Get ahead of the Lead: How Statistical Modeling Can Assist in Identifying Areas of Lead Exposure Risk

Johanna B. Laraway

ABSTRACT

Lead exposure is a threat to all manner of vulnerable population groups across the globe, as it continues to act as both a public health hazard and an environmental justice concern. In low and middle-income countries, poor water lead level (WLL) monitoring infrastructure yields social circumstances wherein at-risk groups and areas of elevated lead content are largely undefined, thereby augmenting the likelihood of chronic lead ingestion and poisoning. In order to mitigate such consequences, more information is necessary regarding the spatial variation in magnitude of lead leaching. This paper presents a statistical approach to model and predict water lead levels (WLL) in any given area using chemical water quality characteristics as well as socioeconomic information. To assess the validity of this proposed methodology, I first selected 1,831 WLL samples taken from public schools within 21 California public water supply systems. I then gathered the necessary indicator variable data for each of these water systems, which included poverty rate and several chemical attributes, and performed multiple regression analysis. This ultimately revealed a statistically insignificant relationship, where only 8% of WLL variance could be explained by the predictor variables. Therefore, at present, this modeling strategy does not represent a viable means of predicting WLL where lead monitoring programs are insufficient. Even so, these results provide important insights into how the model creation process might be altered in the future to generate more accurate predictions and assist in much needed lead exposure hotspot identification.

KEYWORDS

Water lead levels (WLL), disproportionate exposure, model scalability, Multivariable Linear Regression

INTRODUCTION

Despite an outpouring of legislation from the 1970s to 1990s against the use of lead pipes in municipal and residential plumbing (Dignam et al. 2019), recent catastrophes like the Flint Water Crisis have unveiled widespread failure to comply with these regulations. The American Water Works Association estimates that 6.1 million lead service lines (LSLs) currently exist as part of the United States' 160,000 public water systems (NRDC 2016; NRC 2006). Unfortunately, due to the mismanagement of such aging infrastructure, over 5,300 of these systems are in violation of the Environmental Protection Agency's (EPA) Lead and Copper Rule (LCR) (NRDC 2016), a law which stipulates that water systems must take corrective action once water lead levels (WLL) exceed 15 parts per billion (ppb), despite a guideline value of 10 ppb prescribed by the World Health Organization (WHO) (WHO 2018). As a result of this unaddressed contamination, over 18 million people in the United States alone are exposed to WLL above 15 ppb (Olson & Fedenick 2016), thereby posing a significant threat to public health. Though lead poisoning symptoms can be difficult to detect, no blood lead level (BLL) above zero is free of all risk, and exposure can cause neurological and reproductive problems at any age (CDC 2012). Not only can lead's toxicity contribute to fetal deformations, it can hinder brain development, thereby resulting in a loss of IQ (Udedi, 2003; Nevin, 1999), and can even cause severe and permanent brain damage (Duruibe et al., 2007). Furthermore, a multitude of studies have identified a significant link between heightened blood lead levels and the prevalence of crime, both violent and non-violent (Nevin 1999, 2007; Reyes 2007; Stretesky & Lynch 2004). But the issue of persistent lead exposure is a threat to much more than public health alone. A small but growing body of work has considered the relationship between neighborhood characteristics, such as poverty rate or racial composition, and individuals' blood lead level (BLL) (Lanphear et al., 1998), thus revealing a social inequity component to lead exposure risk.

Within the U.S., neighborhood prevalence rates of elevated BLL remain closely linked to socioeconomic, racial, and ethnic segregation (Sampson and Winter 2016), a relationship that may have to do with the allocation of monetary obligation associated with LSL replacement. To clarify, if municipalities detect harmful lead concentrations and take action, local governments are only financially responsible for the portion of the LSL that is under the city street, in other words, the "public" side of the property line, leaving the "private" costs to property owners (LaFrance 2016).

Consequently, even when cities do comply and replace municipal water systems and public service lines, private homeowners must carry the financial burden of replacing lead pipes connected to and within homes (LaFrance 2016). However, for a typical home in the United States, water main replacement can cost between $5,000 and $10,000, not including the external fees associated with repair (LaFrance 2016). For many residents, especially in mid to lower-income communities, this high cost of home-scale pipe replacement makes choosing safe drinking water a difficult and often fiscally impractical choice. Because of this, substantial numbers of low-income civilians and communities of color continue to exhibit unacceptably high blood-lead levels (Meyer et al. 2008 ; Sampson & Winter 2016), further reinforcing issues of environmental injustice, where racial and class-based segregation is directly linked to both environmental hazards and poor health outcomes (Crowder & Downey 2010 ; Downey 2006 ; Sampson & Winter 2016 ; Williams & Collins 2001). Beyond U.S. borders, while historical dissimilarities across countries have resulted in inconsistent correlations between race and BLL, the disproportionate lead exposure based on class present in the United States exists in an analogous way at a global scale.

Low and middle-income countries and communities have been grappling with lead contamination for years (Kordas et al. 2018). In 2000, the World Health Organization published a study estimating that 120 million people had a BLL between 5 and 10 micrograms per deciliter (µg/dL), the majority of which were children (Prus-Ustun et al. 2004); of these children, over 90% lived in low and middle-income countries (Kordas et al. 2018). Yet these countries still struggle to garner recognition, receive restitutions, or design effective prevention efforts (Kordas et al. 2018). This is because the issue is not only one of aging infrastructure but of nonexistent or inadequate lead monitoring programs that leave target groups and sources of exposure undefined (Romeiu et al. 1997; Ngueta et al. 2013). In such countries, very little is known regarding the proportion of homes containing LSLs (Ngueta et al. 2013; Kordas et al. 2018) let alone whether those LSLs are leaching toxic quantities of lead into the drinking water supply. Without the presence or proper enforcement of such programs, it is virtually impossible to understand the true social cost of lead ingestion, which provides little to no incentive for decision-making entities to take remedial action and leaves large swaths of the global population subject to the negative health effects of elevated WLL.

Going forward, in order to alleviate the social inequality and health risks that accompany disproportionate lead exposure, more information is necessary regarding the severity of lead leaching in places where monitoring programs are lacking. With access to this information, governments would be able to quantify and better understand the full extent of the issue, which in turn could instigate much need infrastructure replacements and reduce the likelihood of lead exposure. This study attempted to address this data gap, by applying methods of statistical modeling to publicly available WLL sampling data specific to California and considering both water quality and social values as predictors of WLL. Ultimately, the intention was to quantify the indicator variables' influence on WLL and assess statistical significance, in order to construct an accurate statistical model of lead concentration in drinking water systems, thereby allowing for the identification of high-risk populations at a level of precision useful for jurisdictions lacking mandated lead monitoring programs.

