Science of the Total Environment 769 (2021) 145237

Contents lists available at ScienceDirect

Science of the Total Environment

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

Modeling elevated blood lead level risk across the United States

David C. Wheeler [a,*], Joseph Boyle [a], Shyam Raman [b], Erik J. Nelson [c]

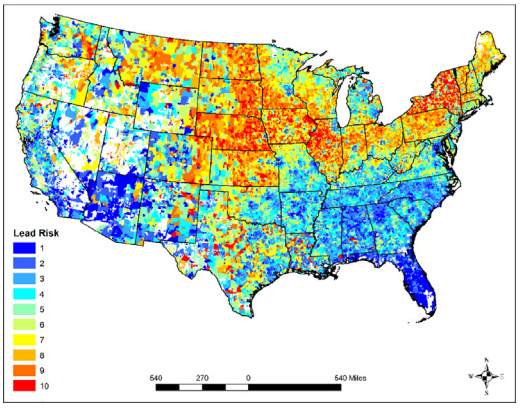
[a] Virginia Commonwealth University, Department of Biostatistics, One Capitol Square, Seventh Floor, 830 East Main Street, Richmond, VA 23219, USA
[b] Cornell University, Department of Policy Analysis & Management, Martha Van Rensselaer Hall, Ithaca, NY 14850, USA
[c] Indiana University, Department of Epidemiology and Biostatistics, 1025 East 7th Street, Bloomington, IN 47405, USA

HIGHLIGHTS

- Lead exposure adversely affects child health and is a major public health concern.
- We combined lead test result data over many states to predict lead exposure risk.
- Lead exposure risk is highest in the Northeast and Midwest US ZIP Codes.
- Percent of houses built before 1940 and median home value are most related to risk.
- The lead exposure risk score can be used for public health intervention efforts.

GRAPHICAL ABSTRACT

Lead risk score predicted for ZIP Codes from a Bayesian hierarchical model based on blood lead testing data and socioeconomic status variables in the United States.



ARTICLE INFO

Article history:
Received 30 September 2020
Received in revised form 17 December 2020
Accepted 13 January 2021
Available online 20 January 2021

Editor: Damia Barcelo

Keywords:
Lead
Socioeconomic status
SES index
Bayesian
Neighborhood deprivation

ABSTRACT

Lead exposure adversely affects child health and continues to be a major public health concern in the United States (US). Lead exposure risk has been linked with older housing and households in poverty, but more studies of neighborhood socioeconomic status (SES) and lead exposure risk over large and diverse geographic areas are needed. In this paper, we combined lead test result data over many states for a majority of the US ZIP Codes in order to estimate its association with many SES variables and predict lead exposure risk in all populated ZIP Codes in the US. The methods used for estimation and prediction of lead risk included the Vox lead exposure risk score, random forest, weighted quantile sum (WQS) regression, and a Bayesian SES index model. The results showed that the Bayesian index model had the best overall performance for modeling elevated blood lead level (EBLL) risk and therefore was used to create a lead exposure risk score for US ZIP Codes. There was a statistically significant association between EBLL risk and the SES index and the most important SES variables for explaining EBLL risk were percentage of houses built before 1940 and median home value. When mapping the lead exposure risk scores, there was a clear pattern of elevated risk in the Northeast and Midwest, but areas in the South and Southwest regions of the US also had high risk. In summary, the Bayesian index model was an effective method for modeling EBLL risk associated with neighborhood deprivation while accounting for additional heterogeneity in risk using lead test result data covering a majority of the US. The resulting lead exposure risk score can be used for targeting public health intervention efforts.

* Corresponding author.
E-mail address: dcwheeler@vcu.edu (D.C. Wheeler).

https://doi.org/10.1016/j.scitotenv.2021.145237
0048-9697/

1. Introduction

Lead (Pb) is a ubiquitous and harmful environmental toxin that causes adverse health effects in children, particularly neurological and neurobehavioral deficits, lower IQ, slowed growth, and anemia (U.S. Department of Health and Human Services, P. H. S, 2007; Canfield et al., 2003; Chiodo et al., 2004; Grandjean and Landrigan, 2014; Lanphear et al., 2000; Lidsky and Schneider, 2003; Miranda et al., 2007; Nelson et al., 2015; Schnaas et al., 2000; Tellez-Rojo et al., 2006; Mielke et al., 1997; Mielke et al., 2017; Mielke et al., 2016). Lead can be ingested from a variety of sources including lead-based paint, household dust containing lead paint, soil, drinking water, and food (Mielke et al., 1997). Although there is no safe blood lead threshold in children, the U.S. Centers for Disease Control and Prevention recommends taking public health actions to reduce future lead exposure for children with blood lead levels (BLLs) at or above 5 μg/dL (Centers for Disease Control and Prevention (CDC), 2012; Wengrovitz et al., 2009). During 2007–2010, the percentage of children aged 1–5 years with BLLs at or above 5 μg/dL was 2.6%, or an estimated 535,000 children in the U.S. with elevated BLLs (EBLLs) (Centers for Disease Control and Prevention (CDC), 2013). Despite efforts by state and local health departments to reduce BLLs in children, the Healthy People 2020 objective to reduce BLLs to an average of 1.6 μg/dL is not likely to be achieved in the near future. (Centers for Disease Control and Prevention (CDC), 2004; US Department of Health and Human Services, 2012). This may be due in part to the difficulty in identifying where to target remediation and prevention efforts because it is not feasible to conduct surveillance that requires obtaining blood from children in a population-based manner.

Prior studies have found evidence that EBLL risk is elevated among persons living in poverty and in older and substandard housing (*A Targeted Approach to Blood Lead Screening in Children, Washington State 2015 Expert Panel Recommendations*, 2016; Jacobs et al., 2002; Raymond et al., 2014). Such housing is often inhabited by racial minorities and socioeconomically disadvantaged persons (Campanella and Mielke, 2008; Leech et al., 2016). Because of these associations, socioeconomic measures of deprivation (e.g., Gini coefficient (Gini, 1997), population below the federal poverty level (U.S. Census Bureau, 2017), or concentrated disadvantage (Sampson et al., 1997)) have been used to estimate risk of EBLLs for a variety of geographic units (e.g., block groups, census tracts, ZIP Codes (Boutwell et al., 2016; Hanna-Attisha et al., 2003; Krieger et al., 2003; Aelion et al., 2013; Carrel et al., 2017)). However, the risk scores and indices that have stemmed from these measures often only use a small subset of area-level covariates, are for only one US state, and also rely upon determinants that have been linked to lead exposure, not actual blood lead test results (*A Targeted Approach to Blood Lead Screening in Children, Washington State 2015 Expert Panel Recommendations*, 2016; Carrel et al., 2017; Jones et al., 2010; Moody and Grady, 2017; Vox Lead Exposure Risk., n.d.; Frostenson and Kliff, 2016). This results in lead exposure risk being portrayed as a function of socioeconomic status (SES) without environmental toxin or serological sampling to substantiate risk. An example of this is the Vox lead exposure risk score (Frostenson and Kliff, 2016), which uses poverty and housing age to construct a risk score for census tracts in the United States (US), but does not use lead test data for the census tracts.

