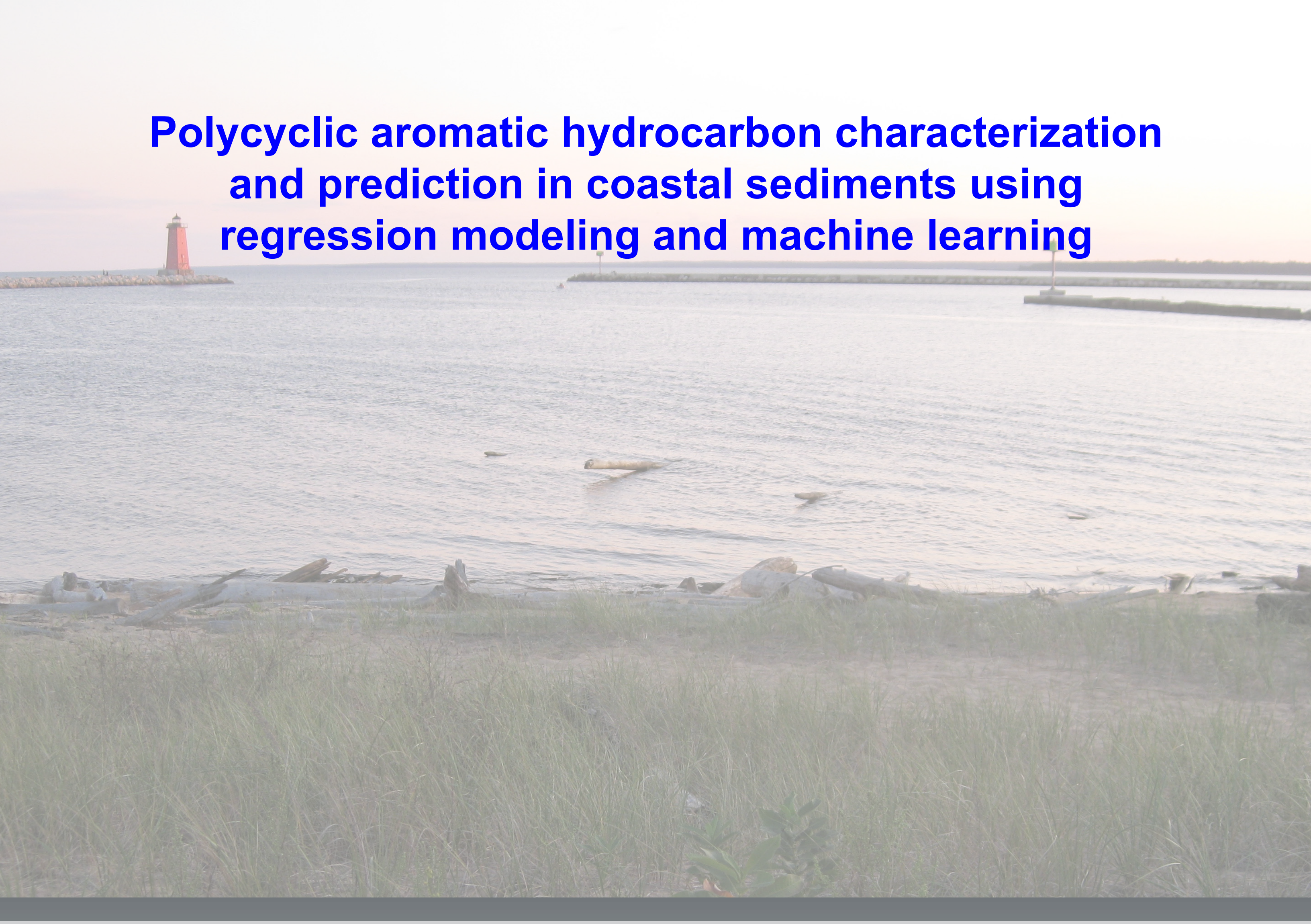
Polycyclic aromatic hydrocarbon characterization and prediction in coastal sediments using regression modeling and machine learning



NOAA National Ocean Service
National Centers for Coastal Ocean Science
Marine Spatial Ecology Division [1]
Stressor Detection and Impacts Division [2]
Consolidated Safety Services, Inc. [3]

Amy Freitag[1]
Seann Regan[3,1]
A. K. Leight[1]
Kimani Kimbrough[2]
Mary Rider[3,2]

June 2021

NOAA TECHNICAL MEMORANDUM NOS NCCOS 293

NOAA NOS NCCOS Monitoring and Assessment Branch

Acknowledgements

The authors would like to extend our appreciation to NCCOS staff broadly for supporting efforts at interdisciplinary work, both in this project and in creating space for growing new collaborations. Much appreciation is also due to the staff of the Mussel Watch and Bioeffects programs from 1985 until present, whose many years hard work at sampling and contaminant testing is highlighted in the PAH dataset used in this study. The pictures in this report are also from their field work. We kindly acknowledge reviewers of this manuscript: Terry McTigue, Tony Pait, and Ed Wirth. We also acknowledge colleagues who have given us much needed advice along the way: Erik Davenport and Michelle Nutt (USGS). Government contract labor was provided by CSS, Inc. under Contract No. EA133C-14-NC-1384 and Contract No. GS-00F-217CA/EA133C17BA0062.

Cover and Back

Front Cover: Photograph provided by Kimani Kimbrough. Manistique River, MI.

Back cover: Photograph provided by MAB photo Archives. Hudson River, NY.

Suggested Citation

Amy Freitag, Seann Regan, A. K. Leight, Kimani Kimbrough, Michael Edwards, Heidi Burkart, and Mary Rider. Polycyclic aromatic hydrocarbon characterization and prediction in coastal sediments using regression modeling and machine learning. Silver Spring, MD. NOAA Technical Memorandum NOS NCCOS 293, 27 pp.

Mention of trade names or commercial products does not constitute endorsement or recommendation for their use by the United States government.

Table of Contents

Table of Figures

Table of Tables

Executive Summary

Since 1986, the National Oceanic and Atmospheric Administration (NOAA) Mussel Watch Program (MWP) has monitored the nation's coastal waters for chemical contaminants and biological indicators of water quality. MWP works with its associated programs, the National Bioeffects Program (BE), Great Lakes Restoration Initiative (GLRI), and Placed Based Assessments (PBA), to support ecosystem management nationwide. Together these programs conduct environmental monitoring, assessment, and research to describe the status and trends in the environmental quality of the nation's estuarine and coastal waters. They utilize a sentinel-based approach to monitoring, by collecting and analyzing sediment and bivalves (oysters and mussels) as surrogates for water pollution and bioaccumulation. Contaminants monitored by the MWP include legacy organic chemicals, such as organochlorine pesticides, industrial contaminants, fossil fuel combustion byproducts, and metals. In recent years, these programs have added contaminants of emerging concern (CECs) to their analyses including pharmaceuticals and personal care products, alternative flame retardants, and alkylphenol and perfluorinated compounds.

In order to answer critical questions about contaminant source and fate, the National Centers for Coastal Ocean Science (NCCOS) is developing new strategies to link the chemical contaminants detected in bivalves and tissue to potential sources of contamination. Through the development of a cross-division collaboration, a team of scientists explored different modeling and machine learning techniques to improve our understanding of the complex interactions between the environment and the chemical contaminants detected. By leveraging the resources of the Monitoring and Assessment Branch (Stressor Detection and Impacts division and the Cooperative Oxford Laboratory and Biogeography Branch (Marine Spatial Ecology Division), scientists have developed nationwide models that draw information from a wide variety of resources to explore factors that drive the fate and distribution of chemical contaminants in our coastal waters. Not only is this information unparalleled in its scope, the resulting models can be applied to the ever-growing database of chemical contaminant data in order to help describe emerging coastal contaminant observations. These types of analyses also help develop programmatic goals by identifying potential data gaps and demonstrating the power of continued cross-disciplinary collaboration.

NCCOS' mission is to provide coastal managers with scientific information and tools needed to balance society's environmental, social, and economic goals. This report supports that goal by developing adaptable tools that can be applied to new and existing datasets to further our understanding of the environmental condition of the nation's coastal waters. By creating these tools, we are strengthening the linkage between science and management as well as highlighting the connectivity between the nation's lands and its coastal waters.

Introduction

This project assessed and characterized the relationship between PAHs and spatial social data, and resulted in:

- Development of a contaminant forecasting model
- Identification of relevant independent variables

The driving force behind this project, and most of the research conducted by the National Centers for Coastal Ocean Sciences (NCCOS) is the recognition that humans and the coastal environment are intimately linked. Humans tend to reside near large bodies of water. Humans also rely heavily on natural resources. These factors make the coastal environment a crucial area economically, socially and ecologically. However, the activities of humans create pressures on the quality, resilience and sustainability of the coastal aquatic environment. One of these pressures comes in the form of chemical contaminants that find their way into coastal aquatic systems and, in many cases, into aquatic biota.

This project utilized existing coastal contaminant data from the Mussel Watch Program (MWP) and Bioeffects (BE) programs, which characterize the distribution of chemical contaminants in coastal environments at different spatial scales. In addition to MWP and BE, current data from the Great Lakes Restoration Initiative (GLRI) and Place Based Assessments (PBA) are also included in this study. MWP is a monitoring program that uses bivalves and sediment to assess the status and trends of contaminants nationally and regionally. BE, GLRI, and PBA projects focus on location-based characterization of contaminants. MAB has accumulated more than three decades of data in estuaries, coastal zones, and the Great Lakes. Along with sustained monitoring and assessment initiatives, MAB has undertaken special studies in response to natural/man-made disasters, oil spills, and executive initiatives that were included in this study. MAB methods and results are characterized in numerous reports, book chapters, and manuscripts; however, this study was a national scale synoptic analysis taking full advantage of the depth and scale of the MAB PAH data.

This project represents a first attempt to combine data from MAB projects and programs for analysis. We use an unsupervised machine learning technique to assess patterns for this data mining effort and compared it to a regression analysis.

In previous reports and in response to program reviewers, stakeholders, and managers, spatial data was used to identify relationships with land use and potential management implications. This research effort expands on this mission to explore all MAB data sources, additional human dimensions factors, and a wider geography to both explore the dynamics of chemical contamination and its relationship to particular types of human activities. Here we utilize spatial human dimensions data to further characterize polycyclic aromatic hydrocarbon (PAH) sediment results to:

1) Characterize the relationship between PAH sediment contamination and human dimensions data.

2) Characterize and predict sediment PAH concentrations nationally based on these relationships.

3) Address data gaps to increase monitoring and assessment efficiency.

Monitoring of waters, sediments, and living tissues from coastal water bodies confirms that land-based chemicals are being released by human activities (Du et al., 2020). However, not all chemicals move through or persist in the environment in the same way. Several classes of chemical contaminants resist degradation and tend to bind to sediments and/or tissues, thus persisting in the environment for long periods of time PAHs resist degradation and tend to bind to sediments and/or tissues, thus persist in the environment for long periods of time. Because of their ubiquity and persistence, PAHs make a good case study for quantitative comparisons to human dimensions.

In 2000, the MWP investigated the relationship between chemical contaminant loads in shellfish at 263 sampling sites around the country and found a moderate correlation between human population size within 20km of a site and PAH concentrations (0.47 Spearman coefficient). Some chemicals showed a much stronger relationship to human population, while others seemed to be more driven by ecosystem effects (O'Connor, T.P., 2002). A synoptic look at national shellfish tissue levels in 2005 showed regional variations in both levels of PAHs and trends, but did not investigate relationships between contaminant concentration and environmental surroundings (Kimbrough et al., 2008).

PAH basics

Polycyclic aromatic hydrocarbons (PAHs) are a common anthropogenic contaminant associated with the use and incomplete combustion of organic matter such as fossil fuels (Roldán-Wong et al., 2020). They constitute a group of several hundred compounds that share a distinct chemical structure containing two or more aromatic rings. PAHs are common components of petroleum products and can be released into the environment through oil spills, and roadway runoff. Once in the aquatic environment they tend to quickly bind to sediment particles but may also accumulate in the tissues of some organisms, particularly oysters, mussels, and other invertebrates (EPA, 2012).