BACKGROUND

Water chemistry

Elevated WLL occurs when lead from LSLs dissolves into the water supply; the magnitude of such leaching depends on several factors. While an absence of corrosion inhibitors or the use of a moderately oxidizing disinfectant residual can impact WLL (WHO 2011), in this study, I focused solely on the variations in the chemical composition of treated water that alter its corrosivity (WHO 2011). For instance, pH, alkalinity, tendency to form a scale on the interior of the pipe, and the comparative presence of chloride to sulfate ions can significantly influence the quantity of lead entering the water supply, with soft, acidic water considered the most plumbosolvent (Schock 1990). The presence of a protective scale, or coating, on the interior of LSLs is crucial, as it lowers the amount of lead leached into the water running through a pipe (Kim & Herrera 2010). So, properties such as total dissolved solids (TDS) and water hardness, which greatly affect scale-forming tendencies, are especially important. Hard water contains higher quantities of calcium and magnesium salts, and is more conducive to scale formation, while soft water is quite corrosive (Rawson & Ayala 2005). Likewise, TDS, which refers to any minerals, salts, metals, cations or anions dissolved in water, tends to contribute to corrosion issues through its effect on scaling (WHO 2011). The relative acidity of the water flowing through a pipe represents an additional influence on WLL, where a more basic pH is desired in order to limit corrosion. For example, if a water system wanted to reduce the level of lead in drinking water, they might consider increasing the pH in the distribution system from <7 to 8-9 (Sherlock 1984). Similarly, the lower the alkalinity, the more likely the water is to be corrosive (Tam & Elefsiniotis 2009). Finally, a chloride-to-sulfate mass ratio (CSMR) greater than 0.5 occurring in drinking water facilities is considered to have the potential to promote galvanic corrosion of leaded connections in the distribution system, and cause hazardous WLLs (Nguyen et al., 2011).

METHODS

Datasets

With the passage of Assembly Bill 746 and the subsequent addition of Section 11627 to the California Health and Safety Code, which went into effect on January 1st, 2018, community water systems have been required to test and record drinking water lead levels at all California public K-12 schools constructed before January 1st of 2010 (A.B. 746, 2017). All sample results are compiled in an excel spreadsheet that can be accessed through the California Waterboards website, and is updated frequently by the Division of Drinking Water in collaboration with the California Department of Education. There are typically several sample sites throughout each school; consequently, most schools have more than one measurement defining the concentration of lead in their drinking water. The sampling results spreadsheet I used had last been updated on October 25th, 2019, and for the purposes of this study, I assumed that these school data accurately reflected the public's general experience with WLL. This dataset contained 40,409 lead level observations for 2,870 public schools in California, with WLL ranging from less than 5 parts per billion to 185 parts per billion and dating from February of 2017 to October of 2019 (see Appendix A1). For each sample taken, community water systems additionally provided the name, address, and district of the school in question, as well as the sampling date and site within the school. If WLL was greater than 15 ppb, the water system responsible for testing recorded this action level exceedance, along with any following corrective efforts, if applicable. While the sampling data was spread out over 146 total California public water supply systems, due to time constraints, I was only able to include twenty-one in this study (see Appendix A2)

To provide a socioeconomic indicator variable, I downloaded poverty rate data for

California census tracts, recorded for the 2000 United States Decennial Census and provided by the U.S. Census Bureau in the form of an excel spreadsheet. Along with poverty rate, the spreadsheet noted the total number of individuals living within a census tract, as well as the total number of individuals living below the poverty level. Within the United States and abroad, poverty thresholds are typically set to reflect the minimum amount of income necessary to cover basic needs, such as food and water, shelter, and clothing (IRP 2016). While nation-specific poverty lines vary widely across the globe, poverty rate, which denotes the percentage of a given population that falls below a respective country's poverty line, allows for a more relative metric of comparison between countries.

Table 1. Summary of Water Quality Variables used in the study. Data was acquired from individual public water systems' annual consumer confidence reports.

Influence	Variable	Units
Scale-formation	Total Dissolved Solids	ppm
	Hardness	ppm
Corrosivity	pH	
	CSMR (ratio of Chloride to sulfate)	
	Alkalinity	ppm

I acquired water quality data from annually published consumer confidence reports (CCR) for individual community water supply systems. The water quality variables I chose to examine were the (a) pH, (b) total dissolved solids (TDS) content, (c) hardness, (d) chloride-to-sulfate-mass ratio, and (e) alkalinity (Table 1) of treated water prior to its distribution. While community water supply systems do additionally report lead contamination in accordance with primary drinking water standards set by the State Water Resources Control Board (SWRCB) and LCR, discrepancies between system-wide and school sampling results pointed to instances of misreporting on CCRs. This may happen for a number of reasons, including sample site choice, sampling technique, or general institutional oversight. For example, in its 2017 CCR, the City of Hanford, California, reported that none of its thirty-five total drinking water samples exceeded federal action levels, despite six of Hanford's fourteen public schools having reported WLL ranging from 15 to 185 ppb (see Appendix, Table 2). This inconsistency motivated my choice to work with school sampling WLL data, as opposed to lead data presented in systems' CCRs.

Data Processing

In order to eliminate multiple sampling results for a single school, I worked in Excel to find the average recorded water lead level for each school. After deleting any remaining duplicates, the database consisted of a single WLL measurement per school for 2,870 schools within 146 public water systems in California. I then imported the processed spreadsheet into ArcMap (An application of Geographic Information Systems) as a CSV file and geocoded the school sampling data. To better understand the distribution of these data across water systems, I acquired a shapefile of California water districts, published by the Department of Water Resources, from the State's Open Data Portal, and overlaid the geocoded WLL samples onto this water systems map. I subsequently used spatial concentrations of data points and variation in WLL to identify and select twenty-one water supply systems on which to focus this study. These twenty-one systems contained a total of 1,831 sampled schools, with WLL measurements ranging from less than 5 ppb to 185 ppb (see Appendix, Table 2).

