Majumdar, S., Reece, A., Walsh, J., Rozier, E., Jorgensen, E., Mansour, R., Ghani, R., 2015. Predictive modeling for public health: preventing childhood lead poisoning. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. https://doi.org/10.1145/2783258.2788629, 2039–2047.

Potash, E., Ghani, R., Walsh, J., Jorgensen, E., Lohff, C., Prachand, N., Mansour, R., 2020. Validation of a machine learning model to predict childhood lead poisoning. JAMA Netw. open 3, e2012734. https://doi.org/10.1001/jamanetworkopen.2020.12734.

Roy, S., Edwards, M.A., 2019. Preventing another lead (Pb) in drinking water crisis: lessons from the Washington D.C. and Flint MI contamination events. Curr. Opin. Environ. Sci. Heal. 7, 34–44. https://doi.org/10.1016/j.coesh.2018.10.002.

Rushton, G., 2003. Public health, GIS, and spatial analytic tools. Annu. Rev. Publ. Health 24, 43–56. https://doi.org/10.1146/annurev.publhealth.24.012902.140843.

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. https://doi.org/10.1017/S1742058X16000151.

Schultz, B.D., Morara, M., Buxton, B.E., Weintraub, M., 2017. Predicting blood-lead levels among U.S. Children at the census tract level. Environ. Justice 10, 129–136. https://doi.org/10.1089/env.2017.0005.

Schwartz, J., 1994. Low-level lead exposure and Children's IQ: a metaanalysis and search for a threshold. Environ. Res. https://doi.org/10.1006/enrs.1994.1020.

Sethi, J.K., Mittal, M., 2019. A new feature selection method based on machine learning technique for air quality dataset. J. Stat. Manag. Syst. 22, 697–705. https://doi.org/10.1080/09720510.2019.1609726.

Taylor, C.M., Golding, J., Hibbeln, J., Emond, A.M., 2013. Environmental factors predicting blood lead levels in pregnant women in the UK: the ALSPAC study. PLoS One 8, 1–8. https://doi.org/10.1371/journal.pone.0072371.

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. https://doi.org/10.1080/10643389.2011.556556.

Trimble, D., 2016. Measuring the Efficacy of Lead Interventions. Millvale, Pennsylvania.

Tsoi, M.F., Cheung, C.L., Cheung, T.T., Cheung, B.M.Y., 2016. Continual decrease in blood lead level in Americans: United States National health Nutrition and Examination survey 1999-2014. Am. J. Med. 129, 1213–1218. https://doi.org/10.1016/j.amjmed.2016.05.042.

U.S. Department of Health and Human Services, 2012. Health effects of low-level lead. Natl. Toxicol. Progr. Monogr.

United States Census Bureau, 2011. Demographic profile with geos summary file dataset. accessed 8.7.21. https://www.census.gov/programs-surveys/decennial-census/data/datasets.html.

US Census Bureau, 2016. American community survey 2015 data profiles. accessed 8.7.21. https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2018/.

Vivier, P.M., Hauptman, M., Weitzen, S.H., Bell, S., Quilliam, D.N., Logan, J.R., 2011. The important health impact of where a child lives: neighborhood characteristics and the burden of lead poisoning. Matern. Child Health J. 15, 1195–1202. https://doi.org/10.1007/s10995-010-0692-6.

Vorvolakos, T., Arseniou, S., Samakouri, M., 2016. There is no safe threshold for lead exposure: a literature review. Psychiatriki 27, 204–214. https://doi.org/10.1177/1461444810365020.

Whitehead, L.S., Buchanan, S.D., 2019. Childhood lead poisoning: a perpetual environmental justice issue? J. Publ. Health Manag. Pract. 25, S115–S120. https://doi.org/10.1097/PHH.0000000000000891.

Wojtusiak, J., Baranova, A., 2011. Model learning from published aggregated data. Stud. Comput. Intell. 375, 369–384. https://doi.org/10.1007/978-3-642-22913-8_17.

Ye, T., Johnson, R., Fu, S., Copeny, J., Donnelly, B., Freeman, A., Lima, M., Walsh, J., Ghani, R., 2019. Using machine learning to help vulnerable tenants in New York City. COMPASS 2019 - Proc. 2019 Conf. Comput. Sustain. Soc. 248–258. https://doi.org/10.1145/3314344.3332484.

Ying, X., 2019. An overview of overfitting and its solutions. J. Phys. Conf. Ser. 1168 https://doi.org/10.1088/1742-6596/1168/2/022022.

Zahran, S., Mushinski, D., McElmurry, S.P., Keyes, C., 2020. Water lead exposure risk in Flint, Michigan after switchback in water source: implications for lead service line replacement policy. Environ. Res. 181, 108928. https://doi.org/10.1016/j.envres.2019.108928.

Zhang, H., Yin, S., Chen, Y., Shao, S., Wu, J., Fan, M., Chen, F., Gao, C., 2020. Machine learning-based source identification and spatial prediction of heavy metals in soil in a rapid urbanization area, eastern China. J. Clean. Prod. 273, 122858. https://doi.org/10.1016/j.jclepro.2020.122858.