Relationship of PAHs to humans

The presence of PAHs in the environment relates to the nearby human population and their fossil fuel burning practices such as cars, home heating, and power generation (Garner et al., 2009). The highest concentrations of PAHs bind to aquatic sediments and organisms in urban areas. As one of many examples, a comprehensive study of contaminants in the Chesapeake Bay found that concentrations of PAHs in sediments were significantly higher in designated industrialized watersheds as compared to rural and agricultural watersheds (Hartwell and Hameedi, 2007).

The general trend of higher PAH concentrations in water bodies near urban areas may be more directly related to specific factors about their urban construction. For example, PAH levels in sediment, oligochaetes, and grass shrimp in South Carolina were found to be related to impervious surface cover in the surrounding watershed, with urban and industrialized creeks showing the highest PAH levels (Garner et al., 2009). Population dynamics outside of land use are also related to PAH levels, such as the presence of industrial farms, roads, and commercial business facilities (Huang et al., 2017), as well as traffic (Jedynska et al., 2014). The meaning of “urban” can also be defined differently depending on the distinctive features considered. The U.S. Census Bureau (Census, 2020) defines an urban area as a place with more than 50,000 people. A group of chemists developed an urbanization indicator, which was correlated to PAH concentrations, consisting of residential building age, population density, road density, and distance from the urban centers (Peng et al., 2013). Overall, the relationships between PAHs and the above variables as discussed in the literature, point to the need for a set of indicators that includes human population size and composition, and some metric of how that population builds and powers the spaces in which it lives.

Modeling efforts to relate PAHs to humans

Given the positive correlation between PAH levels and surrounding human population and activity found in these previous studies, a variety of statistical approaches have been applied to quantify these relationships, following models developed to track environmental fate and transport of chemicals (Peng et al., 2013; Jedynska et al., 2014; and Huang et al., 2017). The first set of models, the most common, is a land use regression approach. A land use regression model developed for a particular study in Europe explained 67% of benzo[a] pyrene concentrations (the most toxic of the PAHs), with large variation detected between study areas (Jedynska et al., 2014).

More complex models, like positive matrix factorization are better able to assess source apportionment (Khairy and Lohmann, 2013). Similarly, other studies have tried regression tree analysis with inputs of population, vegetation types, and soil composition, which explained 71% of PAH concentrations (Kubosova et al., 2009). Similar results were also yielded from a GIS-based correlation analysis (Merbitz et al., 2012), logistic regression (Papritz and Reichard, 2009), and multiple regression (Noth et al., 2011), all of which focused on small-scale interactions including traffic characteristics and home heating fuel type within a few hundred meters of monitoring sites. A partial least squares regression also found a scalar effect, in which PAHs in water were related to land use metrics in the whole watershed while sediment PAH levels were related to local sources (Uher et al., 2016).

Regression methodologies fitted to the input indicators and scale of the study is the most common and most effective choice in modeling the relationship of PAH levels (Hoek et al., 2008). Variation over time and space remains a challenge in all of these examples of predicting contaminant levels. However, several of these statistical models performed well enough to categorize PAH values in sediment as over a threshold of concern, either in need of further testing and source identification or in need of remediation for residential purposes.

The possible models that are applicable for testing and linking the relationship of PAHs to land use and demographic factors also depends on the parameters of the data to be modeled, especially when using secondary data for a purpose differing from that of the original study design. For the purpose of this study, the MWP monitoring data was collected to assess the status and trends of contaminants along coastal areas, as well as the Great Lakes Basin, but does not lend itself to hypothesis testing. The BE, GLRI and PBA often use hypothesis testing techniques at the study level and share sediment sampling techniques, but were not designed for all data to be combined from decades of studies. Machine learning techniques are well suited in finding patterns in the data and identifying gaps across the program, even though data might be generated from different sample designs (e.g. monitoring versus place based).

Methods

Sediment measurements derived from multiple studies over more than three decades were combined and analyzed to gain a better perspective on PAH distribution. Consistent methodologies of the various NS&T programs over time allow for this study to take place.

Sampling sites

This study included MAB data from the continental US, Alaska, Hawaii, and Puerto Rico, with over 4000 samples. Due to the spatial limitations of human-dimensions data, only a subset of samples from the continental US were used (3722 samples). This data set included some monitoring data from the same site collected over many years (in a few rare cases, from Mussel Watch's inception in 1986 to current day) (Figure 1).

The human dimensions data needed to be aggregated to a chemical neighborhood, which can be thought of as an area where human activities within some distance of a sample would be expected to increase the PAH levels in sediment. One approach to defining this neighborhood is to use circular buffers around the sample site, while another is to incorporate hydrologic flow that might carry PAHs in the primary direction of water flow. Both neighborhood types should be relatively small in scale (less than 20 km according to Peng et al., 2013). As such, buffers from 1-5 km were created for each sample site to capture the strongest chemical signal. Buffers of 1, 2, 3, 4, and 5 km are shown in Figure 2.

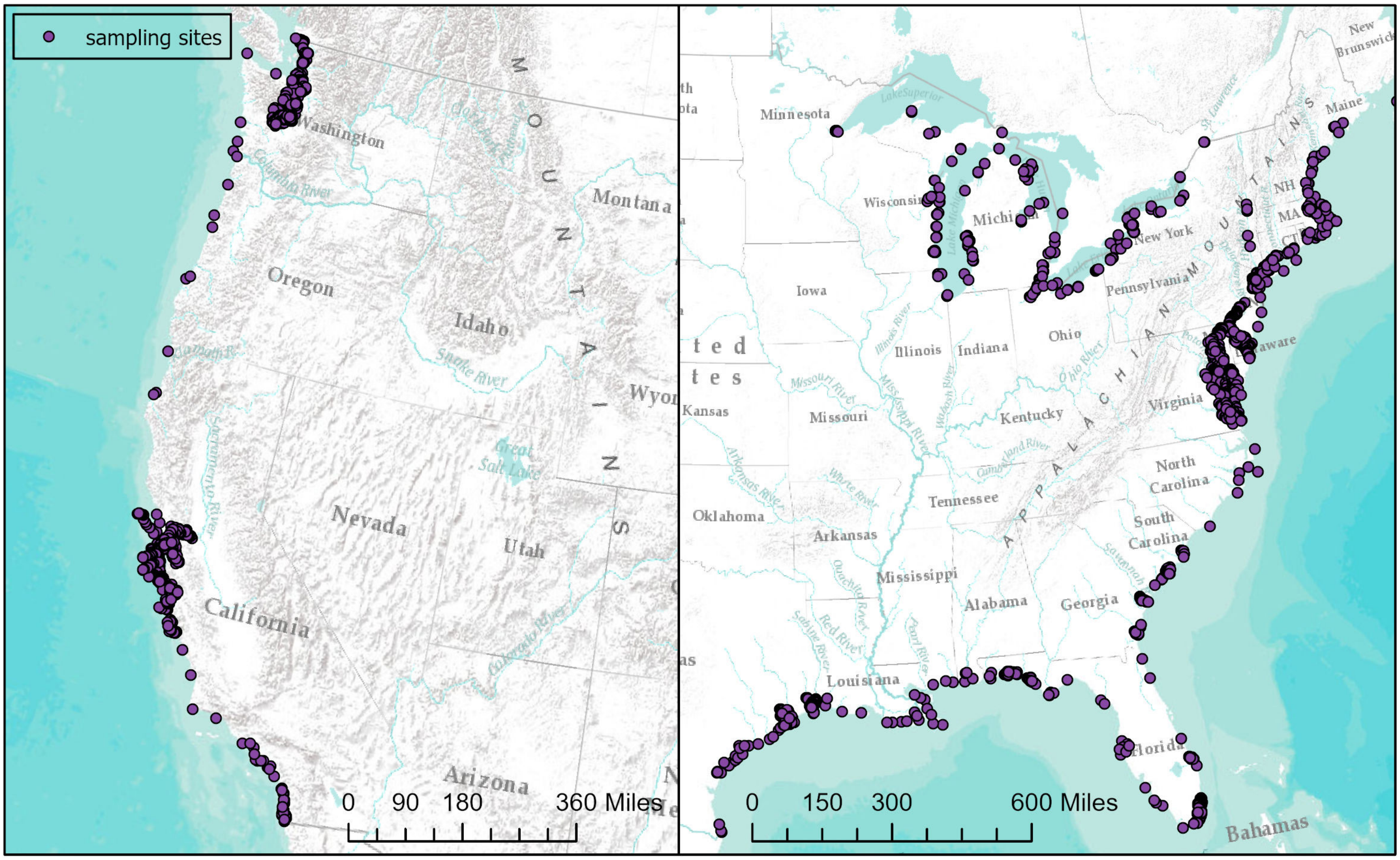


Figure 1. NS&T continental US PAH sediment sampling sites.

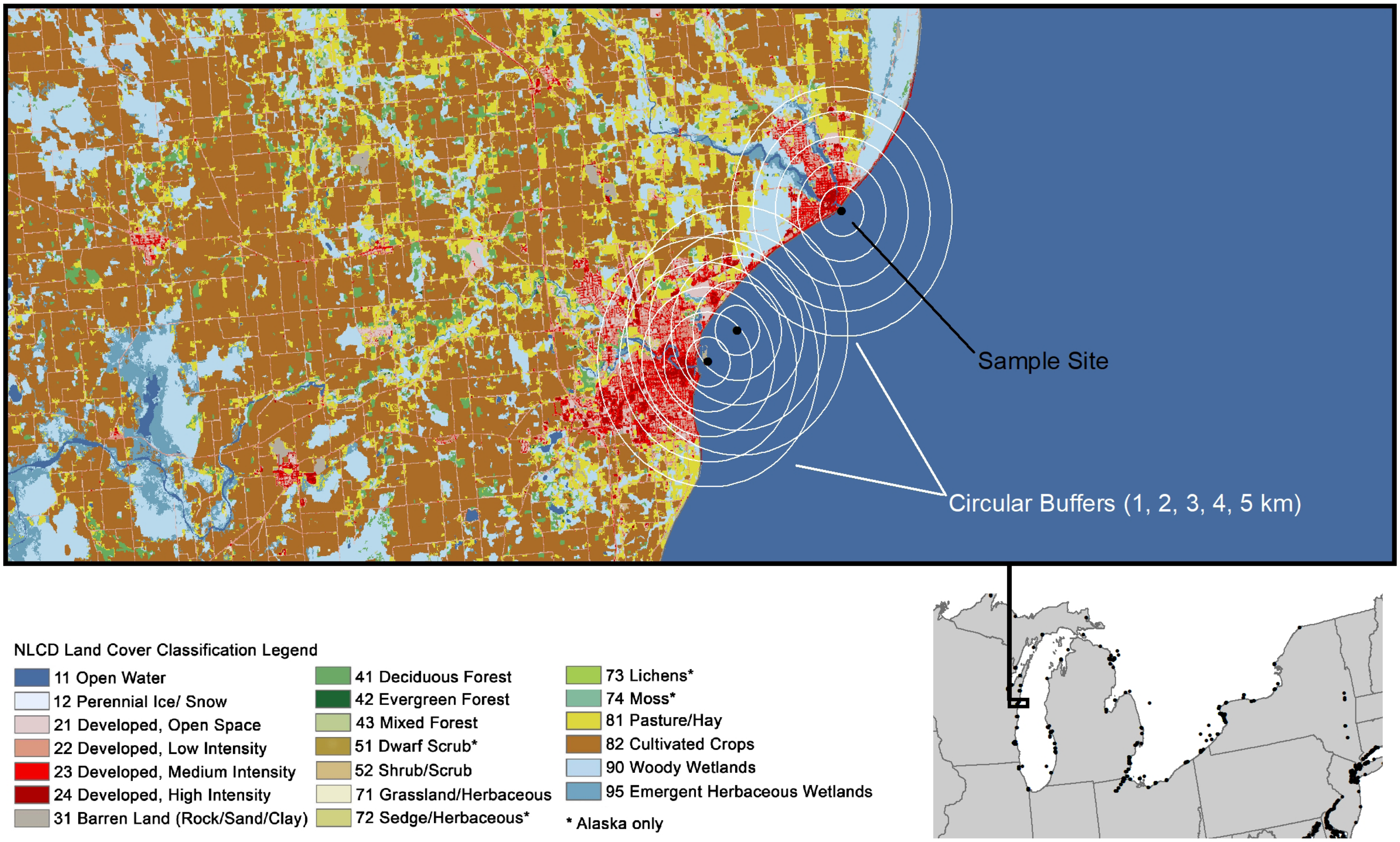


Figure 2. Example of hydrologic unit and buffers used to characterize land use at each NS&T sediment site.