So as to incorporate poverty rate data into the final database, I again utilized ArcMap. I imported poverty rate data as a CSV file, as well as a map of all California census tracts, which I downloaded as a TIGER/Line shapefile, and used the join feature to add poverty rate as a column in the census tract table. After clipping the census tract shapefile to my selected districts, I used the spatial join tool to average the census tract data across each water district polygon, thereby calculating the average poverty rate for each of the chosen twenty-one public water systems. I employed the join function one last time in order to apply district wide poverty data to the school sampling spreadsheet. Once completed, I exported the resulting table back into Excel, where I then manually inputted water quality data.

While I had initially anticipated that I would be able to include all water quality parameters of interest (Table 1), I was only able to incorporate total dissolved solids (TDS), water hardness, and chloride-to-sulfate mass ratio (CSMR) in the final model, since many of the selected water systems did not include the pH or alkalinity of treated water in their annual CCRs. Additionally, because the concentration of chloride and sulfate ions are listed separately in CCRs, I calculated the chloride-to-sulfate mass ratio myself by simply dividing chloride, expressed in parts per million, by the reported sulfate, also in ppm. In the event that a water system had more than one distribution facility or water source, for example if a distribution area received a mixture of local groundwater and water purchased from a regional wholesaler, which is typical throughout California, I calculated the average level detected of each variable, and applied that quantity to the entire system. I was unable to calculate a weighted average, as water systems seldom report the fraction of each water source that is present in combined and distributed water. Out of the twenty-one supply systems for which I collected water quality data, seventeen systems display measurements from 2018 CCRs, one from 2017, one from 2015, and two from 2019. After aggregating all individual datasets and necessary information, the final database consisted of 21 districts and 1,831 schools, where all sampled schools within the same district possessed identical, district-wide water quality and poverty measurements (see Appendix, Table 2)

Analysis

Linear Regression

To better understand how each variable separately affects WLL, I read the final spreadsheet, with all twenty-one districts and accompanying school WLL samples, water quality data, and poverty rate measurements into R (R 2014). I subsequently performed separate regressions for each independent variable, to examine the strength of any linear relationships between lead concentration (the dependent variable) and each individual chemical water quality and poverty variables (the independent variables). I ultimately repeated this process six different times, for both the variables which were later integrated into the final model (TDS, hardness, CSMR, and poverty rate), as well as those which were excluded (pH and alkalinity). The lack of accessible CCR data for chemical properties pH and alkalinity resulted in a significant reduction of analyzed data points; however, I still felt it necessary to run basic regressions for pH and alkalinity, despite their omittance from the final multiple regression model, because understanding – albeit imprecisely – their influence on WLL fostered increased comprehension regarding how inclusion of such variables might have affected the results of this study.

Multiple Regression Modeling

To investigate the proportion of water lead level variation explained by the selected chemical and socioeconomic indicators in combination, I performed multivariable linear regression in R. As an extension of basic linear regression, multiple regression utilizes a similar modeling structure, where the predicted outcome (WLL) is based on a y-intercept and multiple independent variables, each of which possess some unknown beta coefficient that denotes the degree of change in the outcome variable for every unit of change in the respective predictor variable. In this case, I defined the indicator variables as water hardness, TDS, CSMR, and poverty rate, and employed the required functions in R to summarize and plot the resulting model. The purpose of performing such an analysis was two-fold: to first compute the beta coefficients for each variable and identify the appropriate y-intercept, and additionally gather data regarding the accuracy of the model. I compared values of goodness of fit (R^2) in order to determine the best-fit model out of the four viable linear regression models for each isolated variable (not including pH and alkalinity), as well as the multiple regression model. In both linear and multiple regression analysis, my null hypotheses purported that no statistically significant relationship exists, and that any variation in WLL is random, while my alternative hypotheses offered up the contradictory possibility that there is a statistically significant relationship between the input and output variables.

RESULTS

Regressions and Best Fit

Conducting linear regressions between each individual independent variable and lead concentration revealed a poor capacity of the selected water quality and poverty variables to predict changes in WLL. For CSMR, hardness, TDS, and poverty rate, the correlation coefficient values (R) were 0.2819, -0.0849, -0.0106, and -0.0484, respectively. In other words, with the exception of CSMR, the variables on their own have weak statistical relationships with WLL, which in turn elicit low goodness of fit (R^2) values as well (Table _). While these low correlation coefficients make known the insignificant degree to which WLL is explained by the individual models, the p-values suggest the existence of several noteworthy relationships between input variables and WLL. The calculated p-values for both CSMR and hardness were considered statistically significant given a significance level of 0.01; by increasing this threshold to 0.05, the p-value computed for poverty rate also indicated the presence of a significant relationship described by the model (Table

3).

Table 3. Linear Regression Statistics. Result matrix of linear regression analysis for each variable as calculated in R, formatted to display basic regression statistics, including correlation coefficient, coefficient of determination, and p-value for each of the predictor variables.

	Correlation coefficient (R)	*Coefficient of Determination (R^2)*	*p-value*
CSMR	0.2819	0.0795	2.20E-16
Water Hardness	-0.0869	0.0072	0.00028
TDS	-0.0106	0.0001	0.6492
Poverty Rate	-0.0484	0.002	0.0383

In comparing each modeling exercise, I found that multivariable linear regression analysis generated the highest coefficient of determination, with a value that suggests 8.353% of the variation in WLL can be explained by the predictor variables pH, CSMR, TDS, and poverty rate, assuming that every variable included in the model impacts the dependent variable. When the model is adjusted to consider only those independent variables which actually affect WLL, R^2 is slightly lower, at 8.153%. The minor comparative advantage of the multiple regression model over basic regression gives the former the title of best-fit model out of the five investigated in this study. The nature of the relationships identified in multiple regression contrast those found during individual linear regression, particularly with respect to statistical significance. Regardless of whether the significance level is set at 0.05 or 0.01, TDS presents a statistically insignificant p-value (Table 4). Though the p-values for hardness and poverty are also technically insignificant, they are only slightly above a 0.05 significance level; CSMR, on the other hand, maintained a consistently small p-value across both modeling techniques, and is therefore considered to be statistically significant (Table 4). Given the coefficient of determination calculated for CSMR during basic regression analysis, it appears as though the degree to which the model explains WLL relies heavily on CSMR, while the other predicator variables have very little influence. The low R^2 value in combination with largely statistically insignificant p-values disallows rejection of the hypothesis that WLL variation is due to random chance and shows that the final model structure is not an accurate means of predicting WLL.

Table 4. Multiple Regression Statistics. The results of multivariable linear regression analysis as calculated in R, formatted to display basic regression statistics, including correlation, coefficient of determination, and beta coefficients for each of the predictor variables.