Despite these attempts to identify areas of EBLLs through risk mapping and estimation techniques based on utilizing sociodemographic characteristics and housing age, predicting lead exposure across large areas (i.e., many states) may be inaccurate largely because it has been done without area lead test data. Further complicating the matter, lead presence in topsoil has been shown to be naturally decreasing in some rural areas while remaining vigilant in urban centers. However, geographic lead exposure risk based solely on topsoil has not been directly linked to serologically elevated blood lead levels in children and is also not directly associated with known predictors such as age of the housing stock or income (Mielke et al., 1997; Mielke et al., 2017). Importantly, topsoil lead presence is strongly predicted by minority race (Mielke et al., 2020). Recent studies have further estimated the association of SES variables with EBLL risk (Wheeler et al., 2019a; Wheeler et al., 2019b) in census tracts using lead test result data and weighted quantile sum (WQS) and Bayesian index regression models, but these efforts have been only for single states (e.g., Minnesota, Maryland). Given the potential impact for predicting comprehensive lead exposure risk across small areal units, the objective of this study was to use these methods with positive lead serological test results combined over many states in order to predict lead exposure risk across the entire United States at a granular areal level (e.g., ZIP Codes). In this paper, we compared these methods with the Vox lead exposure risk score and random forest to estimate and predict EBLL risk across ZIP Codes in the US using many SES variables potentially related to lead exposure risk. Ultimately, this study produced a lead exposure risk score for all populated ZIP Codes in the US from a comprehensive set of known lead predictors using the best method evaluated.

2. Material and methods

2.1. Data

2.1.1. EBLL data

The blood lead test data used for the outcome variable were provided by Reuters (Pell and Schneyer, 2017), which initially obtained the data from individual state health departments. The number of tests t_y and elevated lead tests e_y for area y were reported. An elevated test was defined as a reading greater than 5 μg/dL. Certain states suppressed small numbers (<5) of tests performed or elevated lead readings for privacy reasons. For these records, we performed a single imputation of test or elevated reading counts. The imputation was within the range specified by the state for count suppression. For example, if a state suppressed the true number of EBLL counts in an area as <5 but not 0, then the imputation of the area's EBLL count was from the discrete set {1, 2, 3, 4} with equal probability given to each number. The proportion of imputed counts on the ZIP, census tract, and county levels was 41%, 36%, and 43%, respectively. To construct the outcome variable, we aggregated area-level data over as many years as were reported between 2005 and 2015 and calculated the proportion of EBLLs in the area, $p_y = \frac{e_y}{t_y}$. The spatial scale and years of blood lead testing data for each state are listed in the supplemental material (Table S1). States reported lead testing data either on the ZIP Code (23 states), county (3 states), or census tract (6 states) level. The states included and their level of reporting are shown in Fig. 1. Much of the United States was covered at the ZIP Code level, with the exception of a cluster of states in the western region of the country that did not report and a cluster in the Midwest and Northeast that did not report at the ZIP level. Lead test data were present for 16,483 of the 31,643 total ZIP Codes in the United States, representing 52% of US ZIP codes. The goal was to produce lead risk scores for every ZIP Code in the US, so prediction was necessary for the states that did not report blood lead test data at the ZIP Code level. To accomplish this goal, we compared the fit of several models to determine which one best predicted EBLL risk across ZIP Codes, and then used this model with all observed data to predict EBLL risk across the entire country.

2.1.2. Demographic data

The SES variables used to construct the SES indices come were 5-year estimates of area-level variables from the 2007–2011 American Community Survey (ACS), an ongoing survey conducted by the U.S. Census Bureau. These variables describe an area's housing stock and the population's economic status, demographic characteristics, and use of governmental assistance programs (Table 1) and were selected for consideration based on their association with EBLLs in the literature