Dependent Variable: PAH levels

Methods for collecting sediment samples for testing followed those described in Apeti et al., 2012. Briefly, the primary method of collection for sediment uses a Ponar grab. Once the grab sampler was aboard the boat, an acetone rinsed spatula was used to collect the top layer of sediment (2-3 cm). The sediment was then homogenized, placed in a glass jar and set on ice for shipping to a contract chemistry laboratory. Specific methods for the analytical measurement of PAH values followed those found in Kimbrough et al. 2007. The suite of individual PAH analytes and the analytical methods applied to measuring them have changed over time to include a broader suite of individual PAHs.

In order to include data from multiple studies over decades, a core set of PAHs common to all sites were summed to represent PAH magnitude (Table 1). Sediment grain size data was not available for all of the samples. This made it impossible to distinguish between samples relatively high in sand, where PAHs are unlikely to accumulate in high concentrations due to low available surface area, and sites with extremely low concentrations due to lower PAH inputs. As a result, samples with summed PAH values in the lowest 1% of all PAH concentrations were removed from the study analysis in order to decrease uncertainty due to measurement error.

Table 1. Parent compounds used for the total PAH concentration.

Acenaphthene	Benzo[a]pyrene	Benzo[g,h,i]perylene	Fluorene
Acenaphthylene	Benzo[e]pyrene	Chrysene	Indeno[1,2,3-c,d]pyrene
Anthracene	Benzo[b]fluoranthene	Dibenzo[a,h]anthracene	Phenanthrene
Benz[a]anthracene	Benzo[k]fluoranthene	Fluoranthene	Pyrene

Independent Variables

PAHs are released to the environment during use and combustion of fossil fuels, both from natural and anthropogenic sources, and are highly persistent in the environment especially in organic sediments in rivers, lakes, and estuaries. In the US, fossil fuel combustion is a primary source of PAHs, and it is considered a nonpoint source pollutant because it accumulates through multiple diffuse sources across the landscape.

With these dynamics in mind and examples from other PAH models, we determined the following list of environmental variables (factors) to test in our model: impervious surface, land use/land cover, boat ramps and marinas, population, parking lot cover, road cover, petroleum industry locations, wastewater treatment facilities, and basic demographics. Definitions and methods for spatial interpretation of each of these variables are presented here; readers should refer to the original source for interactive maps. We utilized the latest available data release for each variable.

Land use/land cover data (Figure 3) came from the Multi-Resolution Land Characteristics (MRLC) consortium in the form of the National Land Cover Database (NLCD; https://www.mrlc.gov/). This NLCD data is derived from Landsat imagery and ancillary geospatial datasets, and utilizes a multi-source integrated training decision-tree to establish land use cover classifications (Yang et al, 2018) . These land use classifications are based on a modified Anderson classification approach with a spatial resolution for the nation at 30 m. The Anderson classification system was initially designed to provide a uniform and standardized classification of remotely sensed data into land use cover types for federal agencies (a full description of this approach can be found at https://pubs.usgs. gov/pp/0964/ report.pdf; Anderson, 1977). For this analysis, we used the 2016 data which covers the contiguous US, but data is not yet available in the same resolution for Alaska, Hawaii, and Puerto Rico, hence our decision to model only the contiguous US. Percent cover of each land use type was calculated for each analytical buffer.

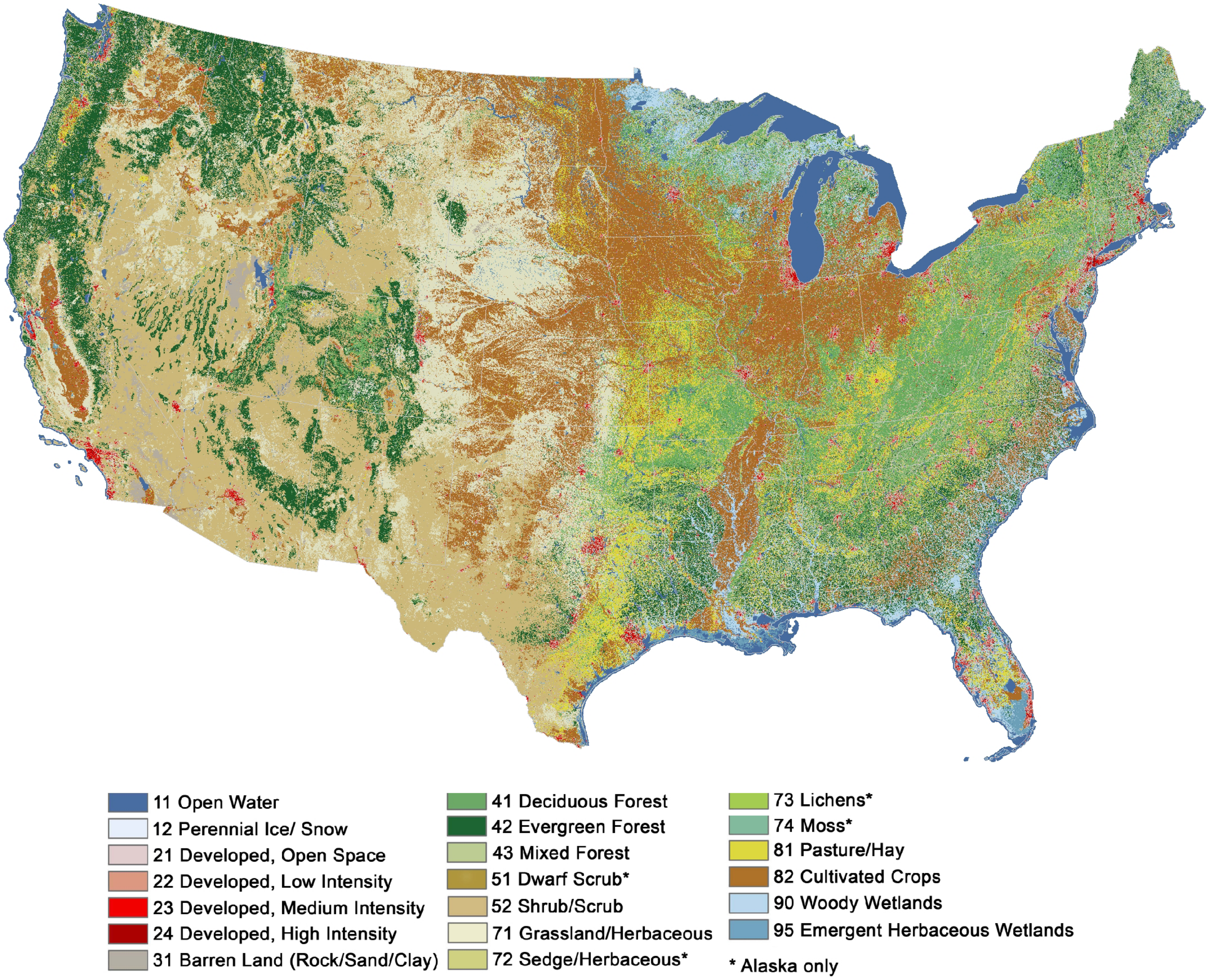


Figure 3. The USGS Land Use/Land Cover data set (2016) used for independent variables at each PAH sediment site.

Data on impervious surfaces (Figure 4) for this analysis were also derived from the MRLC (https://www.mrlc.gov/). This 30-meter pixel resolution data sets includes impervious surfaces as a percentage of developed land covering the entire contiguous United States. The percentage of impervious surfaces were calculated for Euclidean buffer distances, as well as Hydrologic Unit Code (HUC) polygons using an area weighted approach.

Boat ramps and marina locations were derived from the Office of Response and Restoration Environmental Sensitivity Index (ESI) socioeconomic layer (available https://response.restoration.noaa.gov/resources/environmental-sensitivity-indexesi-maps). Point locations for boat ramps and marinas were selected for each state ESI, merged, and then summarized for each study geography with the Summarize Within tool in ArcGIS into a point count, and normalized to a per area basis.

Population data came from the Oak Ridge National Laboratories Land Scan global population data set (https://landscan.ornl.gov/) (Figure 5). At approximately 1 km (30″ X 30″) spatial resolution, it represented an ambient population (average over 24 hours) distribution. Here we use the 2018 data release. LandScan consists of census data converted from summaries by administrative boundaries and combined with primary geospatial data into raster grid cells. Population distribution from this data is a combination of locally adaptive models that are tailored to match the data conditions and geographical nature of each individual country and region.

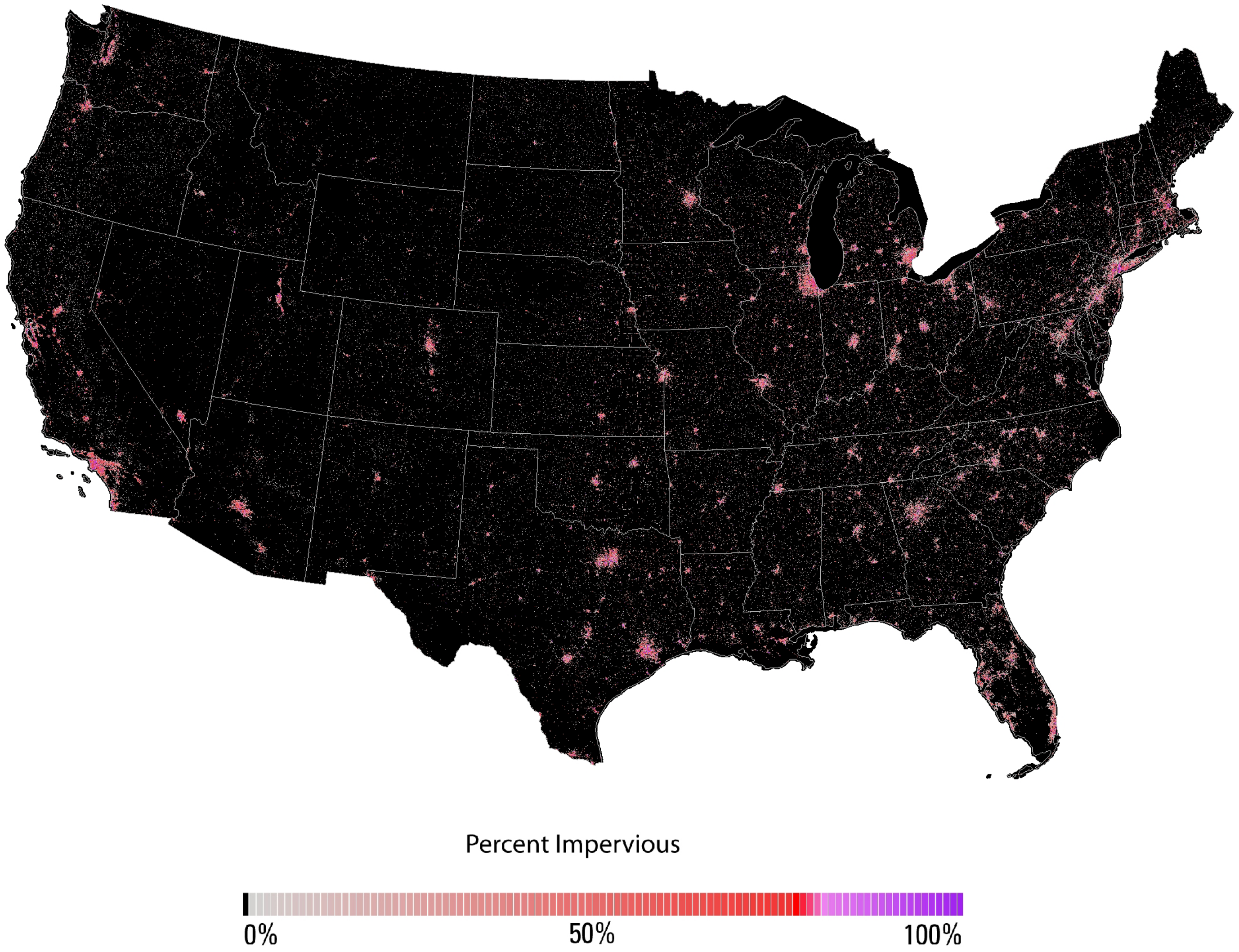


Figure 4. Percent impervious surfaces used in this study as an independent variable for each PAH sediment site.