Multiple R	0.2890
R-squared	0.0835
Adjusted R-squared	0.0815
Standard Error	7.4925 on 1825 degrees of freedom

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	9.1626	1.5576	5.8823	4.801E-09
Poverty Rate (%)	-0.0972	0.0501	-1.9429	0.0522
TDS (ppm)	0.0038	0.0029	1.3060	0.1917
Hardness (ppm)	-0.0125	0.0064	-1.9342	0.0532
CSMR	0.0192	0.0021	9.2353	6.9282E-20

DISCUSSION

In creating the final model, I found that poverty rate and chemical water quality data had a weak influence on the variation in WLL. There was no expectation of perfect correlation between the predictor variables and WLL, but the degree of weakness of the model overall was unanticipated, given common assertions about the effects of water quality characteristics and social circumstances on lead leaching. However, while my results did not display strong statistical significance, and I was therefore unable to move forward with the existing model, this exercise provides new insights into how we might alter the process of model creation in the future to better predict lead contamination under uncertain conditions in an effort to mitigate the disproportionate exposure of vulnerable populations.

Challenges

Limitations of Study Design

Modeling Approach. The findings of multiple regression analysis presented no evidence of any strong linear relationships between the indicator data and WLL. Temporarily neglecting the potential reality that a correlation simply may not exist, it is important to recognize that the method of model creation utilized may not have been appropriate for the data at hand. Multiple regression analysis makes number of assumptions – first and foremost that the relationship between independent and dependent variables are linear. It might be that a relationship does indeed exist and has simply been miscategorized. In the literature surrounding chemical composition and elevated WLL, relationships between water quality data and WLL are typically vaguely discussed with the simple descriptors "high" or "low" (Sherlock 1984; WHO 2011; Tam & Elefsiniotis 2009; Nguyen et al. 2011), as opposed to precise mathematical associations. While the use of such language may loosely imply linearity, exploring options of curvilinear or nonlinear modeling would assist in identifying the nuances of the transition from "high" to "low".

Inadequate sample size also represents a source of potential error in the study's findings. As aforementioned, my research incorporated 1,831 of the total 2,870 public schools for which WLL was reported, in 21 of the total 146 public water supply systems throughout California. While the reduction in school sampling data left a sizeable number of data points, the near seven-fold drop in studied water systems may have diminished the accuracy of the model, as each water system was assigned a single value for each predictor variable. Essentially, the simple multiple regression model was attempting to decipher the extent of variation in the 1,831 sampling results explained by only 21 system-specific measurements for water quality and poverty rate.

Indicators and WLL. Decades of meticulous research have described the ways in which chemical composition can influence the corrosivity of a water supply and react with household plumbing and metal fixtures, resulting in the deterioration of the pipes and increased lead concentration in a residence's drinking water (WHO 2011 ; Xie & Giammar 2011 ; Schock 1989). Such universally accepted truths regarding the chemical processes driving lead leaching were not reflected in the results of this study, subsequently prompting enquires into the limitations of the study's design, particularly with respect to sampling size and the process of calculating water quality indicators. As aforementioned, in the final database, all sampled schools within the same district possessed identical, district-wide water quality measurements. Not only was this a less-than-ideal quantity of water chemistry data, the complex nature of water supply systems throughout California inherently complicated the process of determining the correct values for the selected indicators. More than half of the studied water supply systems, particularly in Southern California, distribute water that is an amalgamation of either multiple water treatment plants, or of local and purchased water, and each source often reports differing water quality values. In these instances, where it was impossible to know the proportions of each water source present in residents' taps, I was only able to calculate a simple average to reflect the water quality conditions of the entire distribution area. This is turn had the potential to negate the data's veracity and damage the accuracy of the model.

In the context of social equity analysis, while there is something to be gained from a larger study site, including a more diverse sampling pool, the averaging of poverty rates to represent entire water supply systems inevitably results in the overgeneralization of socioeconomic circumstances. Fluctuations in wealth occur at small scales, from neighborhood to neighborhood, as much as they do from county to county. Take the case of East Bay Municipal Utility District, for example, which supplies drinking water for the cities of Richmond, Berkeley, Piedmont, Oakland, Walnut Creek, and many others. Just as Oakland and Walnut Creek represent two extremes of wealth and poverty, two neighborhoods residing within Oakland mere miles apart have the capacity to display such a dichotomy as well (USCB 2000). This issue of scale prompts inquiry into the level of geospatial aggregation of data (census tracts, school districts, county, etc.) which would prove most appropriate and beneficial in the creation of a predictive model. On one hand, a small-scale focus may increase the accuracy of predictions but diminish the possibility of scalability, while a larger study site may be more scalable, but involve vast overgeneralization. Such considerations were integral in initially establishing the geographic constraints of this study, and any future choices regarding project area will continue to greatly shape the accuracy and scalability of the model.

Barriers to successful modeling

Social Factors and Model Scalability. The accurate application of this proposed model relies almost entirely on its capacity to maintain its correctness while expanding its scale. While the tenants of water chemistry and corrosivity remain true regardless of geographic location, social circumstances do not, so any environmental inequity element integrated into the final model must account for such discontinuity.

Currently, much of the work exploring environmental injustice in lead exposure has been conducted in the United States, where the burden of lead's adverse health effects is disproportionately borne by low-income and minority residents of older, deteriorating housing stock (Kordas et al. 2018 ; Bernard & McGeehin 2003 ; Leech et al. 2016). Increasingly, this geospatial clustering is understood to represent structural inequality deeply rooted in a history of residential segregation (Sadler et al. 2017; Jacobs 2011). Of course, underlying socioeconomic, cultural, and historical influences differ from country to country, therefore it is expected that explanatory factors for differential lead exposure within low and middle-income countries will diverge from those in the U.S. Consequently, the generation of a model built with data from the U.S. and intended for worldwide application can only incorporate relative social variables that are internationally consistent.

Future directions

Going forward, this study may be modified in a number of ways to increase the likelihood of attaining statistically significant results, which in turn brings closer the reality of an accurate and applicable WLL-predicting model.

At the very least, additional data are necessary to construct a more precise, suitable model for water lead level prediction. In an ongoing effort to improve upon this current study, I have worked alongside my mentor, Gabriel Lobo, to collect water quality data for the remainder of California water systems containing school sampling WLL data, thereby increasing the number of water quality measurements nearly seven-fold. Furthermore, future research should investigate nonlinear methods of statistical analysis, such as multivariate logistic regression, as well as more advanced methods of predictive modeling, such as Decision Trees and Random Forests, which use machine learning to make predictions (Yiu 2019).