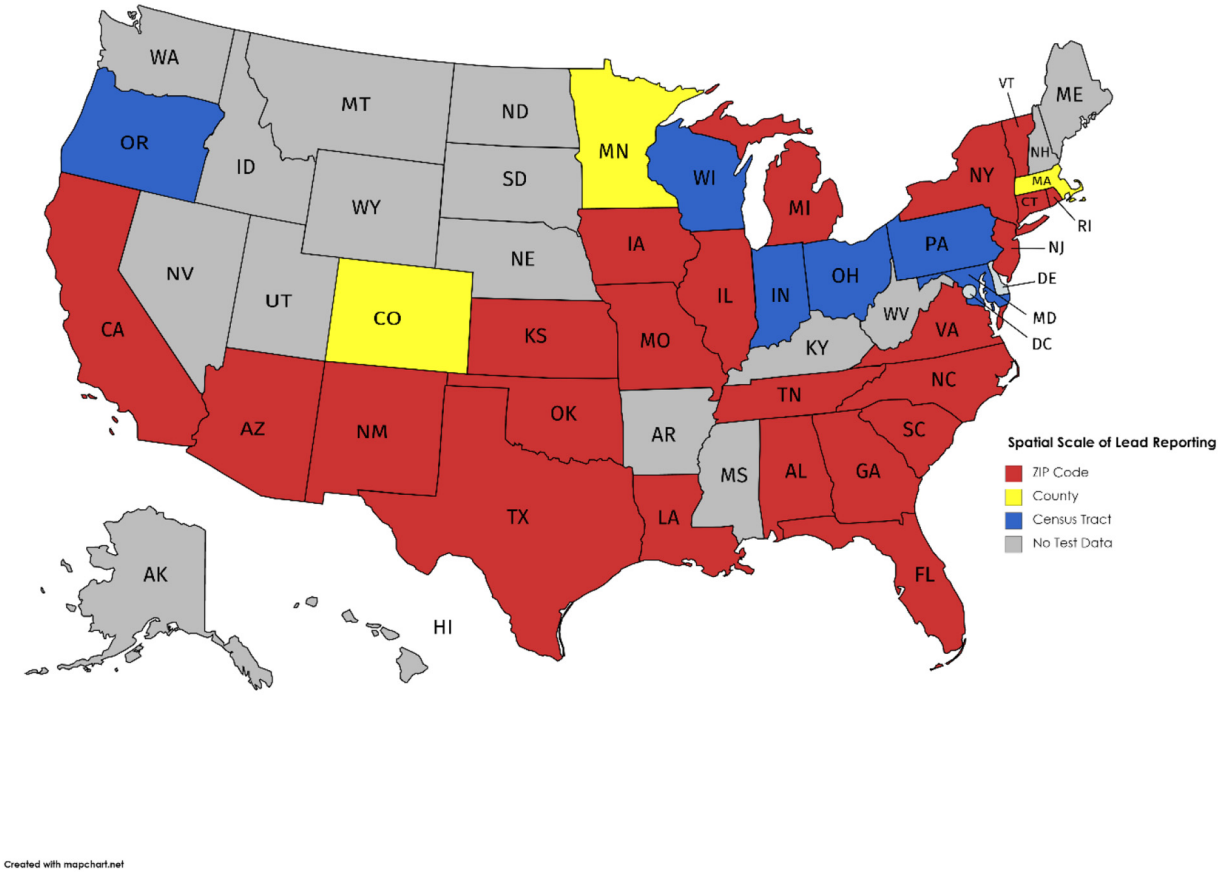


Fig. 1. Coverage of the United States by reported lead test data with shading according to spatial scale of test data.

and use in previous SES indices for EBLL risk (Wheeler et al., 2019a; Wheeler et al., 2019b).

The means and standard deviations of the ACS variables for all the ZIP Codes and the ZIP Codes with blood testing data and without blood testing data show that overall the means of the variables were similar between the ZIP Codes with and without blood testing data (Table 1). Two exceptions are percent Black population and percent structures built before 1940, where the percent of the Black population was 10.2% in the blood testing set and 4.6% in the set without testing data and the percent of structures built before 1940 was 17.4% in the blood testing set and 22.4% in the set without testing data. Noting these descriptive differences, it was important to adjust for these variables in the SES index.

Table 1
American Community Survey variables (2007–2011) considered in socioeconomic status indices with means (standard deviations) by ZIP codes reporting of blood testing data.

Variable	Overall (n = 31,643)	Testing data (n = 16,483)	No testing data (n = 15,160)
Structures built pre-1940 (%)	19.8 (17.7)	17.4 (16.9)	22.4 (18.3)
Structures built 1940–1949 (%)	5.8 (5.6)	5.7 (5.4)	5.9 (5.8)
Structures built 1950–1959 (%)	9.7 (8.1)	10.0 (8.4)	9.3 (7.7)
Structures built 1960–1969 (%)	10.0 (6.9)	10.5 (6.8)	9.4 (6.8)
Structures built 1970–1979 (%)	16.1 (8.9)	15.8 (8.4)	16.5 (9.3)
Structures built 1980–1989 (%)	13.1 (8.8)	13.7 (8.7)	12.5 (8.9)
Structures built 1990–1999 (%)	14.0 (9.7)	14.4 (9.8)	13.5 (9.6)
Structures built 2000–2004 (%)	7.3 (7.3)	7.8 (7.5)	6.8 (7.0)
Structures built post-2004 (%)	4.2 (5.6)	4.7 (6.0)	3.7 (5.0)
Median Household Income (USD)	51,449.1 (22,050.9)	51,642.7 (22,184.4)	51,238.6 (21,903.6)
Median family income (USD)	61,622.7 (25,940.9)	61,745.2 (25,869.8)	61,489.6 (26,018.2)
Gini coefficient for income inequality	0.4 (0.1)	0.4 (0.1)	0.4 (0.1)
Black or African-American residents (%)	7.5 (15.9)	10.2 (17.8)	4.6 (12.9)
Female-headed households (%)	8.7 (7.7)	9.3 (7.5)	8.1 (7.8)
Residents with less than high school degree (%)	15.2 (10.6)	16.2 (10.5)	14.1 (10.5)
Households below federal poverty line (%)	14.3 (10.8)	14.6 (10.3)	14.0 (11.4)
Residents receiving Social Security income benefits (%)	33.2 (12.2)	32.9 (11.6)	33.5 (12.9)
Residents on public Assistance (%)	2.4 (3.3)	2.2 (2.7)	2.6 (3.8)
Residents receiving food stamps/SNAP (%)	10.5 (9.5)	10.8 (8.8)	10.2 (10.2)
Residents using public health insurance (%)	33.0 (12.3)	32.8 (11.5)	33.2 (13.0)
Residents who are unemployed (%)	8.3 (6.4)	8.5 (5.8)	8.1 (7.1)
Vacant structures (%)	17.0 (14.9)	16.5 (13.5)	17.6 (16.2)
Households renting their residence (%)	25.0 (15.6)	25.4 (15.6)	24.6 (15.6)

SNAP = Supplemental Nutrition Assistance Program.

2.2. Statistical modeling

Initial exploration of the lead test dataset on the ZIP Code level (n = 16,483 observations) was done by calculating Spearman correlations between SES variables and the proportion of EBLLs. This was both to assess the strength and direction of their relationship in order to determine the structure of the indices. Median household income and median family income were inverted by subtracting them from their maximum value in the dataset so that all area-level variables would have the same positive direction of association with elevated lead risk. Exploration of the distribution of the proportion of EBLLs revealed that a natural log transformation of the proportion of EBLLs plus a constant of 1 had an approximately normally distributed outcome for modeling. The constant was added before the log transformation due to the presence of zero proportions of EBLLs. This transformed outcome was used in all models.

The following methods were used for modeling proportion of EBLLs: WQS regression, random forests, and Bayesian index models. WQS regression and Bayesian index models were selected due to their previous use in modeling EBLL risk and estimating SES indices (Wheeler et al., 2019a; Wheeler et al., 2019b). Previous research has demonstrated effective performance of WQS regression in modeling sets of correlated variables (Carrico et al., 2015). Random forest was used as a comparison method due to its established performance as a predictive model (Wheeler et al., 2015). When fitting the models, differing sizes of the SES index and different spatial scales of the data (e.g. census tract, county) were considered. The Vox lead exposure risk score was also calculated as a comparison (Vox Lead Exposure Risk., n.d.; Frostenson and Kliff, 2016) (details below).

To evaluate the estimation and prediction performance of the different models and select a final model, the observed ZIP Code data were randomly split into a 70% training set ($n = 11{,}539$) and a 30% prediction set ($n = 4944$). All ZIP-level models were fitted using the 70% training set and then were used to predict in the 30% prediction set. For WQS regression models, the total training set were separated into a 70% set for estimating the SES index weights (training) and 30% for estimating the final model parameters (testing) due to the two-step estimation routine of WQS regression. Datasets for models adding 1) county, 2) census tract, and 3) county and census tract lead test data together were also split according to these rules. Thus, each WQS model had a different number of observations in the training and testing sets. For random forests and the Bayesian index model, no such split of the training set was necessary and all training data were used to fit the models. Counts of observations in each of the data sets are given in Table 2.