For this analysis we used daytime and nighttime population estimates aggregated to buffer distances and HUCs around sampling sites. Daytime and nighttime population modeled estimates were created using census data, input data from the BLS, and InfoUSA databases to estimate worker distributions and flows as well as school children and business travel patterns, both during working daytime hours and evening hours.

Parking lot cover data came from the USGS Wall- to-wall Anthropogenic land-use Trends (NWALT) database (https://www.sciencebase.gov/catalog/item/5c0ea593e4b0c53ecb2af59f) from the most recent data release, 2012. The data has a 60 m resolution, with each pixel attribute representing the estimated percentage of the pixel covered with parking lot. The average percent cover constituting parking lots within each geography was calculated using the Summarize Within tool in ArcGIS (raster version).

Road length was calculated for each geography from the Census Bureau Tiger/Line files for roads (https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-geodatabase-file.html). Each type of road was treated the same (i.e. no differentiation of freeways from neighborhood roads), and total miles of road within each study geography was calculated using the Summarize Within tool in ArcGIS, then normalized to a length per area density measure.

Petroleum industry sources were derived from the Homeland Infrastructure Foundation-Level Data portal maintained by Department of Homeland Security. Points for each of the following categories were included: petroleum terminals, petroleum ports, oil refineries, oil and natural gas wells, oil and natural gas platforms, natural gas storage facilities, natural gas processing plants, natural gas market hubs, and natural gas import/export sites. A count of petroleum industry sites within each study geography was calculated using the Summarize Within tool in ArcGIS and normalized to a per area basis.

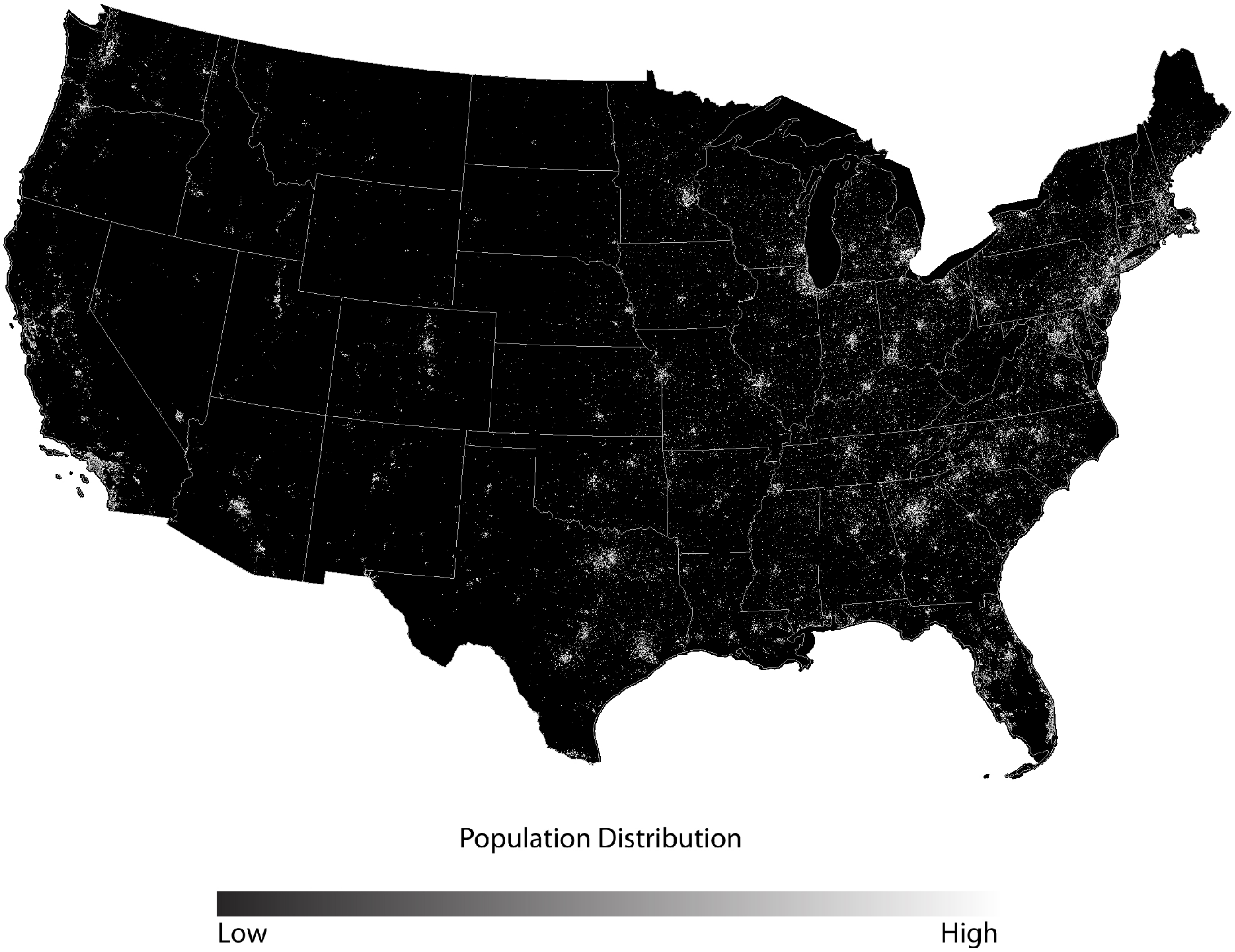


Figure 5. The LandScan population distribution coverage used as an independent variable for each PAH sediment site.

For demographics, each study site was assigned to the nearest Census block group using the Near tool in ArcGIS. From there, data were joined on total population (Census 2010), median age (Census 2010), percent female (calculated from gender counts in Census 2010), number of elderly (65 and over) citizens (2019, ESRI demographics), and median income (2019, ESRI demographics).

Modeling

For the modeling of PAHs in sediment on a national scale, spatial models were considered to account for regional autocorrelation of both the dependent and independent variables in the model using spatial regression and Random Forest modeling tools within ArcGIS Pro 2.3. Data cleaning involved clipping the data set geographically to the contiguous US, as the territories, Hawaii, and Alaska have different or nonexistent data sets for many of the explanatory variables.

Model specification occurred by assessing the fit of non-spatial models. For each sampling site, the independent variables were derived from a set of eight geographies of increasing area around the sample site (except for Census-based data, for which demographics from the nearest block group were used). The first step was to determine which of these geographies best modeled PAH levels by lowest AIC/highest r-squared value in the Exploratory Regression tool. In order to accommodate the very skewed distribution of PAH concentrations (see results), the cube root of PAH values were used as the dependent variable. Although there were significant independent variables, most of these variables explained a low amount of variability in the PAH data. The three significant variables with the highest r^2 values were road density (adjusted $r^2 = 0.27$), impervious surface (adjusted $r^2 = 0.28$), and high intensity development in a 3 km buffer (adjusted $r^2 = 0.27$) (Table 2). Note this exercise was to choose the most related buffer distance; highly correlated variables were removed in the spatial models.

Once there was one specific measure for each category of independent variable, a Poisson regression was used to further specify the model (Generalized Linear Regression tool), beginning with a model including all theorized independent variables: median age, median household income, percent female, percent white, ramp density (1 km), parking lot cover (3 km), road density (5 km), petroleum facility density (1 km), high density development (3 km), medium density development (5 km), low density development (5 km), developed open space (5 km), impervious surface (5 km), daytime population

Table 2. Comparison of r^2 values for various sized spatial buffers and independent variables from regression analysis of PAH concentrations.

Independent Variable	Buffer distance (km)	Adjusted r^2
Impervious surface	5	0.28
Road density	5	0.27
High intensity development	3	0.27
Medium intensity development	5	0.25
Parking cover	3	0.2
Nighttime population	5	0.2
Low intensity development	5	0.14
Daytime population	5	0.13
Developed open space	5	0.04
Petroleum facility density	1	0.03
Boat ramp density	1	0.02

(5 km), and nighttime population (5 km). Poisson regression was used because of the skewed distribution of PAH values, with a vast majority of sites having low PAH concentrations. The initial model with all dependent variables included explained 42.6% of the variance (p = 0, AIC = 37775857) with all variables significant (p = 0 for all), but had eight high VIF scores (above 7.5). Impervious surface (127.1) was found to be multicollinear with road density (VIF = 20.6), high density development (VIF = 11.3), medium density development (VIF = 36.1), low density development (VIF = 13.2), and nighttime population (VIF = 10.2), so it was dropped in the next round of the model.

The model, after removing impervious surface, explained 41.8% of the variance (p = 0). All independent variables were significant (p = 0) and had resulted in four variables with a high VIF (≥ 7.5).

Stepwise, the variable with the highest VIF was removed and the model re-run. Road density and nighttime population were removed accordingly and the final model explained 36.5% of the variance (p = 0), all independent variables were significant, and no variables with a VIF above 7.5 (Table 3). The model is spatially auto-correlated and is used to specify a Geographically Weighted Regression (GWR).

For comparison, a Forest-based Classification and Regression model was also run using the same explanatory variables as the GWR with high VIF variables removed with 100 trees and 25% of data withheld for validation (Breiman 2001). The Random Forest is an ensemble model that uses multiple decision trees with different subsets of data. The use of multiple small decision trees results in a model that does not need to be pruned.

Table 3. Statistics for the final regression model.

Variable	Coefficient	StdError	z-Statistic	Probability	VIF
Median Age	0.049	0.000031	1583	<0.00001	1.73
Median Household Income	-0.000003	0	-471	<0.00001	1.46
Percent Female	-4.31	0.0035	-1230	<0.00001	1.13
Percent White	-1.68	0.0035	-1291	<0.00001	1.61
Ramp Density	-0.23	0.00038	-606	<0.00001	1.15
Mean Parking	0.00056	0.000001	678	<0.00001	2.60
Land Use - High Density Development	1.63	0.0038	432	<0.00001	5.72
Land Use - Open Developed	0.3	0.0079	38	<0.00001	2.62
Land Use - Low Density Development	3.22	0.0057	564	<0.00001	4.09
Land Use - Medium Density Development	6.52	0.0043	1535	<0.00001	5.42
Daytime Population	0.000012	0	316	<0.00001	1.87
Petrochemical Facility Density	0.46	0.00089	512	<0.00001	1.10
Intercept	7.59	0.0019	4059	<0.00001	n/a

Results

There were 3722 individual records of PAH measurements, with 56 of the records missing at least one of the independent variables due to the location of sampling sites. As a result, 3666 sites were used in the national sediment model, with a total of 15 selected independent variables that are theoretically connected to PAH levels found in the coastal environment. The continental national averages and variances for each variable are summarized in Table 4, and the clusters of PAH values are shown in Figure 6. Regression analysis between PAH concentration and selected independent variables appear in Figure 7.