Future research should also consider expanding on both the water quality and social indicators included in this study. For example, the amount of the lead dissolved from the plumbing system depends not only on the characteristics addressed in this study, but on other factors as well, such as the presence of dissolved oxygen, water temperature, and a reduced oxidation-reduction potential from a change in disinfectant (WHO 2011). Additionally, indices like the Langelier Saturation Index (LSI), or Ryznar Stability Index (RSI), amalgamate various water quality properties as a means of evaluating if the water flowing through a pipe has a tendency to form a scale (Awatif et al. 2014). LSI in particular is calculated using pH, temperature, TDS, alkalinity, and calcium hardness, while RSI uses LSI as a component within its own formula (Alsaqqar et al. 2014). Such indices act in a way similar to CSMR, in that they combine values that may otherwise have little impact on WLL to generate a value with an augmented influence on corrosivity and lead leaching. Therefore, going forward, it would be prudent to integrate additional parameters into the model while simultaneously examining how such independent variables might work in tandem to influence water lead levels.

In terms of social variables that may be added to enhance the model's accuracy, future research should work to identify relative measurements of internationally applicable social conditions. The Gini Coefficient, for instance, epitomizes the notion of scalability, as it offers a comparative metric of inequality by summarizing the dispersion of income across a country's entire income distribution (USCB 2016).

Broader Implications

At present, the deterioration of lead plumbing fixtures subjects millions of already vulnerable individuals to the myriad adverse health effects of elevated WLLs, consequently propagating inequality within communities and between countries (Meyer et al. 2008). In low and middle-income countries, where BLL is disproportionately high and WLL compliance programs are often unenforced or altogether absent (Romieu et al. 1997; Ngueta et al. 2013), the issue of lead exposure represents a prominent threat, and is not likely to recede from the global arena in the near future (Kordas et al. 2018). So, in the same way that recent violations of the Lead and Copper Rule have generated a call for corrective action and a focus on social equity within the United States, so too should the global community address the public health crisis and environmental injustice implications associated with lead exposure across the world.

This study represents a small first step in what will no doubt be an iterative process of model development for WLL prediction. There are many elements of this study design that necessitate alteration and improvement; however, the purpose of the proposed methodology remains relevant and essential. In employing more advanced methods of statistical analysis and predictive technologies, it may be possible to estimate lead exposure hotspots in countries where such information is unknown, thus providing governments with the tools and incentive necessary to carry out desperately needed infrastructure replacements and eliminate the presence of lead in drinking water.

ACKNOWLEDGEMENTS

I would first like to thank Samuel Evans, Patina Mendez, and the wonderful GSIs, Roxy, Leslie, and Jessica, for being a constant source of support for all the Environmental Science seniors and providing motivation and encouragement for me throughout the year. To my mentor, Gabriel Lobo, I want to express my gratitude for being such wonderful and patient teacher to me in all of this, for offering feedback and help whenever possible, and giving me the chance to participate in his own research on ALPS. Additionally, having the opportunity to work under Dr. Ashok Gadgil, who inspired my passion for environmental work as a sixth grader, was an absolute honor, and I am grateful that he connected me with Gabriel, as well as Dr. Mike Sohn, whose advice acted as the initial inspiration for this project. And in my jobs, I have been lucky enough to have understanding bosses that always understood the difficulty of being a working student. With regards to my subject matter, I want to thank EdSource, particularly Daniel Willis, for working to bring awareness to the issue of lead contamination in California schools, and directing me towards the dataset on which I based this project. Finally, to my family and friends, especially my roommates, Nicole, Yesenia, and Roya, thank you for always providing top notch moral support and dancing with me at the end of long days.

REFERENCES

Alsaqqar, A.S., B.H. Khudair, & S.K. Ali. 2014. Evaluating Water Stability Indices from Water Treatment Plants in Baghdad City. Journal of Water Resource and Protection. 06:1344–1351.

Bernard, S.M., & M.A. McGeehin. 2003. Prevalence of Blood Lead Levels ≥5 μg/dL among US Children 1 to 5 Years of Age and Socioeconomic and Demographic Factors Associated with Blood of Lead Levels 5 to 10 μg/dL, Third National Health and Nutrition Examination Survey, 1988-1994. Pediatrics. 112:1308–1313.

Bill Text - AB-746 Public health: potable water systems: lead testing: schoolsites. (n.d.). Retrieved March 29, 2020, from https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB746

Brown, M.J., & S. Margolis. 2012. Lead in Drinking Water and Human Blood Lead Levels in the United States. Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/su6104a1.htm

Bureau, U.S.C. (n.d.). Gini Index.

California Water Boards. 2019. Lead Sampling of Drinking Water in California Schools Results.

Crowder, K., & L. Downey. 2010. Interneighborhood migration, race, and environmental hazards: Modeling microlevel processes of environmental inequality. American Journal of Sociology.

Dignam, T., R.B. Kaufmann, L. Lestourgeon, & M.J. Brown. 2019. Control of Lead Sources in the United States, 1970-2017: Public Health Progress and Current Challenges to Eliminating Lead Exposure. Journal of Public Health Management and Practice, 25(Suppl 1 LEAD POISONING PREVENTION), S13–S22.

Downey, L. (2006). Environmental Racial Inequality in Detroit. Social Forces.

Drinking water distribution systems: Assessing and reducing risks. 2007. In Drinking Water Distribution Systems: Assessing and Reducing Risks.

Duruibe, J., M.O. Ogwuegbu, & J.N. Egwurugwu. 2007. Heavy Metal Pollution and Human Biotoxic Effects. International Journal of Physical Sciences. 2:112–118.

Edwards, M., & S. Triantafyllidou. 2007. Chloride-to-sulfate mass ratio and lead leaching to water. American Water Works Association. 99:96–109.

How is poverty measured? – INSTITUTE FOR RESEARCH ON POVERTY – UW–Madison. (n.d.). Retrieved May 11, 2020, from https://www.irp.wisc.edu/resources/how-is-poverty-measured/

Jacobs, D. E. 2011. Environmental health disparities in housing. American Journal of Public Health, 101(SUPPL. 1).

Kim, E. J., & J.E. Herrera. 2010. Characteristics of lead corrosion scales formed during drinking water distribution and their potential influence on the release of lead and other contaminants. Environmental Science and Technology. 44:6054–6061.