2.2.1. Vox lead exposure risk score

In 2016, Vox implemented a lead risk score map for census tracts in the United States based on methodology from the Washington State Department of Health (*A Targeted Approach to Blood Lead Screening in Children, Washington State 2015 Expert Panel Recommendations*, 2016; Vox Lead Exposure Risk., n.d.; Frostenson and Kliff, 2016). They produced a weighted average of two variables for each census tract: the age of the housing stock and the percentage of the population living in poverty. The housing stock variable gave large weight to houses built before 1940, slightly less weight to those built between 1940 and 1959 and increasingly smaller weight to homes built after 1960. The weights for housing age were 0.68 before 1940, 0.43 in 1940–1959, 0.08 in 1960–1979, and 0.03 in 1980–1999 (*A Targeted Approach to Blood Lead Screening in Children, Washington State 2015 Expert Panel Recommendations*, 2016; Vox Lead Exposure Risk., n.d.; Frostenson and Kliff, 2016). These weights represent the percent of housing units with lead hazards by time period and come from a previous study of a nationally representative sample examining lead risk in homes (Wheeler et al., 2019b). The housing age component of the lead exposure risk score is calculated for each area by summing the number of houses built times the weight over the time periods. The poverty variable is defined as the percent of households living in poverty (Vox Lead Exposure Risk., n.d.). The housing age component and poverty component are combined into a weighted sum using weights of 0.58 for age of housing and 0.42 for poverty, and then converted into deciles to represent a lead exposure risk score ranging from 1 to 10. The component weights were derived from differences in mean blood lead levels between children at low and high categories of each variable in a previous study using National Health and Nutrition Examination Survey (NHANES) data (Raymond et al., 2014; Frieden, 2014). Notably, this approach has not been validated on actual lead test data across the country and therefore its ability to predict elevated blood lead levels is unknown. The Vox lead exposure risk scores for census tracts in the United States are publicly available in a web map application (Frostenson and Kliff, 2016). We calculated the Vox score as a validation process and to provide a baseline comparison to our models.

Table 2
Observations counts by model.

Model	Training	Testing	Prediction
WQS: ZIP	8077	3462	4944
WQS: ZIP + County	11,650	4993	4944
WQS: ZIP + Census Tract	12,691	5440	4944
WQS: ZIP + County + Census Tract	12,803	5488	4944
Random Forest: ZIP	11,539		4944
Bayesian Index Model: ZIP	11,539		4944

2.2.2. Weighted quantile sum regression

WQS regression uses weighted quantiles of variables in an index to accommodate different scales of variables, de-correlate the variables, and mitigate some uncertainty in ACS estimation of variables. In this application of WQS, deciles of SES variables and $B = 200$ bootstrap samples of the WQS training data were used to estimate the index weights. Using bootstrap sampling when estimating index weights has been shown to increase sensitivity in detecting important variables in the index (Carrico et al., 2015). In a given bootstrap sample, with $c = 1, \ldots, C$ neighborhood-level variables each separated into $q = 0, 1, \ldots, 9$ quantiles, the WQS model was

$$log(p\_y + 1) = \beta_0 + \beta_1\left(\sum_{c=1}^{C} w_c q_{c,y}\right), \text{where } w_c \geq 0 \forall c \text{ and } \sum_{c=1}^{C} w_c = 1$$

Non-linear optimization was used to estimate the index weights in each bootstrap sample with the solnp library (Ghalanos and Theussl, 2015) in the R computing environment. The final weights were a weighted average of the bootstrap sample estimates, with estimates weighted by the index's test statistic t_b as $\overline{w_c} = \frac{\sum_{b=1}^{B} w_{b,c} t_b}{\sum_{b=1}^{B} t_b}$. The WQS index was constructed using these weight estimates and then its significance was evaluated through the parameter β_1 in the testing set.

To determine whether incorporating lead test data from different spatial scales (i.e., census tract and county) would improve predictive performance, models were fitted incorporating lead test data on 1) only the ZIP level, 2) on the ZIP and county level, 3) on the ZIP and census tract level, and 4) the ZIP, county, and census tract-level with $C = 10$ variables. This initial set of 10 variables consisted of variables having higher Spearman correlations with EBLL proportion in univariate exploratory analyses. Lead test data from states reporting on the ZIP level were common to each of these four models; they differed only in their inclusion of lead test data on the county and/or census tract levels. The prediction set, consisting only of testing data on the ZIP level, was common to all models. To see if a more parsimonious or more complex SES index model would improve prediction of EBLL risk, we also fitted the best spatial scale model with a smaller ($C = 7$) and larger ($C = 15$) index.

2.2.3. Random forest

Random forests were used to model the proportion of elevated tests in a ZIP Code by bootstrapping the data, building regression trees, and aggregating predictions over the trees, each of which was split according to the value of several variables. The number of variables and spatial

scale of data were taken from the most predictive WQS regression model. Each tree in the forest also selected from only a subset of possible variables at each split in order to increase variation between the individual trees. Random forests were fitted over a grid of hyperparameters, including number of variables to try at each split in the tree (2–7), minimal node size (2–7), and number of trees in the forest (300, 400, 500, 600). Random forests were fit in R using the randomForest (Liaw and Wiener, 2002) and ranger (Wright and Ziegler, 2017) packages.

2.2.4. Bayesian index model

A Bayesian index model was used for the mean of the natural log of the proportion of elevated blood lead tests. The transformed outcome followed a normal distribution as $log(p_y + 1) \sim Normal(\mu_y, \tau_p)$ with precision $\tau_p = \frac{1}{\sigma_p^2}$ and $\sigma_p \sim Uniform(0,100)$. The mean was modeled as

$$\mu_y = \beta_0 + \beta_1 \left(\sum_{c=1}^{C} w_c q_{c,y} \right) + u_y,$$

with a neighborhood SES index similar to the WQS models but also incorporating an unstructured random effect u_y. The prior for the unstructured random effect was $u_y \sim Normal(0, \tau_u)$ with precision $\tau_u = \frac{1}{\sigma_u^2}$ and $\sigma_u \sim Uniform(0,10)$. The intercept β_0 had an improper flat prior, and the index coefficient had prior $\beta_1 \sim Normal(0, \tau_1)$ with precision $\tau_1 = \frac{1}{\sigma_1^2}$ and $\sigma_1 \sim Uniform(0,100)$. The weights $w_1, \ldots, w_C$ had a $Dirichlet(\alpha_1, \ldots, \alpha_C)$ prior to ensure that weights would be positive and sum to unity. To predict EBLL proportions for ZIP Codes without reported test data, a sample was drawn from the posterior predictive distribution for the mean, using information from the observed data to estimate model parameters.