Due the uneven distribution of sampling sites at the national scale and spatial autocorrelation of both PAH levels and independent variables, several modeling approaches were tested to see which performed best in this context. These comparisons were meant to answer three primary modeling questions:

1. How does a regression-based approach compare with machine learning techniques, specifically Random Forest when including human dimensions dependent variables?

2. Does a categorical data set improve performance over continuous data?

3. Does a spatial or tabular conceptualization of dependent variables perform better?

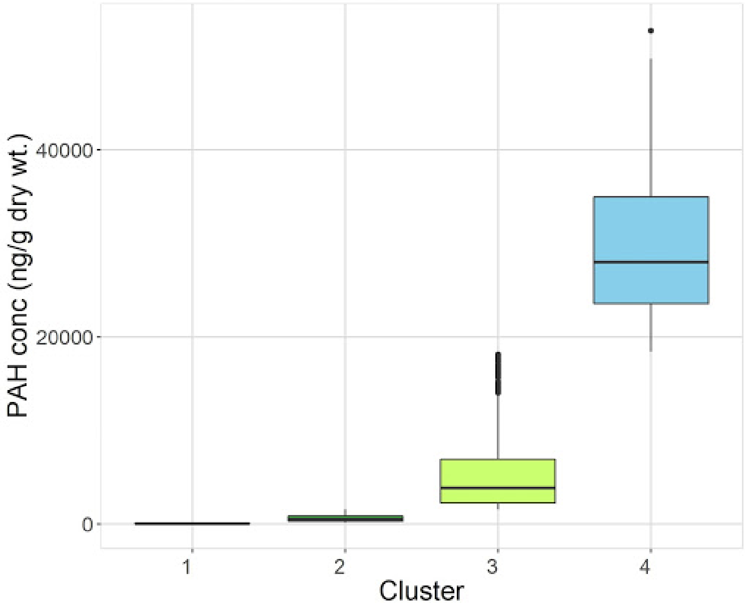


Figure 6. Box plot characterizing PAH clusters.

Table 4. Variables included in modeling runs with summary statistics.

Variable	Mean	Standard Deviation	Range	Units
PAH concentration	3674	33642	0 - 1219089	ng/g dry weight
Median Age	40.09	8.36	13.2 - 79.7	Years
Median Household Income	83558	41686	10331 - 200001	Dollars
Percent Female	0.51	0.05	0.03 - 0.88	proportion
Percent White	0.402	0.26	0.0012 - 1	proportion
Ramp Density (1 km)	0.21	0.66	0 - 8.6	Ramps/sq.km.
Mean Parking (3 km)	265	361	0 - 2013	Percent x 100
Petroleum Facility Density (1 km)	0.05	0.21	0 - 3.5	Count
Road Density (5 km)	3.83	3.46	0 - 18.4	km/km^2
High Intensity Development (3 km)	0.06	0.1	0 - 0.58	proportion
Medium Intensity Development (5 km)	0.08	0.09	0 - 0.43	proportion
Low Intensity Development (5 km)	0.07	0.07	0 - 0.39	proportion
Developed Open Space (5 km)	0.05	0.05	0 - 0.38	proportion
Daytime Population (5 km)	1875	4434	0 - 74227	People
Nighttime Population (5 km)	1277	2308	0 - 32400	People

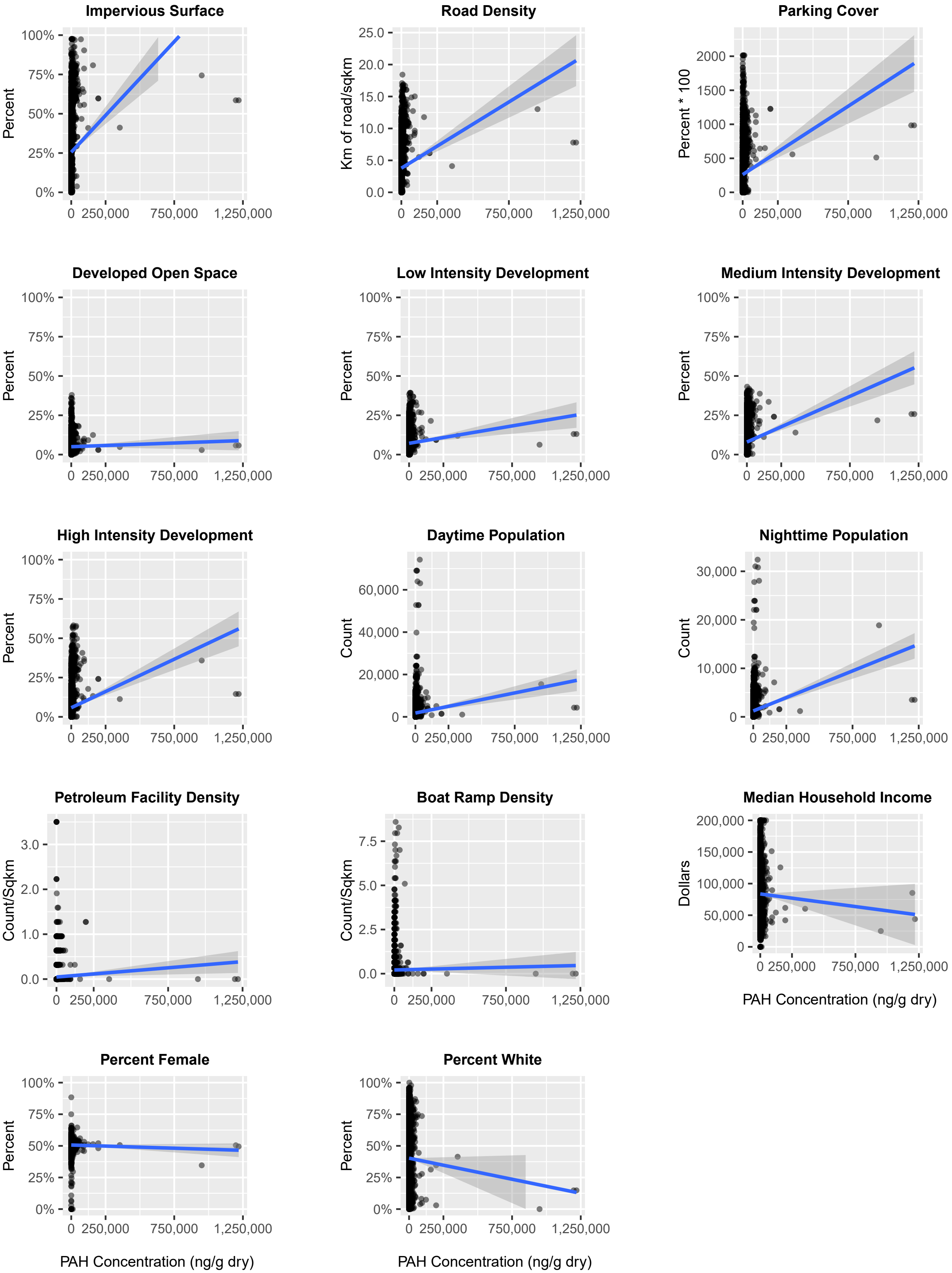


Figure 7. Regression results for relationships between PAH concentration and independent variables.

Geographically Weighted Regression vs. Random Forest

The GWR model explained 60.0% of the variance, which is a 43.7% increase over the non-spatial Poisson regression (Table 5). The GWR, however, would not run because of local multicollinearity when including all the variables specified through standard regression; this should be considered a major shortfall of the GWR approach even though it performs better than the non-spatial regression model. Therefore, those variables (all Census-based variables) that depicted high levels of multicollinearity were dropped in order to complete the model. The distance band required to run the GWR is 460 km, which is similar to the distance between Richmond, VA, and New York City, NY, making the neighborhood of analysis for each point a broad geographic region. This makes sense given that PAH dynamics vary by region based on attributes such as development patterns, coal tar sealant sourcing for roads, and population density. The regions have different correlations with each of the dependent variables, an example of which is shown in Figure 8 depicting the coefficients associated with high density development within a 3 km buffer. An increase in impervious surface in the Mid-Atlantic or Gulf Coast, predicted a bigger increase in PAHs than it would for sampled areas such as the West or Great Lakes Coasts.

Table 5: Model results for Geographically Weighted Regression. Final model: PAH level = mean parking + land use + population.

Distance Band	459654
Deviance explained by the global model (non-spatial)	0.2889
Deviance explained by the local model	0.5996
Deviance explained by the local model vs. global model	0.437
AIC	2.62E+07
Sigma-squared	1.21E+09
Sigma-squared MLE	1.20E+09
Effective degrees of freedom	3532

In terms of model performance, the predicted values versus actual values of each site are shown in Figure 9, with an r^2 value of 0.23. Given the highly skewed distribution of the PAH data, including a few very high sediment readings that highly influence the regression, this poor performance indicates that a geographically weighted regression (GWR) approach is not adequate for the data.

A Random Forest Model using the same dependent variables as the GWR (determined through a model specification process and constraints due to local multicollinearity) yielded a model that explained 47.5% of variation when using 100 trees. The list of variable importance is in Table 6.

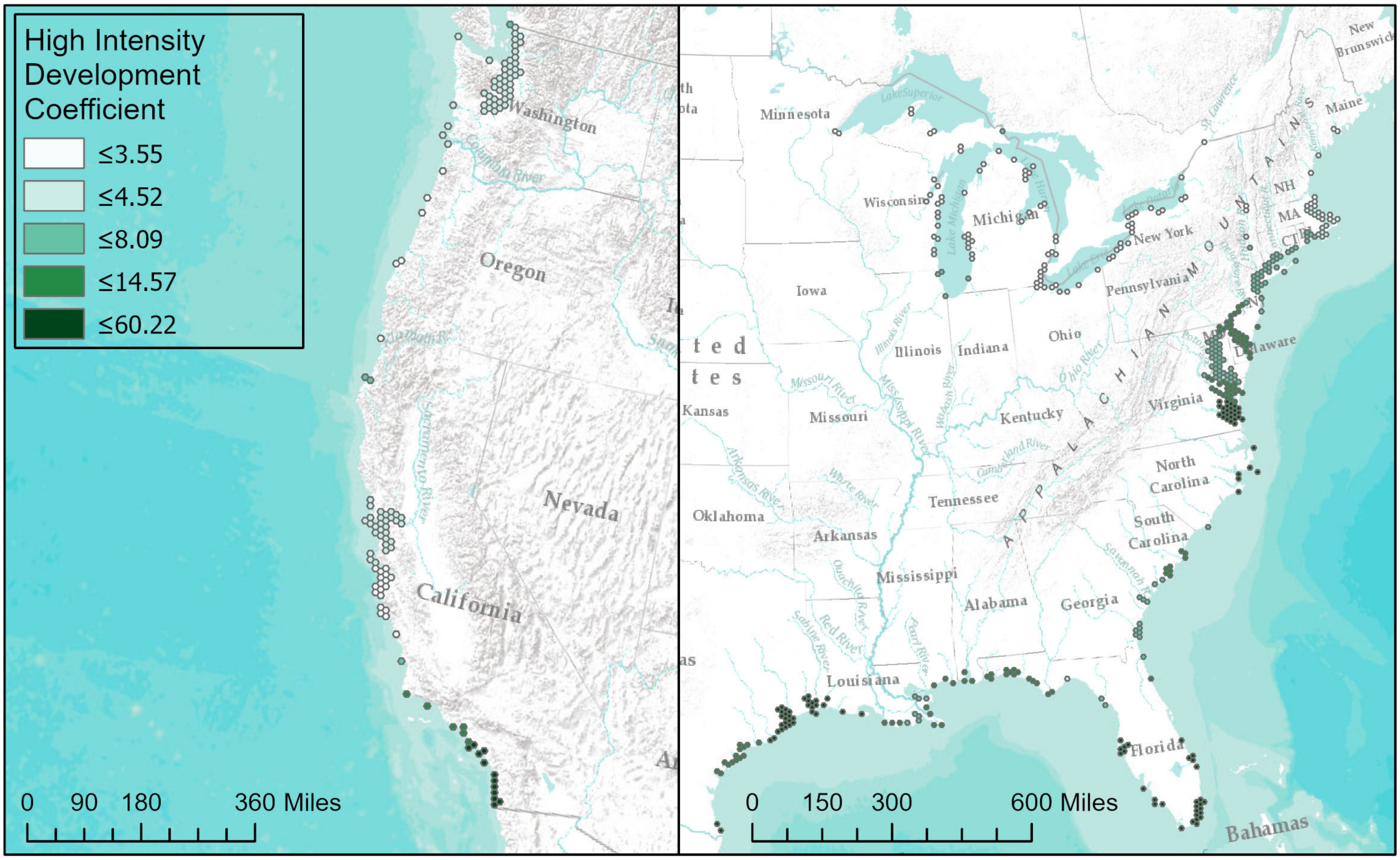


Figure 8. Coefficients for high density development in the GWR model.