Kordas, K., J. Ravenscroft, Y. Cao, & E.V. McLean. 2018. Lead exposure in low and middle-income countries: Perspectives and lessons on patterns, injustices, economics, and politics. International Journal of Environmental Research and Public Health. 15.

LaFrance, D. B. 2016. Open Channel -- Restoring Faith. American Water Works Association. 108:10.

Lanphear, B. P., M. Weitztnan, & S. Eberly. 1996. Racial differences in urban children's environmental exposures to lead. American Journal of Public Health.

Leech, T.G.J., E.A. Adams, T.D. Weathers, L.K. Staten, & G.M. Filippelli. 2016. Inequitable chronic lead exposure: A dual legacy of social and environmental injustice. Family and Community Health. 39:151–159.

Meyer, P. A., M.J. Brown, & H. Falk. 2008. Global approach to reducing lead exposure and poisoning. Mutation Research - Reviews in Mutation Research. 659:166–175.

Nevin, R. 2000. How Lead Exposure Relates to Temporal Changes in IQ, Violent Crime, and Unwed Pregnancy. Environmental Research Section A. 83:1–22.

Ngueta, G., & R. Ndjaboue. 2013. Blood lead concentrations in sub-Saharan African children below 6 years: systematic review. Tropical Medicine & International Health. 18:1283–1291.

Nguyen, C.K., K.R. Stone, & M.A. Edwards. 2011. Chloride-to-sulfate mass ratio: Practical studies in galvanic corrosion of lead solder. Journal / American Water Works Association.

NRDC [National Resource Defense Council] 2016. What's in your water? Flint and Beyond. Analysis of EPA Data Reveals Widespread Lead Crisis Potentially Affecting Millions of Americans. A Report by NRDC, authors Olson E., and Fedenick, K.P. NRDC, New York, N.Y., USA

Prüs-Ustün A., L. Fewtrell, P.J. Landrigan, J.L. Ayuso-Mateos. Lead Exposure. In: Ezzati M., A.D. Lopez, A. Rodgers, C.J.L. Murray, editors. 2004. Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors. Volume 1. World Health Organization; Geneva, Switzerland. 1495–1542.

R Development Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computer, Vienna, Austria. http://www.R-project.org/.

Rawson, J., & R. Ayala. 2005. Method and system for controlling corrosivity of purified water.

Replacing all lead water pipes could cost $30 billion | Water Tech Online. 2016. Retrieved May 11, 2020, from https://www.watertechonline.com/home/article/15549954/replacing-all-lead-water-pipes-could-cost-30-billion

Romieu, I., M. Lacasana, & R. McConnell. 1997. Lead exposure in Latin America and the Caribbean. Environmental Health Perspectives. 105:398–405.

Sadler, R.C., J. LaChance, & M. Hanna-Attisha. 2017. Social and built environmental correlates of predicted blood lead levels in the flint water crisis. American Journal of Public Health. 107:763–769.

Sampson, R.J., & A.S. Winter. 2016. Toxic Inequality in Chicago Neighborhoods. Du Bois Review. 1995–2013.

Schock, M.R. 1990. Causes of temporal variability of lead in domestic plumbing systems. Environmental Monitoring and Assessment. 15:59–82.

Sherlock J.C., D. Ashby, H.T. Delves, G.I. Forbes, M.R. Moore, W.J. Patterson, S.J. Pocock, M.J. Quinn, W.N. Richards, T.S. Wilson. 1984. Reduction in exposure to lead from drinking water and its effect on blood lead concentrations. Human Toxicology. 3:383–392.

Tam, Y.S., & P. Elefsiniotis. 2009. Corrosion control in water supply systems: Effect of pH, alkalinity, and orthophosphate on lead and copper leaching from brass plumbing. Journal of Environmental Science and Health - Part A Toxic/Hazardous Substances and Environmental Engineering. 44:1251–1260.

USCB [United States Census Bureau]. 2000. Employment, Income, and Poverty.

Udedi S.S. 2003. From Guinea Worm Scourge to Metal Toxicity in Ebonyi State, Chemistry in Nigeria as the New Millennium Unfolds. 2:13-14.

Williams, D.R., & C. Collins. 2001. Racial residential segregation: A fundamental cause of racial disparities in health. Public Health Reports.

WHO [World Health Organization]. 2011. Guidelines for Drinking-water Quality, 4th edition. WHO, Geneva, Switzerland.

WHO [World Health Organization]. 2018. A Global Overview of National Regulations and Standards for Drinking-water Quality. WHO, Geneva, Switzerland.