Markov Chain Monte Carlo (MCMC) methods were used to estimate model parameters, with one chain consisting of 20,000 iterations and a 10,000-iteration burn-in period, and convergence was assessed using Geweke's criterion. We fit Bayesian models in WinBUGS version 1.4.3 using the R2WinBUGS package (*J. Stat. Softw.*, n.d.) and completed all other analyses in R version 3.6.1.

2.2.5. Model comparison

Model performance was compared using two criteria. The primary evaluation criterion was the correlation between the estimated and observed EBLL proportions in both the estimation (training/testing) set and the prediction set. The secondary evaluation criterion was the median absolute residual (MAR) between estimated and observed EBLL proportions in the prediction set in order to understand the scale of the models' predictive ability. While performance on the prediction set was important, performance in the estimation sets was also considered because the estimation set comprised a majority of the lead test data, and the goal was to accurately reflect EBLL risk across as much of the country as possible. The best model was determined according to the correlation of estimates with observed EBLL proportions and the MAR. This model was then applied to all ZIP Codes in order to provide a lead risk assessment for the entire United States, including those states that did not report lead test data. To communicate risk in a straightforward manner similar to the Vox lead exposure score, deciles of risk based on the estimated/predicted EBLL proportions were used for mapping results.

3. Results

The average proportion of EBLLs in an area was 0.12, with a standard deviation of 0.19, suggesting a large positive skew in the distribution of EBLLs, with a relatively small number of areas having higher proportions of EBLLs. The large positive skew was corrected with the transformation mentioned above. The results from the WQS models comparing different spatial scales of lead test data on the common prediction set are shown in Table 3. Models performed comparably, having similar correlations between estimated and observed EBLL proportions in the prediction set, as well as nearly identical MAR in such comparisons. However, the model with only ZIP-level lead test data had the highest correlation and lowest MAR. Thus, we retained the ZIP-level model for subsequent WQS regression models, as well as random forest and Bayesian index models.

Table 3
Comparing WQS models in the prediction set according to Spearman's correlation and median absolute residual (MAR).

Quantity	Z	Z + C	Z + T	Z + C + T
Correlation	0.3306	0.3306	0.3235	0.3237
MAR	0.0617	0.0625	0.0631	0.0660

Note: Z = ZIP Code, C = County, T = Census Tract.

The results of WQS models with different index sizes modeled only at the ZIP Code level are shown in Table 4. These models performed very similarly, with correlations of approximately 0.33 between estimated and observed EBLL proportions and MARs of 0.06 in the prediction set. The model with C = 7 variables had the highest correlation, suggesting a more parsimonious set of area-level variables was preferable in predicting EBLL risk. Estimated weights for the variables across these three models are listed in Table 4, also showing the importance of structures built pre-1940 and median household value with EBLL risk. These two variables make up most of the weights in the SES index. Notably, almost no weight is given to houses built in 1940–1949, meaning that percent of homes built before 1940 account for the housing age effect. In all WQS models, there was a significant positive association between the SES index and EBLL risk, according to the index coefficient and *p*-values (Table 4).

The random forest and Bayesian index model achieved higher correlations with EBLL proportions than the WQS models on the estimation data and performed comparably on the prediction set (Table 5). The best random forest model randomly sampled two variables at each split, had a minimum size of six terminal nodes, and contained 500 trees in the forest. The Bayesian index model converged according to Geweke's diagnostic. The Bayesian index model had the highest overall correlation with all the data and was subsequently applied to all ZIP Codes in the United States. The variables with the highest weights in the Bayesian index model were structures built pre-1940 (0.893) and median house value (0.105). These variables composed almost the entirety of the index weight. The index effect in this model was highly significant ($\widehat{\beta_1} = 0.022, 95\%$credible interval $(0.020, 0.024)$ and positively related with EBLL risk. Although the Bayesian model performed comparably to other models on the prediction set, it was superior in estimating EBLL proportions in the estimation set (training and testing combined). Thus, it

Table 4
Prediction set performance of ZIP-only WQS models according to Spearman's correlation and median absolute residual (MAR) along with the estimated SES index weights and regression coefficients.

Quantity	C = 7	C = 10	C = 15
Correlation	0.3308	0.3306	0.3303
MAR	0.0616	0.0617	0.0618
Index Coefficient	0.0191	0.0191	0.0196
Index P-value	< 0.001	< 0.001	< 0.001
Structures Pre-1940%	0.79	0.79	0.77
Structures 1940–1949%		0.01	0.01
Median House Value, Inv	0.17	0.16	0.14
Median Household Income, Inv	0.00	0.00	0.00
Gini Coefficient	0.01	0.01	0.00
Black %	0.02	0.01	0.00
Supplemental Security Income %		0.01	0.00
Food Stamps %	0.00	0.00	0.00
Public Insurance %	0.01	0.01	0.00
Unemployed %		0.00	0.00
Female Households %			0.04
< HS Graduate %			0.00
Poverty %			0.00
Vacant Structures %			0.03
Renters %			0.00

Table 5
Performance of models in estimation set, prediction set, and all data combined.

Model	Variables	Correlation: Prediction	Correlation: Test	Correlation: Overall
Vox	2	0.25	0.23	0.24
WQS: Z	10	0.33	0.30	0.32
WQS: Z + C	10	0.33	0.30	0.32
WQS: Z + T	10	0.32	0.25	0.28
WQS: Z + C + T	10	0.32	0.23	0.28
WQS: Z	7	0.33	0.33	0.33
WQS: Z	10	0.33	0.30	0.32
WQS: Z	15	0.33	0.33	0.33
Random Forest	7	0.34	0.95[a]	0.77
Bayesian Index	7	0.33	0.99[a]	0.85

[a] Calculated for training and testing sets combined.

was best able to quantify EBLL risk across ZIP Codes in the estimation and prediction sets. The Vox lead exposure risk score performed the worst in all instances, which shows the limitation of this overly simplistic approach to calculating the lead exposure risk score, as well as the benefit in using lead test result data to fit models. The correlation of the estimates and observed values in all the data combined was much higher with the Bayesian index model (0.85) than with the Vox score (0.24).