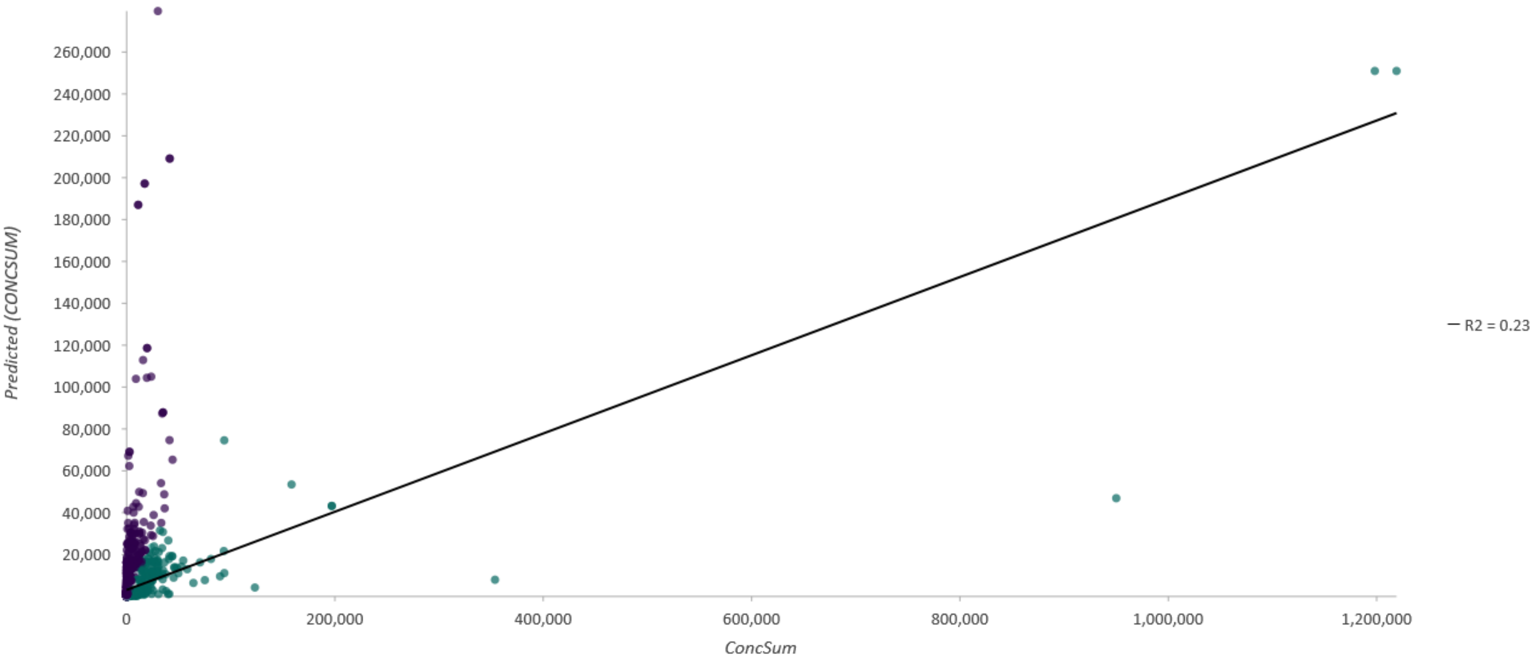


Figure 9. PAH values (x-axis) compared to predicted values from Geographically Weighted Regression (y-axis).

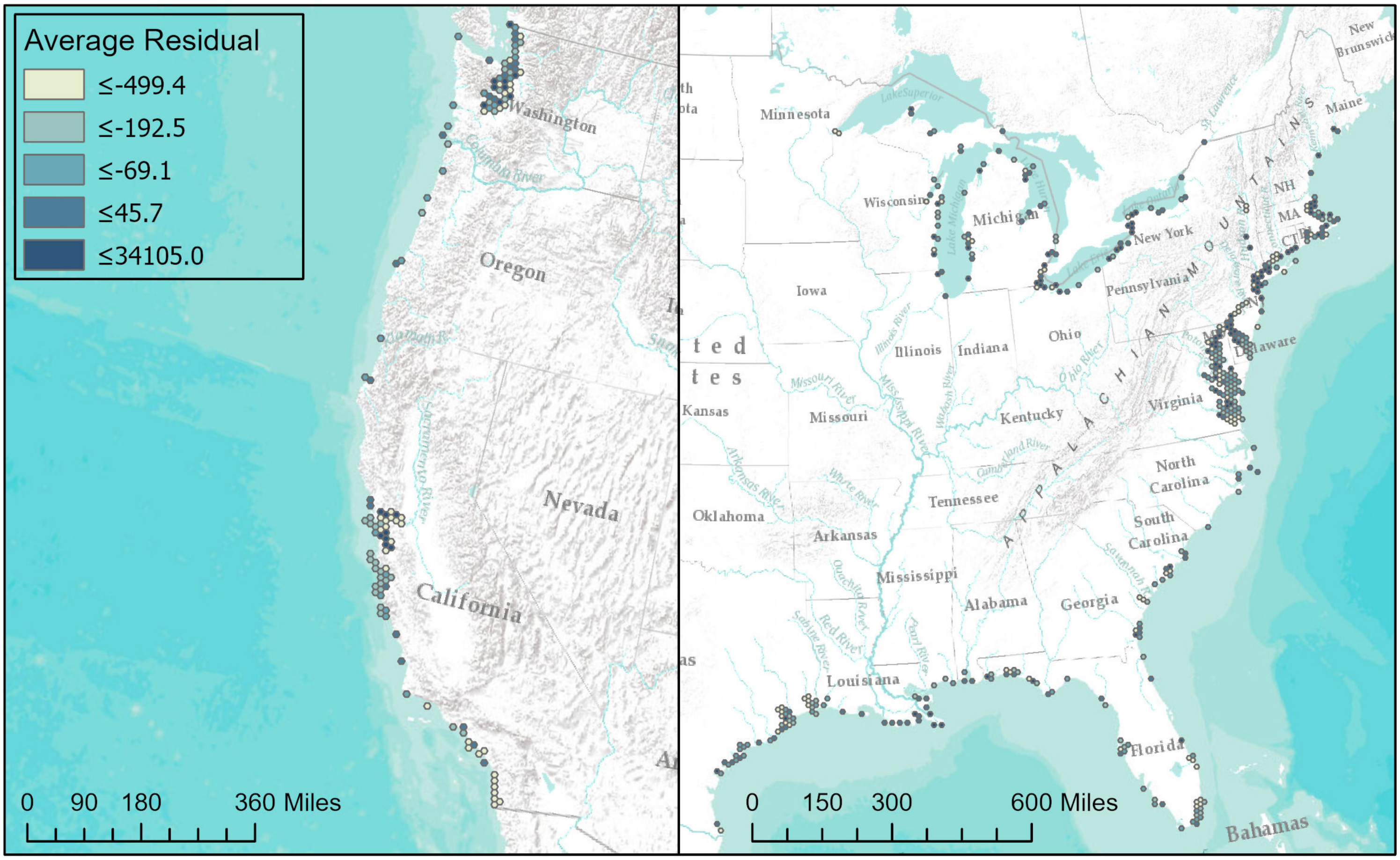


Figure 10. Random Forest average residual values.

Table 6. Variable Importance from Random Forest model using same variables as those in the Geographically Weighted Regression model.

Variable	Importance
Land Use - High Intensity Development	0.24
Mean Parking Cover	0.23
Land Use - Open Developed	0.16
Land Use - Medium Intensity Development	0.14
Land Use - Low Intensity Development	0.12
Daytime Population	0.11

There was little variation in how well the Random Forest model performed over space, as shown by the map of model residuals in Figure 10.

In terms of model performance, the predicted versus actual values of PAH concentrations are as follows for points included in the training, with an $r^2 = 0.94$. For testing values withheld (25%), performance was highly variable, ranging from 0.04 to 0.91 over 10 repetitions.

Random Forest: Categorical vs. Continuous Data
Since Random Forest can incorporate all of the variables from the data reduction process into the model, this comparison will return to the full set of independent variables specified from the Generalized Linear Regression: parking cover, daytime population, high density development, medium density development, low density development, developed open space, percent white, median household income, median age, density of petrochemical facilities, boat ramp density, and percent female. The continuous data model explains 42.3% of the variance with 100 trees. The regression diagnostic show predicted versus actual values of training data with an $r^2 = 0.95$ ($p = 0$) and predicted versus actual values of validation data ranging from an r^2 of 0.02 to 0.90 ($p = 0$). The variables are ranked in Table 7.

Using a categorical version of Random Forest yielded very different results. PAH values were clustered using Mclust, a Bayesian-based tool (Fraley & Raftery, 1999), into 4 clusters with 1191, 1533, 882, and 116 sites. The cluster number was input as the dependent variable for the Random Forest based on the classification tool, run with compensation for sparse categories. The resulting model had a mean squared error (MSE) of 42.2 with 100 trees. As shown in Table 8, the variable importance is more well spread across the different independent variables, with accuracy of the predicted versus actual values of validation data ranging from r^2 of 0.62 to 0.89.

As expected, given the variability in both measurement capacity and independent variables at low PAH levels,

Table 7. Variable importance for the continuous version of the Random Forest model.

Variable	Importance
Percent White	0.16
Mean Parking Cover	0.15
Land Use - High Intensity Development	0.15
Land Use - Low Intensity Development	0.12
Daytime Population	0.11
Land Use - Medium Intensity Development	0.09
Median Age	0.08
Percent Female	0.04
Median Household Income	0.04
Land Use - Open Developed	0.04
Petrochemical Facility Density	0.00
Boat Ramp Density	0.00

Table 8. Variable importance for the categorical version of the Random Forest model.

Variable	Importance
Petrochemical Facility Density	0.11
Boat Ramp Density	0.11
Land Use - High Intensity Development	0.11
Land Use - Medium Intensity Development	0.10
Daytime Population	0.09
Mean Parking Cover	0.09
Land Use - Low Intensity Development	0.08
Land Use - Open Developed	0.08
Percent Female	0.06
Median Age	0.06
Median Household Income	0.06
Percent White	0.05

Table 9. Classification of PAH values into clusters based on actual values and values predicted by the categorical Random Forest model.

		Actuals			
	Cluster	1	2	3	4
Predicted	1	147	104	31	16
	2	66	197	94	27
	3	6	52	129	34
	4	0	1	12	16

performance was best in cluster 4 (the highest PAH values), with predicted versus actual values for training and validation data at 0.92 and 0.89, respectively. The confusion matrix is shown in Table 9.

Overall, the continuous model explained about the same variance in the data, but was far less consistent in predictive values when evaluated with validation data. Repeated runs of the model are possible in order to identify versions of the model with best predictive power.

The categorical model offered more consistent performance when evaluated with validation data and performed very well when applied to the highest PAH values. Therefore, depending on the application of the model and where one needs to prioritize performance and consistency, either version of the model may be preferable.