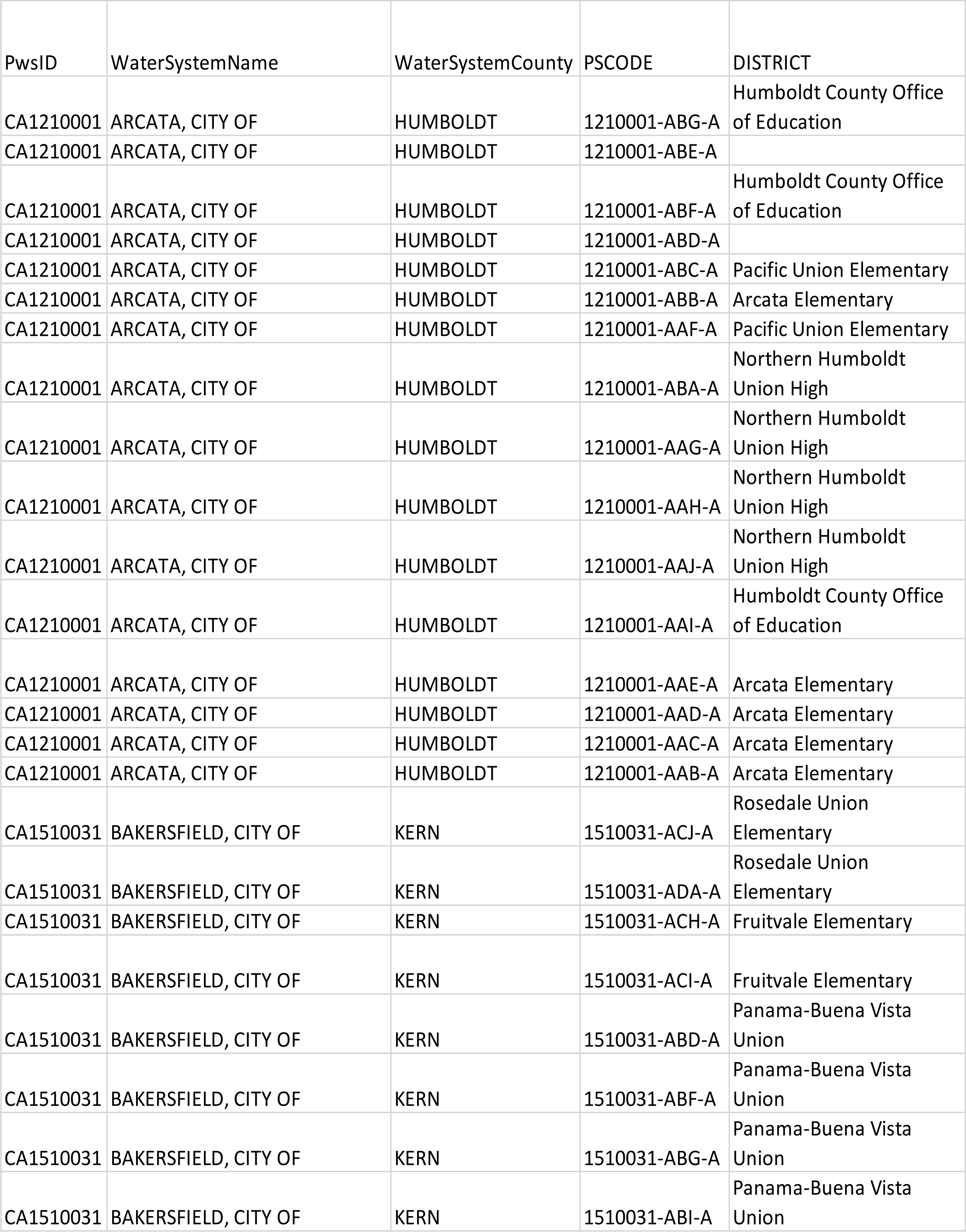
Reyes, J., & A. College. 2007. Environmental Policy as Social Policy? The Impact of Childhood Lead Exposure on Crime. In The B.E. Journal of Economic Analysis & Policy. 7.

Xie, Y., & D.E. Giammar. 2011. Effects of flow and water chemistry on lead release rates from pipe scales. Water Research. 45:6525–6534.

Yiu, T. 2019. Understanding Random Forest. Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

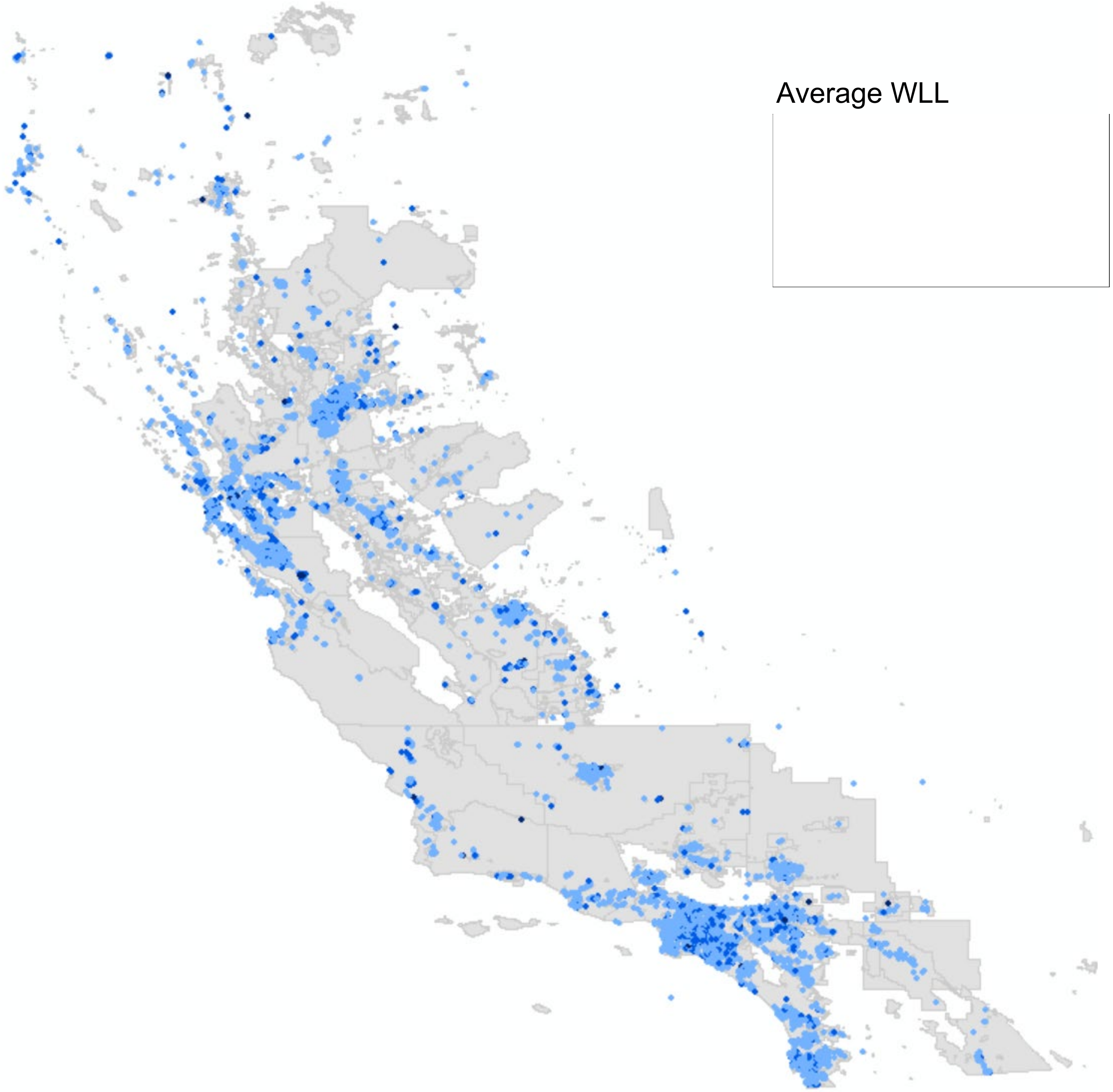
APPENDIX

Table 2. Final Database. Double click the table to navigate the final database used in this study.