After fitting the Bayesian index model to all the observed lead test data and predicting EBLL proportions for those ZIP Codes without reported test data, we mapped the deciles of risk based on the estimated/predicted EBLL proportions. The lead risk scores for ZIP Codes (Fig. 2) across the United States shows a clear spatial pattern with highest risk generally found across the rust belt in the Northeast and Midwest and lowest risk across the South and Southwest. However, there are also ZIP Codes of highest risk in parts of the southwest, including in New Mexico, Arizona, Texas, Utah, Nevada, and California. A closer look at the Mid-Atlantic region (Fig. 3) shows variation in risk both across and within states, with lower risk in Virginia and Maryland compared with West Virginia, Ohio, and Pennsylvania. The high risk in urban areas and lower risk in surrounding suburban areas is evident in Washington, DC and Baltimore, MD. To better visualize lead risk in ZIP Codes, we have created a web map application that allows for interactive exploration of these predicted lead risk scores. This tool is available free online (Boyle, 2020).

4. Discussion

This study investigated the feasibility and efficacy of predictive modeling for lead exposure risk related to neighborhood SES variables across the United States. We combined lead test result data from many states to enable estimating and predicting lead exposure risk for all populated ZIP Codes in the United States to create an easily understandable risk score. We compared the performance of several different methods including the Vox lead exposure risk score, random forest, WQS regression, and Bayesian index models for estimating and predicting elevated blood lead level risk. The results showed that the Bayesian index model performed the best, followed fairly closely by random forest. The Vox lead exposure risk score performed the worst. Among the SES variables considered, we found that percent of homes built before 1940 was the most important, followed by median house value. The other SES variables received little weight in the index model. There were very clear patterns in the lead exposure risk score, with the highest values found in the rust belt of the US and lower values in the south and southwest. There were also clusters of high risk in cities such as Baltimore and Washington, DC with low risk in the surrounding suburban areas.

We approached this study with a major asset of having lead test result data for the majority of ZIP Codes in the US. The processing and combination of test data from many states is a major strength of this study. In fact, this is the first study to use the novel approach of combining lead test result data across many states to estimate EBLL risk. Previous work, such as the Vox lead exposure risk score (Frostenson and Kliff,

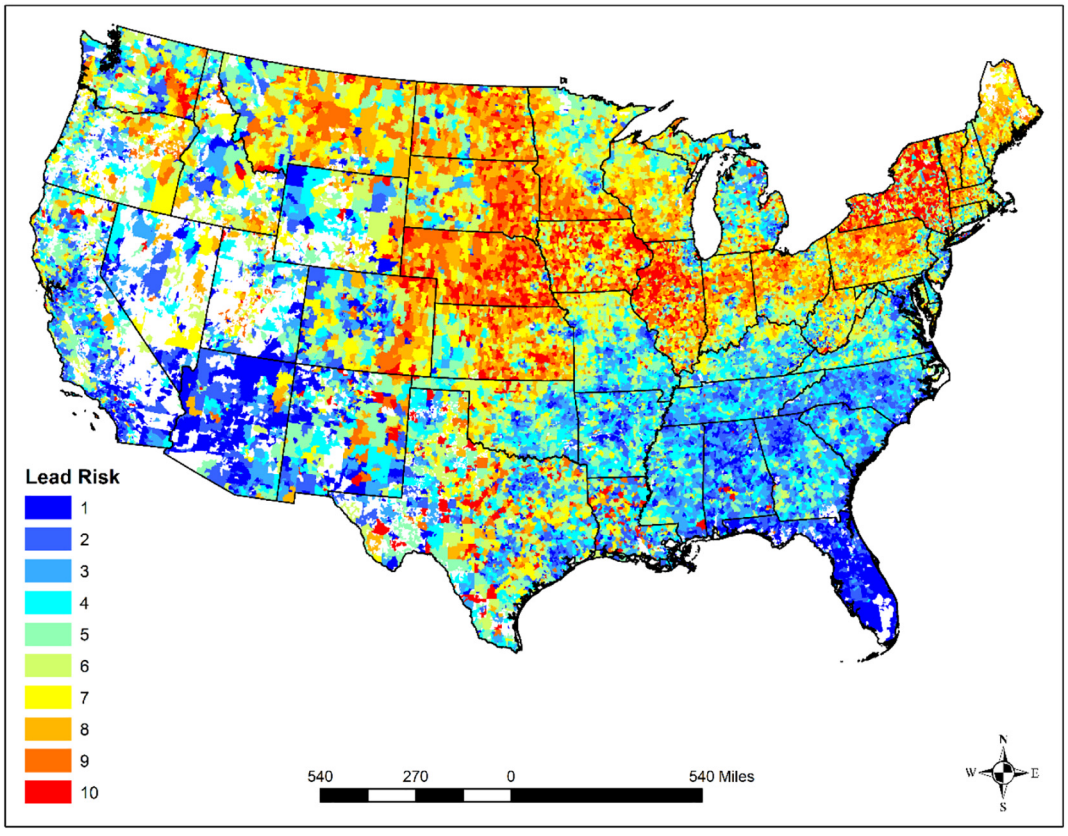


Fig. 2. Lead risk score for ZIP Codes in the United States estimated from the Bayesian index model.