Spatial vs. Tabular Conceptualization of Dependent Variables

Dependent variable inputs for the spatial version used in the Random Forest tool can also be the raw data files and represent a different way of conceptualizing the dependent variables. The tool calculates the distance to points and lines and uses the underlying individual raster data, leaving the model to be more spatially precise, but less integrative of the nearby land use. The resulting model had a MSE of 66 with 100 trees. Unlike the tabular conception of the independent variables, the performance of the model on training data was similar to performance on test data (0.57 - 0.71 across all four clusters, highest in cluster 1). Using the same independent variables as in the cluster model above. Table 10 shows the importance values in a spatial model, and Table 11 the confusion matrix. Given the importance of performance for cluster 4 (the highest PAH values), the spatial version of this model is not preferred.

Table 10. Random Forest spatial model variable importance.

Variable	Importance %
Daytime Population	0.15
Parking Cover	0.15
Land Use Land Cover	0.15
Boat Ramps	0.10
Petroleum Facilities	0.10
Median Age	0.09
Median Household Income	0.09
Percent Female	0.09
Percent White	0.09

Table 11. Classification of PAH Values into clusters based on actual values and values predicted by a Random Forest model using select independent variables.

		Actuals			
	Cluster	1	2	3	4
Predicted	1	82	15	78	101
	2	59	44	101	178
	3	23	8	93	97
	4	2	3	6	18

Data Reduction Strategies

Random Forest is widely used due to its inherent protections against overfitting. However, some debate exists whether one should perform data reduction before running Random Forest in order to improve performance, even if both options are statistically valid. The results thus far have all relied upon some data reduction to remove multicollinear independent variables. For comparison, Table 12 shows the clustered, tabular Random Forest model with all possible variables included; MSE error is 37.5 at 100 trees.

Table 12. Importance of variables from Random Forest model with clustered PAH values and all potential independent variables

Variable	Importance %
Land Use - High Intensity Development	0.09
Land Use - Medium Intensity Development	0.09
Petroleum Facility Density	0.08
Mean Parking Cover	0.08
Boat Ramp Density	0.08
Impervious Surface	0.08
Daytime Population	0.07
Nighttime Population	0.07
Land Use - Low Intensity Development	0.07
Median Age	0.05
Median Household Income	0.05
Land Use - Open Developed	0.05
Percent White	0.04
Percent Female	0.04

Table 13. Classification of PAH values into clusters based on actual values and values predicted using Random Forest with all potential independent variables.

		Actuals			
	Cluster	1	2	3	4
Predicted	1	166	66	42	11
	2	84	166	104	29
	3	15	34	133	39
	4	0	1	10	18

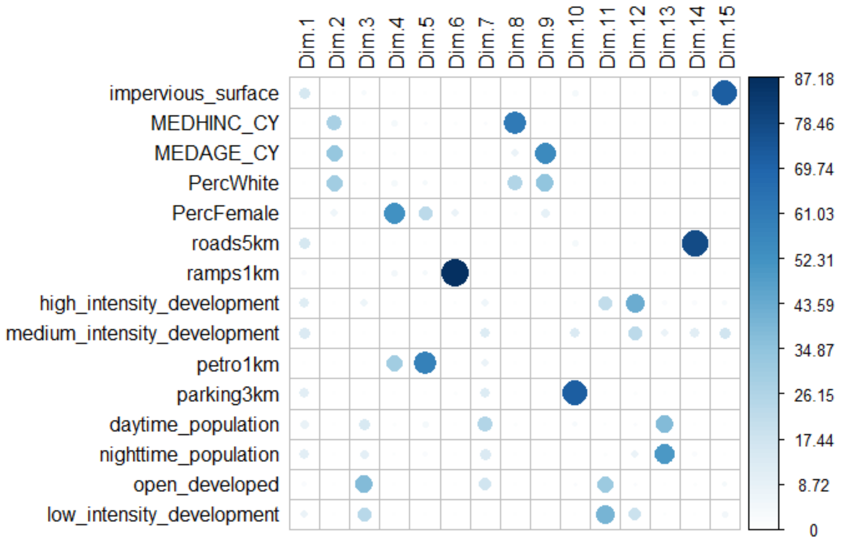


Figure 11. Contribution of each independent variable to PCA components.

Table14. Variance explained by each of the PCA components

Dimension	Eigenvalue	Variance (%)
1	6.07	40.5
2	2.08	13.8
3	1.71	11.4
4	1.02	6.8
5	0.98	6.5
6	0.88	5.8
7	0.62	4.1
8	0.51	3.4
9	0.4	2.7
10	0.34	2.2
11	0.19	1.3
12	0.1	0.7
13	0.06	0.4
14	0.04	0.3
15	0.01	0.04

Table 15. Variable importance for Random Forest run on PCA components.

Component	Importance (%)
6	9
7	8
13	8
10	7
15	7
8	7
11	7
3	7
5	6
2	6
12	6
14	6
9	6
4	6
1	5

Table 16. Classification of PAH sites into clusters based on actual PAH values and PAH values predicted based on a Random FOrest model with PCA components.

		Actuals			
	Cluster	1	2	3	4
Predicted	1	146	80	49	23
	2	99	158	86	41
	3	18	56	98	49
	4	3	1	8	17

Results depicted in the confusion matrix (Table 13) looks similar to the regression- reduced model, however accuracy for validation data was slightly better for the highest PAH cluster (0.92 versus 0.89) but slightly different for the lowest three clusters (between 0.67 and 0.78, versus 0.62 to 0.76).

Another type of data reduction that is more comprehensive in its statistical approach is a Principal Components Analysis (PCA). Since many of the independent variables in this study are both theoretically and mathematically correlated, data reduction in the form of a PCA is a common approach for removing the effects of multicollinearity and an added protection against overfitting (even though Random Forest protects against overfitting fairly well on its own). A PCA of the independent variables yields the 15 dimensions depicted in Figure 11.

The first four components all have eigenvalues greater than > 1 and together explain 75% of the variance in the independent variable data set (see Table 14). These first four components together include at least some portion of each of original independent variables, suggesting the value of including all variables in this statistical analysis.

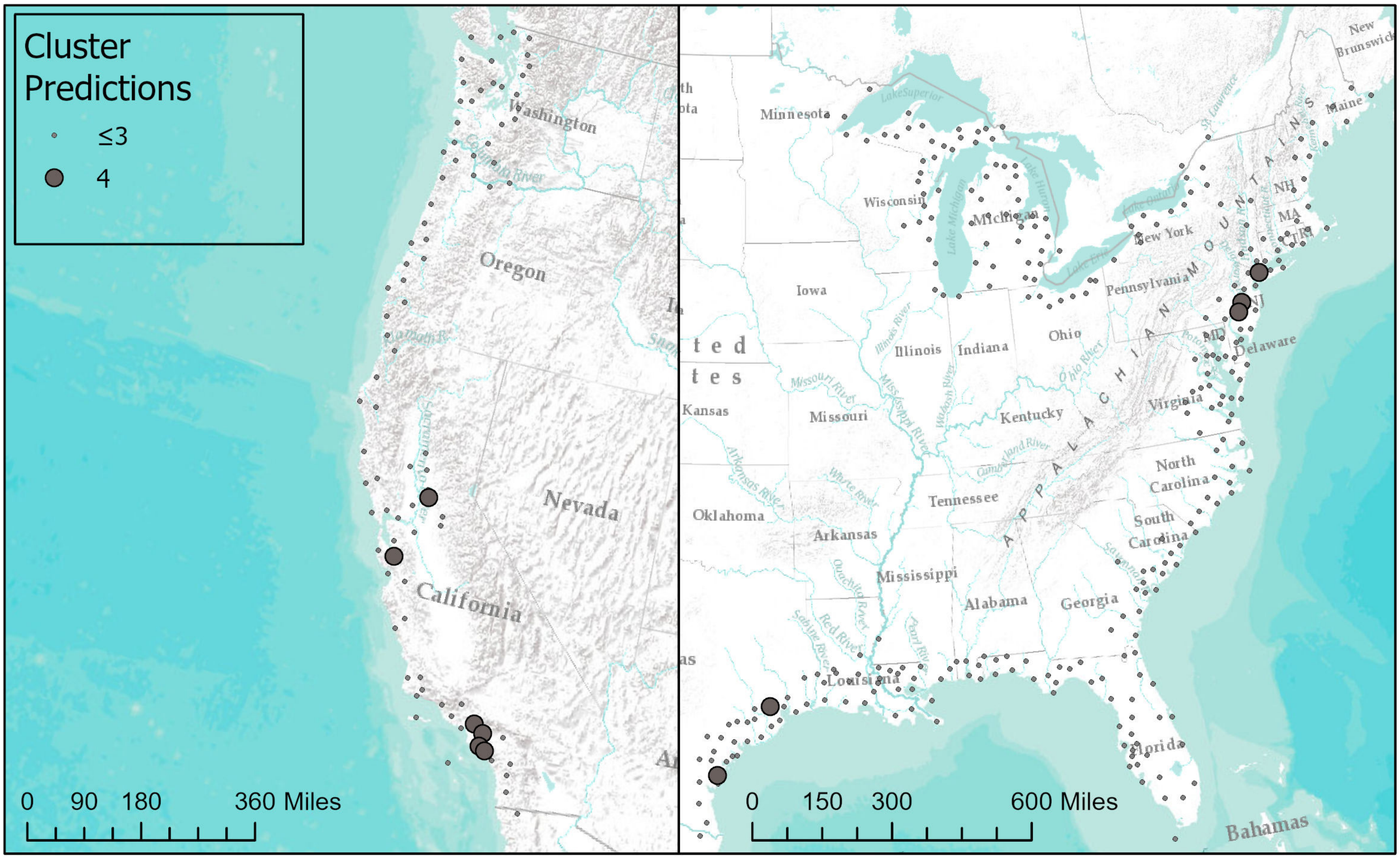


Figure 12. Predicted high PAH sediment levels as modeled for the centroid of each watershed. Cluster 4 depicts high levels.

However, the requirement for all variables to be populated suggest that it is necessary to reduce the complexity of the data inputs.

In this case, a PCA in which all 15 resulting components are kept yields the variable importance as shown in table 15, with 38.9 MSE at 100 trees and 0.61 - 0.87 accuracy for each of the four clusters (highest for the highest PAH cluster). The confusion matrix is in Table 16. The performance is very similar in performance to the unreduced model, but more difficult to interpret. Both the PCA and regression dimension reduction methods yielded similar results to those calculated from the Random Forest models. This introduces several considerations in model specification such as prioritizing which cluster performs the best and how easily interpretable the results need to be. Given the importance of model performance for the highest cluster of PAH levels, the non-reduced model will be used for the predictive map in the next section.

Predictive map

Using the best performing model structure (no data reduction, categorical chemical data, and using 5 km buffers to determine a neighborhood around each sampling site) and existing sampling sites as a training data set, we used Random Forest predictive capabilities to predict the level of PAH concentration at the centroid of each watershed (or HUC8). Random Forest was run 50 times for validation, with the highest performing iteration used for predictions. The final model had an MSE of 36.9 and accuracy of 0.90 for the highest cluster of PAHs. Figure 12 depicts the centroids that are expected to be in cluster 4 (high PAH levels) highlighted in large gray circles; the centroids predicted to have less PAHs are shown in small circles. The predicted high PAH levels are primarily (but not entirely) surrounded by urban land use and present primarily in Southern California and seaside New York, both of which boast high commercial vessel port activity. In Southern California, this prediction is also reflected in regional monitoring efforts (Du et al 2020).