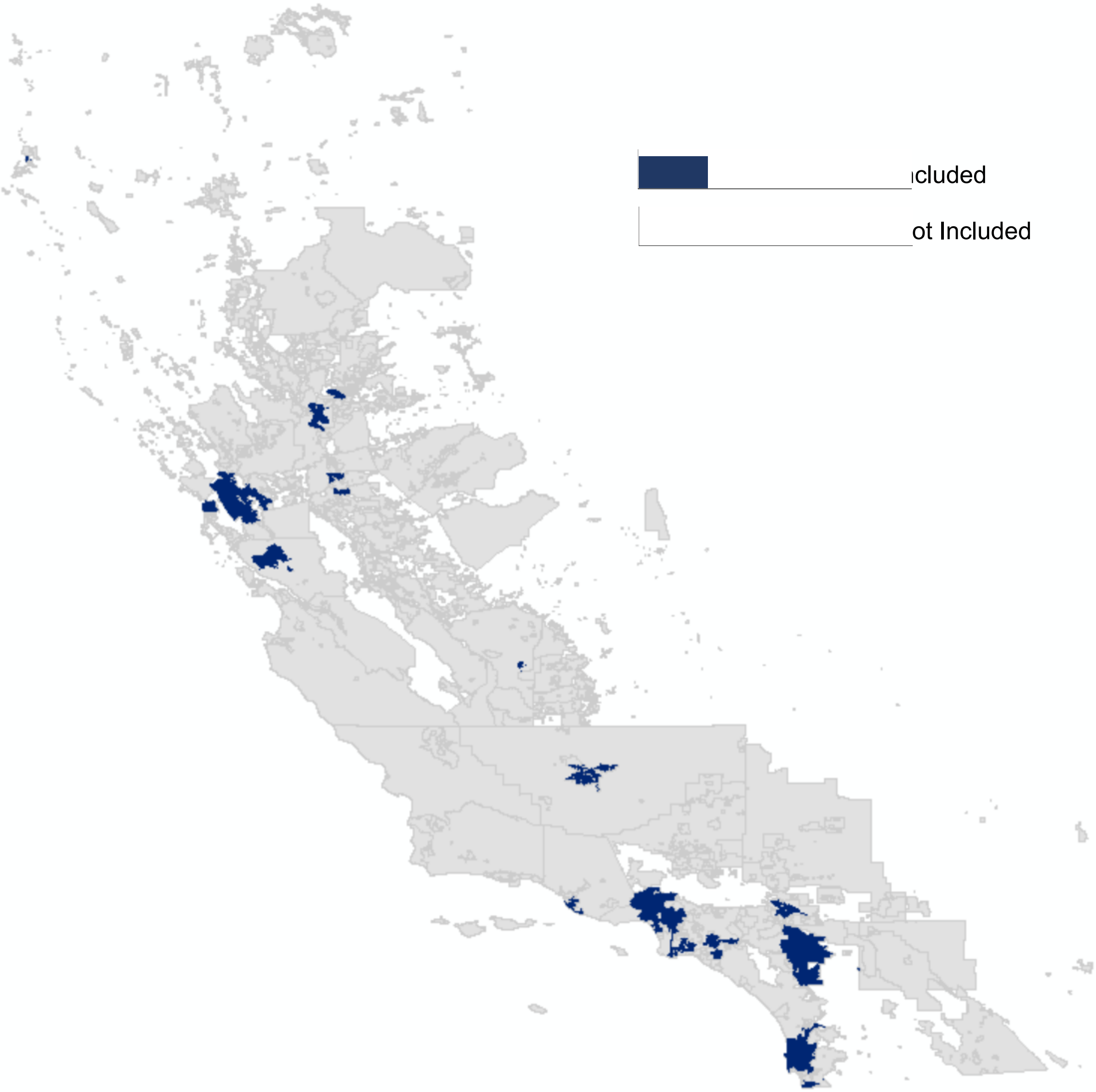


PwsID	WaterSystemName	WaterSystemCounty	PSCODE	DISTRICT
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-ABG-A	Humboldt County Office of Education
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-ABE-A	
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-ABF-A	Humboldt County Office of Education
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-ABD-A	
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-ABC-A	Pacific Union Elementary
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-ABB-A	Arcata Elementary
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-AAF-A	Pacific Union Elementary
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-ABA-A	Northern Humboldt Union High
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-AAG-A	Northern Humboldt Union High
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-AAH-A	Northern Humboldt Union High
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-AAJ-A	Northern Humboldt Union High
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-AAI-A	Humboldt County Office of Education
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-AAE-A	Arcata Elementary
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-AAD-A	Arcata Elementary
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-AAC-A	Arcata Elementary
CA1210001	ARCATA, CITY OF	HUMBOLDT	1210001-AAB-A	Arcata Elementary
CA1510031	BAKERSFIELD, CITY OF	KERN	1510031-ACJ-A	Rosedale Union Elementary
CA1510031	BAKERSFIELD, CITY OF	KERN	1510031-ADA-A	Rosedale Union Elementary
CA1510031	BAKERSFIELD, CITY OF	KERN	1510031-ACH-A	Fruitvale Elementary
CA1510031	BAKERSFIELD, CITY OF	KERN	1510031-ACI-A	Fruitvale Elementary
CA1510031	BAKERSFIELD, CITY OF	KERN	1510031-ABD-A	Panama-Buena Vista Union
CA1510031	BAKERSFIELD, CITY OF	KERN	1510031-ABF-A	Panama-Buena Vista Union
CA1510031	BAKERSFIELD, CITY OF	KERN	1510031-ABG-A	Panama-Buena Vista Union
CA1510031	BAKERSFIELD, CITY OF	KERN	1510031-ABI-A	Panama-Buena Vista Union

A 1. Map of School Sampling WLL Results. This map displays the 2,870 total public schools for which sampling took place in compliance with AB 746 throughout California.



A 2. Map of Final Selected Water Supply Systems included in this study. Selected public water supply systems included the Cities of (1) Anaheim, (2) Arcata, (3) Bakersfield, (4) Fullerton, (5) Hanford, (6) Hayward, (7) Long Beach, (8) Los Angeles (department of water and power), (9) Oxnard, (10) Roseville, (11) Sacramento (Main), (12) San Bernardino, (13) San Diego, (14) San Jose (Main), (15) San Jose (Evg/Edv/Coy service areas), (16) Santa Ana, and (17) Stockton; (18) the East Bay MUD (serves East Bay cities in the SF Bay), (19) East Valley Water District (San Bernardino), (20) Eastern Municipal Water District (Riverside), and (21) San Francisco Public Utilities Commission (SFPUC)


ıcluded
ot Included