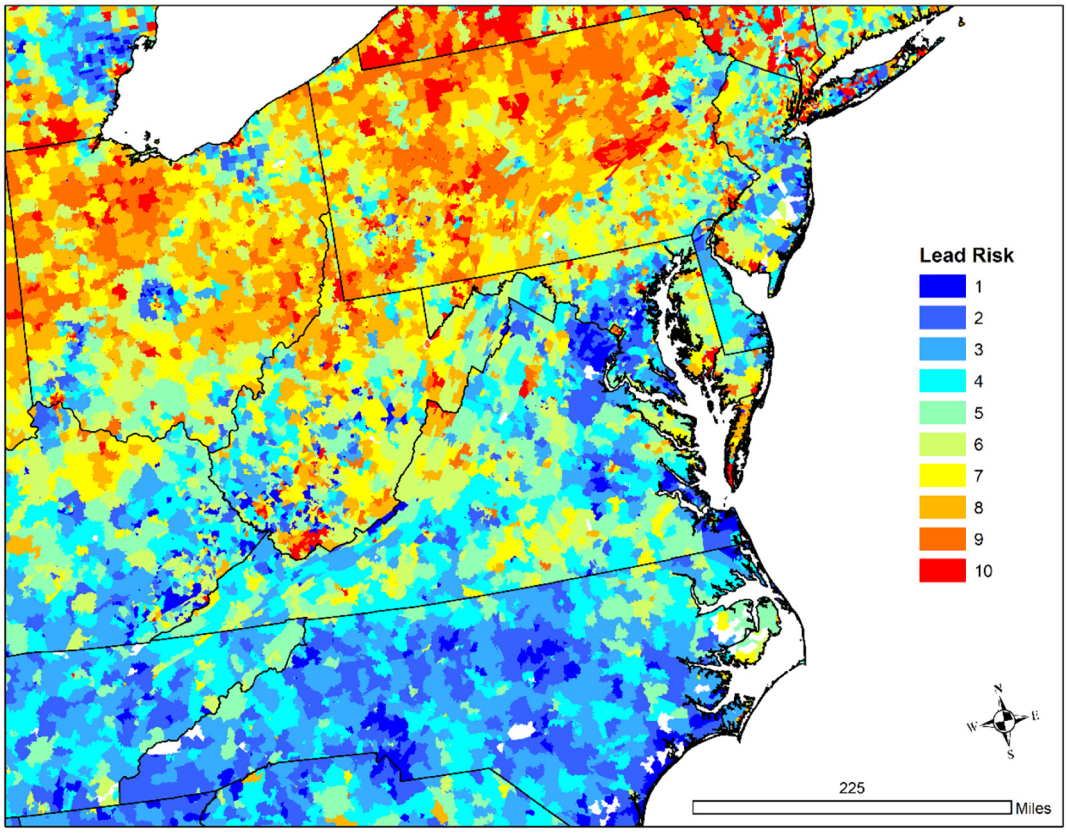


Fig. 3. Lead risk score for ZIP Codes in the United States centered on the Mid-Atlantic region.

2016), has used census data to construct a risk score but did not use geographical-referenced lead test data, nor did it rigorously estimate the weights for the housing age and poverty variables that are the basis for the risk score. Our previous work did use lead test result data to estimate local lead exposure risk explained by SES variables, but these studies were for single states (Minnesota and Maryland (Wheeler et al., 2019a; Wheeler et al., 2019b)) and no prediction of risk was necessary. In the present study, we borrowed information from the areas with observed test data to predict risk in areas without reported lead test data. We also allowed for heterogeneity in risk beyond what the SES index could explain in our Bayesian index model. These model approaches enabled the Bayesian index model to outperform the Vox lead exposure risk score and provide a more effective means of calculating lead risk for ZIP Codes. In addition, we determined which area SES variables are related to lead exposure risk, most notably houses build before 1940 and home value.

Limitations exist for our study, as the data used in our models are subject to inherent bias. The data were collected by state health departments and are generally non-random. Due to this, children seen as having a higher risk for having elevated lead levels (for example, living in housing constructed before 1950 or living in poverty) may have been targeted for testing. In addition, the media outlet Reuters requested lead test result data from all states, but not all states reported data to Reuters and thus we did not have complete coverage of test results for the United States. In addition, not all states reported data at the ZIP Code level. Also, we used an area-level model because individual-level data were unavailable.

5. Conclusions

In conclusion, we have used independent lead test result data across many states and Bayesian index models to estimate lead exposure risk for all populated ZIP Codes in the United States. Our method is an improvement over an existing lead exposure risk score for small areas in the United States. Further, the Bayesian modeling approach is better for identifying ZIP Codes for targeted intervention to reduce lead exposure among children. More efficient allocation of resources for prevention of elevated blood lead cases can be accomplished through advocacy to target geographic clusters of neighborhoods with elevated risk.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2021.145237.

CRediT authorship contribution statement

David C. Wheeler: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing. **Joseph Boyle:** Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Shyam Raman:** Data curation, Investigation, Writing – review & editing. **Erik J. Nelson:** Conceptualization, Data curation, Project administration, Writing – original draft, Writing – review & editing.

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.

References

A Targeted Approach to Blood Lead Screening in Children, Washington State 2015 Expert Panel Recommendations. Washington State Department of Health; DOH, pp. 334–383.

Aelion, C., Davis, H., Lawson, A., Cai, B., McDermott, S., 2013. Associations between soil lead concentrations and populations by race/ethnicity and income-to-poverty ratio in urban and rural areas. Environ. Geochem. Health 35 (1), 1–12.

Boutwell, B., Nelson, E., Emo, B., 2016. The intersection of aggregate-level lead exposure and crime. Environ. Res. 148, 79–85.