Discussion

Of the modeling options conducted in this study, categorical Random Forest produced the most reliable and accurate results. When evaluating the Random Forest models, choosing between continuous and categorical versions impacted model reliability. The continuous variables produced a more reliable set of importance rankings but predicted PAH values with far less accuracy, especially for sites with PAH values in the highest cluster. When sites with high PAH levels are prioritized, it is suggested that the categorical model would be better suited. However, if the research question is to understand the dynamics of each independent variable individually, the decision may fall toward continuous modeling without using PCA for data reduction. Furthermore, if one is really interested in just one of the independent variables, the GWR might have something more to offer - many of the demographic variables were locally collinear. For example, road cover in a region similarly shows regional trends that correspond with regional road sealant sources.

This modeling effort depended entirely on secondary data by design, as it was meant to help prioritize future chemical monitoring efforts and deliver some context to the spatial dynamics observed in the nationwide PAH monitoring effort. As a result, the independent variables were a best estimate of the underlying demographic and land use factors that influence PAH contamination, but were at times inexact and not comprehensive. For example, studies in China link particular types of business and industry to PAH contamination (Huang et al 2017), but lack of a similar comprehensive nationwide business database limited the ability for this effort to focus on business sources. Wastewater treatment outflows are also known sources of many contaminants, including PAHs, but georeferenced wastewater facilities generally locate the facility, not the outfall, and these can be a significant distance from one another. There is also no national database of wastewater facilities, and so would have required merging different data sets or scaling down to smaller models with more complete regional collections. In addition, as it pertains to chemical data site locations, the priorities for sampling have changed over time and there is not equal representation across the nation or the types of land use surrounding the sites. For example, the historic focus on industrial contaminants means there are few sites surrounded by agricultural land uses.

For other environmental variables assessed in this study, such as population, several data options addressed these attributes, and we had to choose between the available options based on attributes of the data and spatial resolution. To measure land use and land cover data, we used the MRLC NLCD 2016 coverage because of its spatial consistency across the entire country, and the need to be consistent with similar work conducted by the USGS (MRLC 2020). LandScan was chosen as a population data set due to the potential to tease and extract daytime, and nighttime population estimates, and because of the high degree of spatial accuracy. Additionally we utilized two data sets that are easily available for both government agency and public use.

While the modeling approach using all independent variables in a PCA followed by Random Forest worked best for this particular study, this may not be the case for all data sets and contexts. Decisions concerning data reduction need to be part of structuring the theoretical approach to modeling this type of data. Random Forest is renowned for its ability to avoid overfitting and adding PCA to the workflow increases the model’s overall ability to handle large numbers of independent variables that are likely related to one another. However, even depending on mathematical approaches to avoid duplication, one must still decide which data sources to seek out according to what we know about the modeling context. In this case, PAH sources are well documented, as is the environmental fate and behavior of PAHs once introduced to the environment.

In determining which independent variables to include, we looked to other PAH modeling efforts and source tracings in order to prioritize which data were required. This is a fundamentally subjective effort, and requires deep knowledge of both available data sets and the current state of the science.

Additionally, dimension reduction prior to modeling efforts were conducted in this study was based on a concept of the “local environment” around each sampling site in the form of a buffer. We tested differently sized circular buffers and HUC system boundaries to see which captured the best relationship between the variables used in this modeling exercise. Many, if not most, attempts at this type of modeling included a water systems approach, and therefore the best relationship between the variables tends to be hydrologic boundaries. However, buffers can be logistically easier to work with and may perform better for quick-to-settle chemicals, such as PAHs where previous studies have found the immediate environment makes the most difference in sediment-based contaminant level (Uher et al., 2016). One can, in theory, include multiple conceptualizations as separate independent variables that will be grouped together via PCA if related. However, we tested this ahead of time to limit the number of interactions that might be happening at any given time. Only then did the final round of dimension reduction occur via PCA.

This modeling effort utilized PAHs because of their ubiquity and persistence in aquatic environment, but few chemical contaminants share these convenient attributes, raising the question of how well this modeling approach will apply to chemicals with different environmental behavior and presence. First, additional variables are likely needed to better match their sources. For example, a climate and/or weather variable may be needed to capture the dynamics of shorter-lived contaminants that are deposited into the coastal ecosystem following rain events. Depending on the environmental fate and transport behavior of different types of chemicals, a different size or shape of buffer may perform better and so this step in dimension reduction remains necessary. And finally, for less ubiquitous chemicals, nationwide data may be unavailable; a set of regional models may be necessary in order to remove large spatial gaps in the chemical data. In addition to limitations surrounding how to process various types of chemical data, there is likely more variability in human dimensions than we are able to capture utilizing secondary data derived from sources not designed explicitly for this purpose (such as the Census).

At the end of the modeling exercise here, it is helpful to step back and remember why the model is important: to help predict areas with high levels of PAHs that have not been consistently monitored over time and researchers, managers, and policy makers can focus future monitoring efforts. Even the best model presented here has limited accuracy for the lower level PAH clusters. Several decisions made in the modeling process obscure the direct relationships between individual independent variables and PAH level. However, the accuracy for the highest cluster is high enough to feel confident using it in future sampling plans and the effects of individual variables do not need to be teased apart since they are all easily available, nationwide data sources that can be reliably deployed in future model runs and studies.

References:

Anderson, J.R. 1977. Land use and land cover changes—A framework for monitoring. J. Res. US Geology Survey, 5, pp. 142-152.

Apeti, D.A., W.E. Johnson, K.L. Kimbrough and G.G. Lauenstein. 2012. National Status and Trends Mussel Watch Program: Field Methods 2012 Update. NOAA National Centers for Coastal Ocean Science, Center for Coastal Monitoring and Assessment. NOAA NCCOS Technical Memorandum 134. Silver Spring, MD.

Breiman, L. 2001. Random forests. Machine Learning, 45(1): pp. 5–32.

Census. 2010. https://www.census.gov/programs-surveys/decennial-census/data/datasets.2010.html. Accessed 5/2020.

Census. 2020. https://www.census.gov/programs-surveys/geography/about/faq/2010-urban-area-faq.html#par_textimage_10. Accessed: 5/2020.

Du, Bowen, Charles S. Wong, Karen McLaughlin, and Kenneth Schiff. 2020. Southern California Bight 2018 Regional Monitoring Program: Volume II. Sediment Chemistry. Southern California Coastal Water Research Project, Technical Report 1130. https://ftp.sccwrp.org/pub/download/DOCUMENTS/TechnicalReports/1130_B18Chemistry_Full-res.pdf.

EPA (U.S. Environmental Protection Agency). 2012. Toxic Contaminants in the Chesapeake Bay and its Watershed: Extent and Severity of Occurrence and Potential Biological Effects. Annapolis, MD. 175 pp.

ESRI Demogrpahic Data. 2019. https://doc.arcgis.com/en/esri-demographics/data/updated-demographics.htm

Fraley, C., and A. E. Raftery.1999. MCLUST: Software for model-based clustering, Journal of Classification, 16, pp. 297-306

Garner, T.R., J.E. Weinstein and D.M. Sanger. 2009. Polycyclic Aromatic Hydrocarbon Contamination in South Carolina Salt Marsh-Tidal Creek Systems: Relationships Among Sediments, Biota, and Watershed Land Use. Arch Environ Contam Toxicol, 57, pp. 103–115.

Hartwell, S.I. and J. Hameedi. 2007. Magnitude and extent of contaminated sediment and toxicity in Chesapeake Bay, NOAA technical memorandum NOS NCCOS 47: 234 pp.

Hoek, G., R. Beelen, K. de Hoogh, D. Vienneau, J. Gulliver, P. Fischer and D. Briggs. 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment, 42 (33), pp. 7561-7578.

Huang, Y., M. Liu, R. Wang, S. Khalil Khan, D. Gao and Y. Zhang. 2017. Characterization and source apportionment of PAHs from a highly urbanized river sediments based on land use analysis. Chemosphere, 184, pp. 1334-1345.

Jedynska, A., G. Hoek, M. Wang, M. Eeftens, J. Cyrys, M. Keuken, C. Ampe, R. Beele et al. 2014. Development of Land Use Regression Models for Elemental, Organic Carbon, PAH, and Hopanes/Steranes in 10 ESCAPE/TRANSPHORM European Study Areas. Environmental Science & Technology, 48 (24), pp. 14435-14444.

Khairy, M.A., R. Lohmann (2013). Source apportionment and risk assessment of polycyclic aromatic hydrocarbons in the atmospheric environment of Alexandria, Egypt. Chemosphere, 91, pp. 895-903.

Kimbrough, K.L., W.E. Johnson, G.G. Lauenstein, J.D. Christensen and D.A. Apeti. 2008. An Assessment of Two Decades of Contaminant Monitoring in the Nation's Coastal Zone. Silver Spring, MD. NOAA Technical Memorandum NOS NCCOS 74. 105 pp.

Kimbrough, K.L., G.G. Lauenstein and W.E. Johnson. 2007. Organic contaminant analytical methods of the National Status and Trends Program: 2000-2006. NOAA Technical Memorandum NOS NCCOS 30. Silver Spring, MD.

Kubošová, K., J. Komprda, J. Jarkovský, M. Sáňka, O. Hájek, L. Dušek, I. Holoubek and J. Klánová. 2009. Spatially Resolved Distribution Models of POP Concentrations in Soil: A Stochastic Approach Using Regression Trees. Environmental Science & Technology, 43 (24), pp. 9230-9236.

Merbitz, H., M. Buttstädt, S. Michael, W. Dott, and C. Schneider. 2012. GIS-based identification of spatial variables enhancing heat and poor air quality in urban areas. Applied Geography, 33, pp. 94-106.

MLRC (Multi-Resolution Land Characteristics). 2020. https://www.mrlc.gov/. Accessed: 5/2020.

Noth, E.M., S.K. Hammond, G.S. Biging and I.B. Tager. 2011. A spatial-temporal regression model to predict daily outdoor residential PAH concentrations in an epidemiologic study in Fresno, CA. Atmospheric Environment, 45 (14), pp. 2394-2403.

O'Connor, T.P., 2002. National distribution of chemical concentrations in mussels and oysters in the USA. Marine Environmental Research, 53(2), pp. 117-143.

Papritz, A. and P.U. Reichard. 2009. Modelling the risk of Pb and PAH intervention value exceedance in allotment soils by robust logistic regression. Environmental Pollution, 157 (7), pp. 2019-2022.

Peng, C., Z. Ouyang, M. Wang, W. Chen, X. Li and J. Crittenden. 2013. Assessing the combined risks of PAHs and metals in urban soils by urbanization indicators. Environmental Pollution, 178, pp. 426-432.

Roldán-Wong, N. T., K. A. Kidd., B. P. Ceballos-Vázquez., A. R. Rivera-Camacho., and M. Arellano-Martínez. 2020. Polycyclic aromatic hydrocarbons (PAHs) in mussels (Modiolus capax) from sites with increasing anthropogenic impact in La Paz Bay, Gulf of California. Regional Studies in Marine Science, 33, 100948.

Uher, E., C. Mirande-Bret and C. Gourlay-Francé. 2016. Assessing the relation between anthropogenic pressure and PAH concentrations in surface water in the Seine River basin using multivariate analysis. Science of The Total Environment, 557–558, pp. 551-561.

Yang, L., S. Jin, P. Danielson, C. Homer, L. Gass, A. Case, C. Costello, J. Dewitz, J. Fry, M. Funk, B. Grannemann, M. Rigge and G. Xian. 2018. A New Generation of the United States National Land Cover Database: Requirements, Research Priorities, Design, and Implementation Strategies, ISPRS Journal of Photogrammetry and Remote Sensing, 146, pp. 108-123.

U.S. Department of Commerce

Gina M. Raimondo, *Secretary*

National Oceanic and Atmospheric Administration

Benjamin Friedman, *Under Secretary for Oceans and Atmosphere (Acting)*

National Ocean Service

Nicole R. LeBoeuf, *Assistant Administrator for Ocean Service and Coastal Zone Management (Acting)*

The mission of the National Centers for Coastal Ocean Science is to provide managers with scientific information and tools needed to balance society's environmental, social and economic goals. For more information, visit: **http://www.coastalscience.noaa.gov/**