Boyle, J., 2020. Lead in the USA. Mapbox Digital Map Accessed December 17, 2020. Available at:. https://api.mapbox.com/styles/v1/boylejr/ckfvgrprv41i319nymtaem8oo.html?fresh=true&title=view&access_token=pk.eyJ1IjoiYm95bGVqciIsImEiOiJja2NwYW11d2swYWU4MzNvY2dnbGp0ajBnIn0.yfF0WqHPwXxdWCs68rsBdQ#4.07/38.31/-95.7.
Campanella, R., Mielke, H., 2008. Human geography of New Orleans' high-lead geochemical setting. Environ. Geochem. Health 30 (6), 531–540.
Canfield, R., Henderson Jr., C., Cory-Slechta, D., Cox, C., Jusko, T., Lanphear, B., 2003. Intellectual impairment in children with blood lead concentrations below 10 microg per deciliter. N. Engl. J. Med. 348 (16), 1517–1526.
Carrel, M., Zahrieh, D., Young, S., 2017. High prevalence of elevated blood lead levels in both rural and urban Iowa newborns: spatial patterns and area-level covariates. PLoS One 12 (5) (e0177930).
Carrico, C., Gennings, C., Wheeler, D., Factor-Litvak, P., 2015. Characterization of weighted quantile sum regression for highly correlated data in a Risk analysis setting. J. Agric. Biol. Environ. Stat. 20 (1), 100–120.
Centers for Disease Control and Prevention (CDC), 2004. Preventing lead exposure in young children: a housing-based approach to primary prevention of lead poisoning. Atlanta, GA. https://www.cdc.gov/nceh/lead/publications/primarypreventiondocument.pdf.
Centers for Disease Control and Prevention (CDC), 2012. Response to the advisory committee on childhood Lead poisoning Prevention report, low level lead exposure harms children: a renewed call for primary prevention. MMWR Morb. Mortal. Wkly Rep. 61 (20), 383.
Centers for Disease Control and Prevention (CDC), 2013. Blood lead levels in children aged 1-5 years—United States, 1999-2010. MMWR Morb. Mortal. Wkly Rep. 62 (13), 245–248.
Chiodo, L., Jacobson, S., Jacobson, J., 2004. Neurodevelopmental effects of postnatal lead exposure at very low levels. Neurotoxicol. Teratol. 26 (3), 359–371.
Frieden, T., 2014. Use of selected clinical preventive services to improve the health of infants, children, and adolescents—United States, 1999–2011. MMWR Supplements 63 (2), 1–2.
Frostenson, S., Kliff, S., 2016. Where Is the Lead Exposure Risk in your Community? Vox https://www.vox.com/a/lead-exposure-risk-map
Ghalanos, A., Theussl, S., 2015. Rsolnp: general non-linear optimization using augmented Lagrange multiplier method. R package version 1, 16.
Gini, C., 1997. Concentration and dependency ratios [1909, in Italian]. Rivista Di Politica Economica 769–789.
Grandjean, P., Landrigan, P., 2014. Neurobehavioural effects of developmental toxicity. Lancet Neurol. 13 (3), 330–338.
Hanna-Attisha, M., LaChance, J., Sadler, R., Champney Schnepp, A., 2003. 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 (2), 186–199.
R2WinBUGS: a package for running WinBUGS from R. J. Stat. Softw., 12(3), 1–16.
Jacobs, D., Clickner, R., Zhou, J., Viet, S., Marker, D., Rogers, J., Zeldin, D., Broene, P., Friedman, W., 2002. The prevalence of lead-based paint hazards in U.S. housing. Environ. Health Perspect. 110 (10).
Jones, E., Wright, J., Rice, G., 2010. Metal exposures in an inner-city neonatal population. Environ. Int. 36 (7), 649–654.
Krieger, N., Chen, J., Waterman, P., Soobader, M., Subramanian, S., Carson, R., 2003. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: the public health disparities geocoding project (US). J. Epidemiol. Community Health 57 (3), 186–199.
Lanphear, B., Dietrich, K., Auinger, P., Cox, C., 2000. Cognitive deficits associated with blood lead concentrations <10 microg/dL in US children and adolescents. Public Health Rep. 115 (6), 521–529.
Leech, T., Adams, E., Weathers, T., Staten, L., Filippelli, G., 2016. Inequitable chronic lead exposure: a dual legacy of social and environmental injustice. Fam Community Health 39 (3) (151–159.24).
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.
Lidsky, T., Schneider, J., 2003. Lead neurotoxicity in children: basic mechanisms and clinical correlates. Brain 126 (Pt 1), 5–19.
Mielke, H., Dugas, D., Mielke Jr., P., Smith, K., Gonzales, C., 1997. Associations between soil lead and childhood blood lead in urban New Orleans and rural Lafourche parish of Louisiana. Environ. Health Perspect. 105 (9), 950–954.
Mielke, H.W., Gonzales, C.R., Powell, E.T., Mielke Jr., P.W., 2016. Spatiotemporal dynamic transformations of soil lead and children's blood lead ten years after hurricane Katrina: new grounds for primary prevention. Environ. Int. 94, 567–575.
Mielke, H., Gonzales, C., Powell, E., 2017. Soil Lead and Children's blood Lead disparities in pre- and post-hurricane Katrina new Orleans (USA). Int. J. Environ. Res. Public Health 14 (4).
Mielke, H.W., Gonzales, C.R., Powell, E.T., Shah, A., Berry, K.J., Richter, D.D., 2020. Spatial-temporal association of soil Pb and children's blood Pb in the Detroit Tri-County area of Michigan (USA). Environ. Res. 191, 110112.
Miranda, M., Kim, D., Galeano, M., Paul, C., Hull, A., Morgan, S., 2007. The relationship between early childhood blood lead levels and performance on end-of-grade tests. Environ. Health Perspect. 115 (8), 1242–1247.
Moody, H., Grady, S., 2017. Lead emissions and population vulnerability in the Detroit (Michigan, USA) metropolitan area, 2006–2013: a spatial and temporal analysis. Int. J. Environ. Res. Public Health 14 (12).
Nelson, E., Shacham, E., Boutwell, B., 2015. Childhood lead exposure and sexually transmitted infections: new evidence. Environ. Res. 143 (Pt A), 131–137.
Pell, M., Schneyer, J., 2017. Reuters finds 3,810 U.S. Areas with Lead Poisoning Double Flint's. Reuters, p. 2017 November 14. https://www.reuters.com/article/us-usa-lead-map/reuters-finds-3810-u-s-areas-with-lead-poisoning-double-flints-idUSKBN1DE1H2.
Raymond, J., Wheeler, W., Brown, M., 2014. Lead screening and prevalence of blood lead levels in children aged 1–2 years—child blood lead surveillance system, United States, 2002–2010 and national health and nutrition examination survey, United States, 1999–2010. MMWR Supplements 63 (2), 36–42.
Sampson, R., Raudenbush, S., Earls, F., 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277 (05), 918–924.
Schnaas, L., Rothenberg, S., Perroni, E., Martinez, S., Hernandez, C., Hernandez, R., 2000. Temporal pattern in the effect of postnatal blood lead level on intellectual development of young children. Neurotoxicol. Teratol. 22 (6), 805–810.
Tellez-Rojo, M., Bellinger, D., Arroyo-Quiroz, C., 2006. Longitudinal associations between blood lead concentrations lower than 10 microg/dL and neurobehavioral development in environmentally exposed children in Mexico City. Pediatrics 118 (2), e323–e330.
U.S. Census Bureau, 2017. American Community Survey. https://www.census.gov/programs-surveys/acs/.26.
U.S. Department of Health and Human Services, P. H. S, 2007. Toxicological Profile for Lead. Agency for Toxic Substances and Disease Registry (ATSDR) https://www.atsdr.cdc.gov/toxprofiles/TP.asp?id=96&tid=22#bookmark16.
US Department of Health and Human Services, 2012. Healthy people 2020: Topics and objectives index. Washington, DC. http://www.healthypeople.gov/2020/topicsobjectives2020.
Vox Lead Exposure Risk. (n.d.). Retrieved September 1, 2020, from https://github.com/voxmedia/data-projects/tree/master/vox-lead-exposure-risk
Wengrovitz, A., Brown, M., Advisory Committee on Childhood Lead Poisoning DoE, Emergency Health Services NCfEH, Centers for Disease C, Prevention, 2009. Recommendations for blood lead screening of Medicaid-eligible children aged 1–5 years: an updated approach to targeting a group at high risk. MMWR Recomm. Rep. 58 (RR-9), 1–11.
Wheeler, D., Nolan, B., DellaValle, C., Ward, M., 2015. Modeling groundwater nitrate concentrations in private wells in Iowa. Sci. Total Environ. 536, 481–488.
Wheeler, D., Raman, S., Jones, R., Schootman, M., Nelson, E., 2019a. Bayesian deprivation index models for explaining variation in elevated blood lead levels among children in Maryland. Spat. Spatiotemporal. Epidemiol. 30, 100286.
Wheeler, D., Jones, R., Schootman, M., Nelson, E., 2019b. Explaining variation in elevated blood lead levels among children in Minnesota using neighborhood socioeconomic variables. Sci. Total Environ. 650 (Pt 1), 970–977.
Wright, M., Ziegler, A., 2017. Ranger: a fast implementation of random forests for high-dimensional data in C++ and R. J. Stat. Softw. 77 (1), 1–1744 Sturtz, S., Ligges, U., & Gelman, A. (n.d.).